2

INTRICACIES OF SPACED RETRIEVAL

A Resolution

Henry L. Roediger III and Jeffrey D. Karpicke

Robert Bjork has spent his entire professional life studying learning and memory, and many of us have spent our lives (in part) reading his path-breaking research. One interesting characteristic of Bob's work, much of it conducted in collaboration with Elizabeth Bjork, is the often counterintuitive nature of the findings emanating from their lab. At the risk of overstatement, one can view many of the important contributions that the Bjorks have made as creating a paradox and then mounting a satisfactory explanation for it.

Our chapter will deal with several paradoxes raised by the Bjorks' work. There are three interrelated puzzles. First, remembering an event that is repeated is greatly aided if the first presentation is forgotten to some extent before the repetition occurs. (Yes, you read it correctly—good remembering of an event can depend on its forgetting.) Second, retrieving an event can be a more potent learning opportunity than restudying it, which flies in the face of educational wisdom that studying creates learning and testing merely measures it. Third, putting these two paradoxes together, testing an event has a greater effect if one waits for some forgetting to make retrieval more effortful and difficult.

This last claim seems especially puzzling, because if we want to test people, shouldn't we want to do it under conditions in which they cannot make errors? After all, the idea of learning through "errorless retrieval" is a hallmark of certain approaches to memory remediation

in brain-damaged individuals. As we shall show, these approaches advocating errorless retrieval imply a wrong assumption, at least in healthy people (the case may be different in older adults and brain-damaged individuals). We can thank Bob and Elizabeth Bjork for these insights. In this chapter we unpack them and show how and when they are true.

FORGETTING AN EVENT CAN ENHANCE ITS RELEARNING

The theme of this volume is how successful forgetting can sometimes enhance remembering. The case is most obvious in studies of directed (or intentional) forgetting. If people must remember two sets of information successively, they can learn and remember the second set better if they have been told just before learning it that they can forget the first set of material they recently learned. That is, if two lists are presented, getting a forget instruction for the first list improves retention of a second list relative to the case where subjects feel responsible for remembering the first list while learning the second list. Establishing this fact was one of Bob Bjork's first major scientific contributions (e.g., Bjork, LaBerge, & LeGrand, 1968; Bjork, 1970). Intentional forgetting has been examined in many studies over the years, and whole volumes are devoted to it (Golding & MacLeod, 1998).

Forgetting of information can lead to successful remembering in another, more paradoxical way, too. Strangely, successful remembering of information can depend—in certain situations—on having successfully forgotten (to some degree) the same information earlier. The previous statement may seem weird or even patently absurd, but we review evidence here that it is true. Once again, Bob Bjork was responsible for this critical insight (Bjork & Allen, 1970). The condition in which the previous statement holds true occurs when an event to be remembered is repeated in some form, either restudied or tested. To the extent that a first presentation is forgotten, its repetition will be well remembered. Bjork gleaned this insight from research on the spacing effect and then extended it. The spacing effect (e.g., Glenberg, 1976; Madigan, 1969; Melton, 1970) refers to the situation when events are repeated and the spacing or lag between repetitions is varied. When an event occurs, its repetition has little effect on retention when the repetition occurs immediately (when the event is still fresh from its first presentation), but the impact grows as the repetition is delayed. Figure 2.1 shows a typical spacing effect from a careful study by Madigan (1969) in which words were presented once or twice in a long list and the spacing between repetitions was manipulated. Free recall of the list items was the dependent

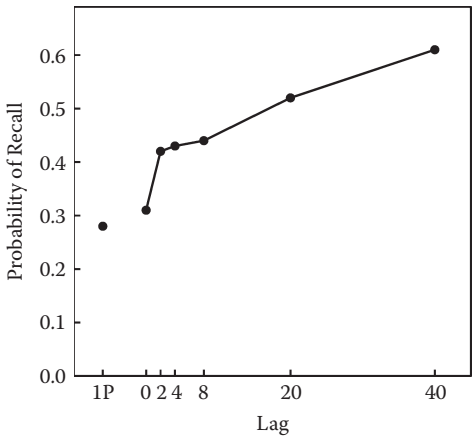

Figure 2.1 The basic spacing (or lag) effect in free recall. Words were either presented once (1P) or repeated. When repeated, items occurred at various spacings indicated on the abscissa. Later recall increased as a function of lag between presentations. (Data are adapted from Madigan, S. A., *Journal of Verbal Learning and Verbal Behavior*, *8*, 828–835, 1969.)

measure. Of course, as the lag increases since the first presentation in these kinds of experiments, the first presentation is increasingly forgotten. So the second presentation creates more durable learning when it occurs with an increasing amount of forgetting of the first presentation (up to some limiting condition). Crowder (1976, Chapter 9) spells out the logic quite clearly.

Bjork and Allen (1970) took this observation from the spacing effect and created an experiment in which the time between presentations would be held constant, but forgetting could still be manipulated by varying the difficulty of the task given to the subject between presentations. Subjects either performed a difficult task (one causing more forgetting) after the first presentation or performed an easy task (with less forgetting) between presentations. Sure enough, the second presentation led to greater recall on a final criterial test when it occurred after the difficult rather than the easy interpolated task. Ergo, forgetting of the information causes its greater retention after a repetition. Others replicated this finding (Robbins & Wise, 1972; Tzeng, 1973), but it may not extend to all situations (Roediger & Crowder, 1975). However, Logan, Roediger, and McDermott (2010) have shown how this principle—greater forgetting prior to a representation leading to greater recall—may benefit foreign language vocabulary learning.

AU: 2009 in references.

More recently, Storm, Bjork, and Bjork (2008) examined recall of items after two presentations. After the first presentation, some items

were subjected to retrieval-induced forgetting (using the Anderson, Bjork, & Bjork, 1994, technique) and others were not. All items were repeated and then recalled on a later test. Storm et al. found that "items that were relearned benefited more from that relearning if they had previously been forgotten" (2006, p. 230). They commented that this outcome "is very surprising from a common sense standpoint" (ibid.). Of course, so are all the findings reviewed here: How and why should greater forgetting of an event before it is presented again cause better later retention? That mystery runs throughout this chapter (see Crowder, 1976, pp. 273–314, for ideas in the context of spacing effect research).

AU: 2008 meant?

RETRIEVAL AS A MEMORY MODIFIER

The remainder of this chapter is about the effects of testing one's memory on later retention. This is not a new topic. In fact, it predates the festschrift for Bob Bjork by exactly 100 years, if we take the date as being of the first empirical papers we can find on the topic (Abbott, 1909; see Roediger & Karpicke, 2006a, for a review). The discovery made by Abbott and replicated by countless others is that the effect of taking a test is not neutral but alters later retention. When information is correctly retrieved on a test, this act makes the probability of future retention on a delayed test greater than if no test had occurred or even if the person had restudied the material rather than being tested on it (see Roediger & Karpicke, 2006b; Whitten & Bjork, 1977, among many others).

In the cognitive psychology of memory, the 1970s were the heyday of studies of retrieval, with many important papers on topics such as the encoding specificity principle (Tulving & Thomson, 1973) and transfer-appropriate processing (Morris, Bransford, & Franks, 1977). Another milestone publication of that era was R. A. Bjork's (1975) chapter that has the same title as this section heading. He argued that educators and psychologists both tended to ignore the importance of testing. He wrote:

> Retrieval from memory is often assumed, implicitly or explicitly, as a process analogous to the way in which the contents of a memory location in a computer are read out, that is, as a process that does not, by itself, modify the state of the retrieved item in memory. In my opinion, however, there is ample evidence for a kind of Heisenberg principle with respect to retrieval processes: an item can seldom, if ever, be retrieved from memory without modifying the representation of that item in memory in significant ways. (1975, p. 123)

Bjork's chapter went on to report research on retrieval as a memory modifier. He interpreted the phenomenon of the testing/retrieval effect through the lens of the then-new levels of processing ideas of Craik and Lockhart (1972), maintaining that there could be levels of processing during retrieval just like there were during encoding. Specifically, when retrieval occurred under easy, superficial conditions, it did not benefit later retention. However, when retrieval involved more difficult and complex processes, the effects on later recall were much greater. Thus, all acts of retrieval are not equal: Some confer great benefit and some provide little or no benefit. We return to this theme, too, later in the chapter.

A couple of years later, Whitten and Bjork (1977) reported an elegant experiment that documented Bjork's earlier points quite well. We report only a sketch of the logic here; the actual experiment was more complex. The authors presented subjects with two words to be remembered and then had them perform a distracter task for varying amounts of time afterwards: 4, 8, or 14 seconds. At this point, the items were either presented again or tested. Subjects had to recall the pair of words on test trials. When items were tested, recall declined from .72 to .61 to .54 across the three intervals. No feedback was given, so the level of retrieval success is critical in the case of tested events. Of course, when items were restudied, subjects were reexposed to 100% of the original items, so testing put items in that condition at a disadvantage relative to repeated study conditions, especially in the long-delayed conditions. A final test was given a bit later, after many items had been presented in these conditions.

We consider here the final test results for items that were studied twice or studied once and then tested. For simplicity, we consider only the extreme lags in this figure, those items that had been tested or restudied after lags of 4 or 14 seconds during the initial learning phase. The results can be seen in Figure 2.2, with data points estimated from Whitten and Bjork's (1977) Figure 1. Final recall showed a spacing effect in both cases: Performance was better when the second presentation or the test occurred after 14 seconds of distracter activity rather than only 4 seconds, which shows the usual lag effect (in both restudy and testing). In addition, final recall performance was better in the condition in which subjects had taken a test during learning than when they had restudied the item. Note that this testing effect occurred despite the fact that, as the delay increased, recall on that first test became increasingly poor, such that barely more than half the items (54%) were recalled on the initial test after 14 seconds of distracter activity. Thus when overt recall occurred early (after 4 seconds), it had less of a positive effect on

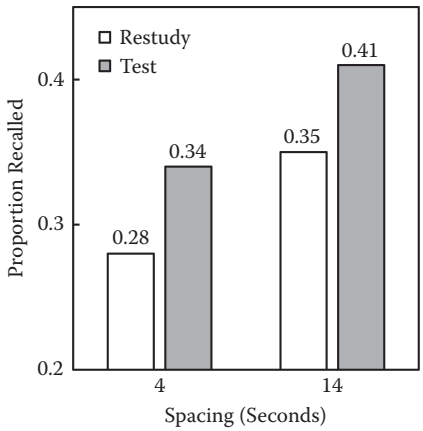

Figure 2.2 Final recall results as a function of the lag between study of a word pair and its restudy (white bars) or test (gray bars). Both spacing (lag) and testing had positive effects. (Data adapted from Whitten, W. B., & Bjork, R. A., *Journal of Verbal Learning and Verbal Behavior*, *16*, 465–478, 1977, Figure 1.)

a final test than when recall occurred after 14 seconds (despite the fact that recall on the initial test dropped during this period). Whitten and Bjork interpreted the results as indicating that retrieval difficulty was the critical component and cited related research by Jacoby and Bartz (1972) as reinforcing their point (see too Gardiner, Craik, & Bleasdale, 1973; Jacoby, 1978).

Whitten and Bjork's (1977) results may look rather slender. The advantage of testing to restudy was only 6% or so, although it was consistent. However, the problem of low performance on the initial test should be borne in mind. When Whitten and Bjork performed conditional analyses, examining final recall performance conditional on subjects successfully recalling items on the first test, the testing effect was much larger. Yet such conditional analyses raise the specter of item selection effects. The general logic is that "easier" items are, by definition, the ones recalled on the initial test. Therefore, any resulting advantage of recalling these items at a higher level on the delayed test may be due to selection of ease items in this condition rather than an effect of testing. That is true, but many analyses have shown convincingly that testing effects are not due to item selection effects and are not restricted only to easy items (Karpicke, 2009; Karpicke & Roediger, 2007b, 2008; Roediger & Karpicke, 2006), and even in Whitten and Bjork's study there was an absolute advantage in the testing conditions when all items

AU: Please indicate 2006a or 2006b.

were included. Testing effects are often quite large in other experiments (see Roediger & Karpicke, 2006a).

EXPANDING RETRIEVAL SCHEDULES

In 1978 Landauer and Bjork provided another important empirical contribution that has guided research and thinking in the intervening years. They asked: If testing aids retention (and it does), and if multiple tests provide greater benefits to retention than do single tests (also true), what schedule of testing provides the best performance? If we want to learn a person's name, or foreign language vocabulary, or definitions of scientific concepts, what is the best way to schedule our self-testing? This question is of critical importance for students who must learn a large body of factual knowledge.

Two experiments by Landauer and Bjork (1978) sought an answer. The authors contrasted several different possible schedules of testing that will be described momentarily. The materials subjects learned were paired associates (either first names with last names in Experiment 1 or face-name pairs in Experiment 2). After studying a pair, students received various schedules of repeated tests.

We will describe selected conditions of their Experiment 1 here. In one condition, items were presented only once. In four other conditions, four schedules of repeated testing with various schedules of spacing between tests were used. Three tests were given in all conditions, but the lags between tests varied according to the four schedules of spacing. The conditions of repeated testing were uniform-short spacing, uniform-moderate spacing, expanding spacing, and contracting spacing. (We provide an operational explanation of these labels shortly.) During tests, students were given the first name of the person and asked to produce the last name (in Experiment 1). In the two uniform conditions, the three tests were given with equal intervals between them. Thus, in the uniform-short condition, students were tested three times immediately after studying a pair. Following Landauer and Bjork (1978), we will refer to this condition as the 0-0-0 condition, because no intervening items occurred between tests. The uniform-moderate condition employed a 5-5-5 schedule of spacing, meaning that five intervening study or test events occurred between the tests of a particular pair in this condition. This condition is also called an equal interval condition, because the interval between tests is equivalent. The expanding test condition used a 1-4-10 spacing, indicating that a pair was first tested after only intervening item, then after 4 more, and finally after 10 intervening items. In the contracting condition, the spacing was reversed: 10, 4, and 1. Many

items were tested in these various conditions. In addition, as a baseline control condition, some items were presented a single time and never tested, which permits an answer to the question of what benefits the various testing schedules have over and above a single presentation of a pair with no testing. A final point is that all tests in these experiments were given without feedback.

After the acquisition phase of the experiment just described, students were given a final test 30 minutes later (with a lecture occurring during the interval). They were again given the first name of the pair and asked to produce the last name. The results (estimated from Figure 2 of Landauer & Bjork, 1978) are presented in Figure 2.3. All the testing conditions produced better final recall than the single-presentation study condition, but performance differed widely among the testing conditions (despite the fact that the number of prior tests was held constant at 3). The uniform-short (0-0-0) condition was poorest, the uniform-long (5-5-5) and contracting (10-4-1) conditions were intermediate, and the expanding condition (1-4-10) was best. Note that the three latter

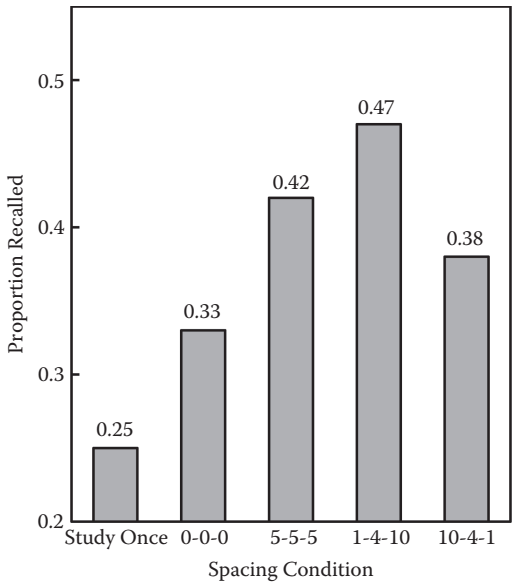

Figure 2.3 Final recall after either a single presentation (study once) or a single presentation and three tests. Schedules of the three tests had a large effect on recall. All testing conditions aided recall relative to the single presentation condition, but the massed testing condition conferred the least benefit, and the expanding retrieval condition produced the most benefit. (Data adapted from Landauer, T. K., & Bjork, R. A., in M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical Aspects of Memory*, Academic Press, London, England, 1978, pp. 625–632, Figure 2.)

conditions all have equivalent numbers of total events between tests (15); the critical point is how they were distributed. The expanding retrieval schedule was best. Experiment 2, which used face-name pairs, reported the same finding.

Landauer and Bjork (1978) extolled expanding retrieval as the best method to learn new information such as names and faces, and probably everyone reading their report quickly agreed. (The first author of this paper certainly did, when he read it; the second author here was not yet born.) The underlying rationale seems so straightforward and the benefit seems commonsensical (after the fact). The advice would be that when you meet a new person and hear her name, you should retrieve it rather quickly before the name is lost from immediate (working) memory. That initial retrieval ensures you encoded the item. After that initial retrieval, you should then wait a bit longer and retrieve it again, to practice retrieval at an intermediate time span. Then, finally, you should wait even longer for a further retrieval that would solidify or consolidate the memory more permanently. Landauer and Bjork wrote: "The expanding procedure may thus be seen as an effective shaping procedure for successively approximating the desired behavior of unaided recall at long delays" (Landauer & Bjork, 1978, p. 631).

For many psychologists who had learned about shaping of behavior through reinforcement by successive approximations to the desired behavior (e.g., Skinner, 1953, Chapter VI), the principle seemed intuitive (at least with the 20/20 wisdom of hindsight). Many textbook writers and teachers (again, the first author included) began to preach that expanding retrieval was the best way to practice new information to retain it best.

The main thrust in the remainder of this chapter is to claim that, despite the early rush to embrace expanding retrieval as a central technique in using retrieval-enhanced learning via testing, the idea is fundamentally flawed. As it has usually been operationalized in extant research, expanding retrieval has a fatal flaw: The first test given (often after lags of zero or one intervening item from initial presentation) makes retrieval "too easy," and making retrieval easy undermines its positive effect. We provide evidence below to support this claim, but of course it took many years for researchers to understand this point. After all, the data in Landauer and Bjork's (1978) paper showed that expanding retrieval was better than equal interval retrieval, so what is the problem? We describe it below.

Although Landauer and Bjork's (1978) claims now seem wrong to us, Bob Bjork actually anticipated the problem in his writings before that 1978 paper, ones we reviewed above. In his 1975 chapter, Bjork argued

that retrieval difficulty is critical to the testing effect—the more difficult the retrieval on a first test, the better the later recall on a second test. However, in the wisdom of hindsight, the expanded retrieval technique makes an initial retrieval very easy: In any schedule in which the first retrieval occurs after a lag of zero intervening items, it is essentially perfect, and with one intervening item performance does not drop much. These are the standard lags for initial tests in expanding retrieval conditions. As we shall see later in the chapter, the difficulty of the first retrieval in the typical expanding scheme is critical to later performance. But we are getting ahead of ourselves. Before we discuss this later part of the story, we will review (albeit briefly) the 30-year historical impact of the Landauer and Bjork paper by selectively reviewing research from 1978 to 2007.

EXPANDING RETRIEVAL: RESEARCH AND CONTROVERSY

A strange thing happened to research in this area after publication of Landauer and Bjork's (1978) landmark paper: nothing. For many years no one did research on the issue of expanding retrieval, at least not compared to that for equally spaced retrieval. The matter seemed to have been considered a closed case; no further research seemed needed. Why? Our guess is that the findings (although new) made so much sense that everyone nodded and said "of course." The fact that the findings were compelling and intuitive seemed to choke off further inquiry into the matter for about a decade. On the positive side, many people talked about the findings and included them in lectures and books, which was hardly surprising because they were interesting and were directed at an important practical problem.

In this section, we provide a selective overview of research directed at this issue after the 1978 paper until 2007, when a spate of new research was published. Balota, Duchek, and Logan (2007) have provided a much more thorough review of work during this period, which should be consulted for additional detail.

A few researchers did examine expanding retrieval sequences as a mode of learning. Rea and Modigliani (1985) tested third grade school children as they learned multiplication facts and spelling words. However, their control condition was massed testing—four tests with no other items between tests (0-0-0-0, using the notation above). Rea and Modigliani (1985) showed that an expanded retrieval sequence (0-1-2-4) was more effective than massed retrieval, but they did not have the critical equally spaced condition, and so total spacing was confounded with condition. Other researchers also compared expanding schedules

of retrieval to various other conditions, again usually massed testing or sometimes expanding study (rather than testing) schedules. They generally concluded that the expanding testing schedule was better either in neurologically impaired patients (e.g., Camp & McKitrick, 1992) or in healthy adults learning names (e.g., Morris & Fritz, 2002) than were massed schedules or multiple presentations without testing. However, the critical equal interval testing condition was not included.

In the first study since Landauer and Bjork's original one comparing expanding retrieval to equally spaced retrieval, Shaughnessy and Zechmeister (1992) were able to replicate Landauer and Bjork and showed a small positive effect of expanding retrieval over equally spaced retrieval on a test given soon after acquisition. However, a few years later Cull, Shaughnessy, and Zechmeister (1996) obtained quite mixed evidence across a series of five experiments. The results were puzzling, so Cull (2000) followed up this work with a more dedicated effort. Without going into the details of all the experiments (see Balota et al., 2007), suffice it to say that Cull found no evidence that expanding retrieval schedules provided any benefit to recall relative to equal interval schedules (although both led to better performance than did massed testing schedules). Carpenter and DeLosh (2005) also showed no superiority of expanding to equal interval training. In fact, the trend (during both acquisition and retention phases) was for the equal interval condition to be superior.

Balota, Duchek, Sergent-Marshall, and Roediger (2006) mounted a study with large numbers of young adults, healthy older adults, and other older adults with early-stage Alzheimer's disease. Because the subjects had widely different memory abilities, Balota et al. began all subjects with two massed tests of paired associates to ensure subjects had encoded the material well before implementing further massed, spaced, or equal interval schedules. Thus, all subjects received five tests after a single presentation with the following schedules: 0-0-0-0-0, 0-0-3-3-3, or 0-0-1-3-5. During acquisition, massed testing produced essentially perfect performance in all subject groups, whereas the expanding condition led to greater performance on the last test than did the equal interval condition. Because expanding retrieval led to better performance during learning, one might expect this benefit to carry forward to the final criterial test at the end of the session. However, this did not happen. Despite the fact that spaced retrieval produced much greater final recall than did massed retrieval for all three groups of subjects, expanding retrieval was not better than equal interval retrieval in any of the groups. Thus, once again, no evidence was found support-

ing Landauer and Bjork's hypothesis that expanding retrieval could "shape" later recall.

Two other points are worth making about the Balota et al. (2006) results. First, in the massed condition, subjects were tested on and successfully recalled all items. Thus, there were five successful retrievals under conditions that fostered errorless retrieval, thought on some accounts to be optimal for later performance (because subjects never make an error or draw a blank). However, this massed condition produced the worst performance on the final test, probably because the retrievals were effortless and shallow (Bjork, 1975). The second point is more subtle: Recall that at the end of learning, the expanding retrieval condition produced higher performance than the equal interval condition, yet on the delayed test, the two conditions were equivalent. What this pattern must indicate is that forgetting occurs more rapidly after expanding retrieval than after equal interval retrieval. In fact, this same pattern occurred in Landauer and Bjork's (1978) original study. In almost all the experiments discussed thus far, the final criterial test was at the end of one experimental session (but see Cull, 2000). The pattern of differential forgetting between conditions suggests that, with much longer retention intervals, there may be a reversal—retention may actually be better following equal interval retrieval practice relative to expanding retrieval practice. We consider designs with such delays in the next section.

THE MYSTERY OF EXPANDING RETRIEVAL PRACTICE AND ITS VICISSITUDES: A PARTIAL SOLUTION

At this point in the chapter, the reader is rightfully confused. Landauer and Bjork (1978) found that expanding retrieval is superior to massed or equal interval retrieval, and their finding accords well with other ideas in the learning and memory literature, such as shaping and errorless retrieval. Although their conclusion about expanding retrieval was accepted for many years (and all studies show that it is superior to massed retrievals), evidence since the mid-1990s paints a mixed picture. Why? We attempt to answer that question in this section by relying on two related concepts championed by Bob Bjork.

Recently Bjork (1999) has advocated an important and counterintuitive idea about the relation between initial learning performance and long-term retention. There are many instances where the rate

and level of initial learning is very good relative to some other condition, yet these seemingly beneficial conditions ultimately produce poor long-term retention as assessed on delayed tests (again, relative to a companion condition in which learning was slower). Stated another way, conditions that make initial learning slower and more difficult might produce worse initial learning performance but lead to gains in long-term retention. Bjork has called this the idea of creating "desirable difficulties" to promote learning, and he has gathered a variety of evidence supporting this concept (see Bjork, 1999; Schmidt & Bjork, 1992). Some difficulty that makes initial learning slower and more effortful can make long-term retention better.

An example of desirable difficulties relevant to this chapter is the spacing effect: When repeated presentations are massed together, they often produce better performance on an immediate test (one soon after the second presentation) than does spacing the presentations (Peterson, Wampler, Kirkpatrick, & Saltzman, 1963). However, as is well known, spaced repetition produces better retention on delayed criterial tests than does massed practice (see Figure 2.1). This spacing × retention interval interaction for studied materials is both replicable and important (see Balota, Duchek, & Paullin, 1989; Balota et al., 2007). The same pattern occurs if we consider spaced retrieval practice. Performance is essentially perfect on massed repeated tests (e.g., with a 0-0-0 schedule) and will be better than performance on equally spaced tests because forgetting will have occurred before the first retrieval attempt (e.g., with a 5-5-5 schedule). Yet invariably the spaced retrieval conditions produce better performance on delayed retention tests than does massed retrieval.

In short, Bjork's key point from the concept of desirable difficulties is that performance during initial learning is not necessarily diagnostic of long-term retention. This fact has profound implications for education and other training scenarios, because instructors often use initial learning performance as the metric by which they evaluate the effectiveness of learning and training activities. They rarely test performance long after the learning episode to determine what is retained.

Returning to the focus of this chapter—schedules of retrieval practice—an expanding retrieval condition is bound to perform better during the initial learning phase than an equally spaced condition. That is, subjects are likely to recall more items in an expanding condition than in an equally spaced condition because the first retrieval attempt occurs soon after study in the expanding condition. In most experiments on spaced retrieval, subjects are not given feedback after each test, but there is also very little (if any) forgetting across tests after the first one. Therefore, the position of the first test determines the level of

performance on subsequent tests. If 80% of items are recalled on test 1, then approximately 80% will be recalled on repeated tests. If 60% are recalled on test 1, then about 60% will be recalled on repeated tests. And so on. This fact is independent of the schedule of repeated tests and is apparent in Landauer and Bjork's (1978) data (Figure 3 in their paper) and in other experiments, too. The difference in level of performance across conditions is entirely due to the position of the first test. Yet the surprising finding is that the forgetting rate seems faster in the expanding than in the equally spaced condition. This is indicated in studies where there are large advantages of expanding relative to equally spaced conditions during initial learning, but no differences between the conditions on retention tests given at the end of the experimental session (e.g., Balota et al., 2006). Again, the same pattern can also be seen in Landauer and Bjork's data (their Figure 3).

Does the concept of desirable difficulties help explain the puzzling effects of retrieval practice schedules? That is, does expanding retrieval promote good performance during initial learning (greater retrieval success than equally spaced schedules) but result in relatively poor long-term retention? A number of recent experiments have addressed this question and suggested that the answer is yes.

We carried out a series of experiments in which subjects learned difficult vocabulary words under a variety of spaced retrieval conditions (Karpicke & Roediger, 2007a). We examined massed (0-0-0), expanding (1-5-9), and equally spaced (5-5-5) conditions, and we also included two conditions in which subjects took just a single test during initial learning: The single test occurred either after a lag of one trial or after a lag of five trials. The latter two conditions are conceptually similar to those used by Whitten and Bjork (1977) and others (e.g., Jacoby, 1978). The critical aspect of the experiment was that we manipulated the retention interval that occurred between the initial learning phase and the final criterial test: Half of the subjects took the final test at the end of the experimental session (about 10 minutes after the initial learning phase) and half took the final test 2 days later.

Figure 2.4 shows the proportion of word pairs recalled on the final tests in each spacing condition at the two different retention intervals. First, it is worth pointing out that at both retention intervals the spaced retrieval conditions (expanding and equal interval) led to better recall than did massed retrieval. The left panel of Figure 2.4 provides recall on the final test that occurred shortly after learning, and the data show an advantage of expanding retrieval relative to equal spacing. This outcome replicates Landauer and Bjork's (1978) original finding and is due to greater retrieval success during the learning phase, because

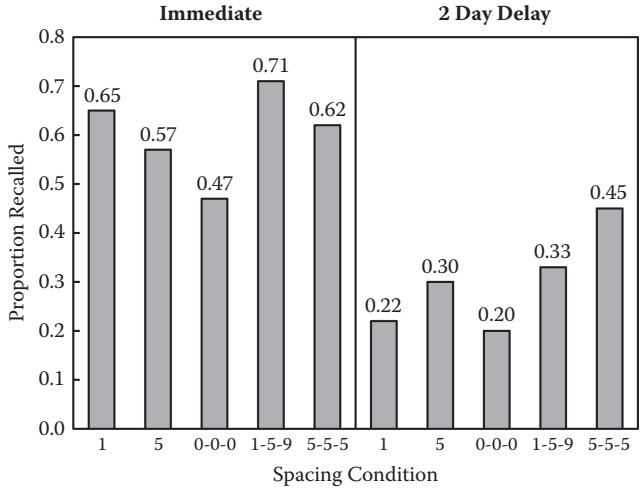

Figure 2.4 Final recall as a function of various schedules of retrieval practice. The left panel shows final recall 10 minutes after the learning phase, and the right panel shows final recall 2 days after the learning phase. Expanding retrieval (1-5-9) produced a short-term benefit relative to equally spaced retrieval (5-5-5), but equally spaced retrieval produced better long-term retention than expanding retrieval. (Data adapted from Karpicke & Roediger, 2007, Experiment 1.)

AU: Please indicate 2007a or 2007b, and then give full reference info in source line.

the expanding condition recalled more items initially than the equally spaced condition. However, two days after learning the pattern had reversed: Now the equally spaced condition produced better long-term retention than expanding retrieval.

Note that a similar interaction occurred when considering just the single-test conditions: A single test after a short delay during acquisition (one intervening item) produced better recall than a single test after a somewhat longer delay (five intervening items) both during acquisition and on the immediate test, but on the test given two days later, the single, more effortful initial test (the one after five intervening items) led to better retention than the easier initial test (the one given after one item).

Another feature of the data in Figure 2.4 documents the fact that giving several tests under conditions that are too easy undermines the positive effects of testing. In the 0-0-0 condition subjects were required to recall items three times under conditions in which they were essentially always correct. However, these three (easy) retrievals led to later retention that was even worse than a *single* test given under more difficult conditions (the five conditions at both delays).

Logan and Balota (2008) also recently conducted an experiment examining the effects of expanding and equally spaced retrieval schedules at short and long retention intervals. They tested both younger and older adults and examined several different spacing schedules. The subjects in their experiments learned weakly associated word pairs under different schedules and took a final test either at the end of the experimental session (immediate) or one day later. The results are shown in Figure 2.5. Overall, Logan and Balota did not find a consistent advantage of expanding retrieval over equally spaced retrieval in either subject group at either retention interval. In fact, they found that equally spaced retrieval was often better than expanding retrieval on the delayed final test.

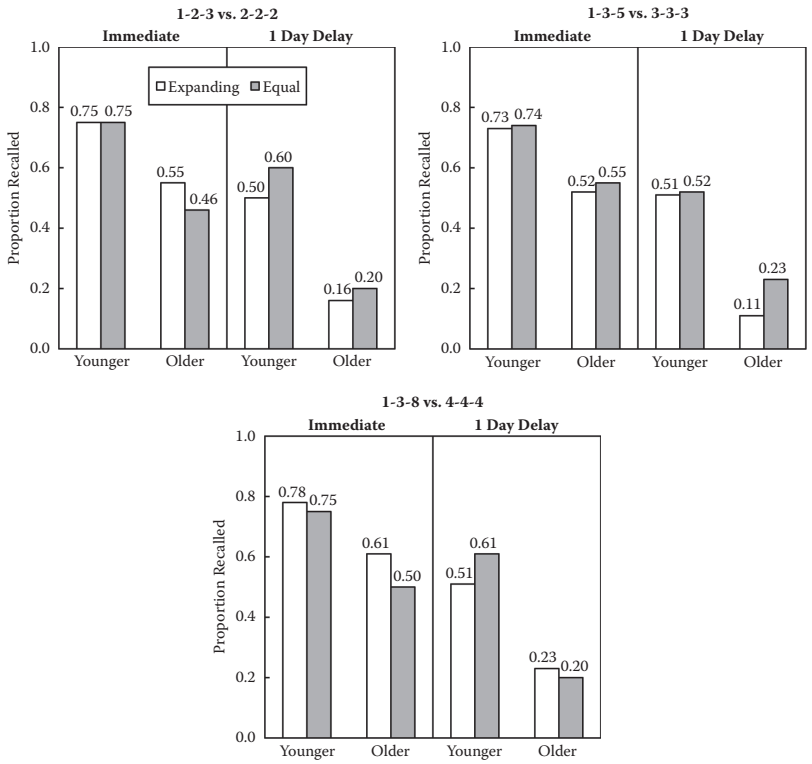

Figure 2.5 Final recall after expanding or equally spaced retrieval practice on immediate or one-day delayed tests. The figure shows results for both younger and older adults. The top, middle, and bottom panels show performance for different expanding and equally spaced schedules that are matched in total spacing. No advantage of expanding retrieval was evident, and equally spaced retrieval often produced better final recall than expanding retrieval on the one-day delayed test. (Data adapted from **Logan**, J. M., & Balota, D. A., *Aging, Neuropsychology, and Cognition*, *15*, 257–280, 2008.)

The Karpicke and Roediger (2007a) and Logan and Balota (2008) results might seem strange given the belief that expanding retrieval is supposed to improve long-term retention. But the findings are consistent with Bjork's concept of learning tasks that produce desirable difficulties. The desirable condition, however, is the equally spaced retrieval schedule, not expanding retrieval.

This pattern of results must force us to reconsider the theory about why expanding retrieval ought to work. The standard theory of expanding retrieval practice is that the schedule combines the positive features of retrieval success and retrieval difficulty. Of course, difficult retrieval is important, but unless subjects are given feedback (and they are not in most spaced retrieval studies), retrieval practice can only promote learning when a person is able to successfully recover the desired item. Therefore, expanding retrieval is thought to work in part because the early first retrieval promotes retrieval success and, as noted above, this determines the level of performance on repeated tests. Retrieval difficulty comes into play because it is assumed that gradually increasing the spacing of repeated tests should increase retrieval difficulty on the tests. However, Karpicke and Roediger (2007a) and Logan and Balota (2008) examined response times on tests during initial learning and showed that retrieval grew increasingly faster across repeated tests. This does not accord with the idea that retrieval grew increasingly difficult across tests regardless of the schedule of repeated tests.

The alternative hypothesis we have proposed is that the position of the first test is the important difficulty for improving long-term retention, not the schedule of repeated tests (see Karpicke & Roediger, 2007a). In expanding retrieval conditions, the first retrieval attempt often occurs almost immediately after studying the item (lags of zero or one trial). This retrieval attempt might not be effective because retrieval occurs while items still reside in immediate memory. Therefore, equally spaced retrieval practice might enhance retention because that schedule involves a delayed first test (e.g., a lag of five trials between study and a first test).

The crux of the problem in virtually all comparisons of expanding and equal interval retrieval is that the position of the first retrieval attempt is confounded with the schedule of repeated tests. Expanding retrieval conditions involve an immediate first test (e.g., 1-5-9), and equally spaced conditions involve a delayed first test (e.g., 5-5-5). We conducted an experiment that eliminated this confound (Karpicke & Roediger, 2007a, Experiment 3; see too Carpenter & DeLosh, 2005). Two conditions involved an immediate first test (after a lag of zero trials), and two involved a delayed first test (after a lag of five trials). Then the repeated tests were either expanding (1-5-9) or equal (5-5-5). The results are shown

in Figure 2.6. When we controlled for the position of the first test, the advantage of expanding retrieval practice disappeared on an immediate final test (cf. Carpenter & DeLosh, 2005) and there was no difference as a function of placement of the first test (0 or 5). However, on the test two days later, an overall advantage of the two conditions with a delayed first test (5-1-5-9 and 5-5-5-5) appeared (relative to the conditions in which the first test was immediate). Thus, in delayed recall, the effect of position of the first test mattered, but the schedule of repeated tests (expanding or equally spaced) did not have any effect. This result falls perfectly in line with the results of Whitten and Bjork (1977) and accords with Bjork's (1975) notion that difficult retrieval is critical for promoting learning, but once again, it does not support the idea that expanding retrieval is the best schedule of retrieval practice for long-term retention.

We end this section by describing an experiment that explored the effects of different schedules of retrieval on learning educational texts. Landauer and Bjork's (1978) original study was focused on a rather specific applied scenario: learning faces and names when it is inappropriate or impossible to receive feedback after an initial presentation.

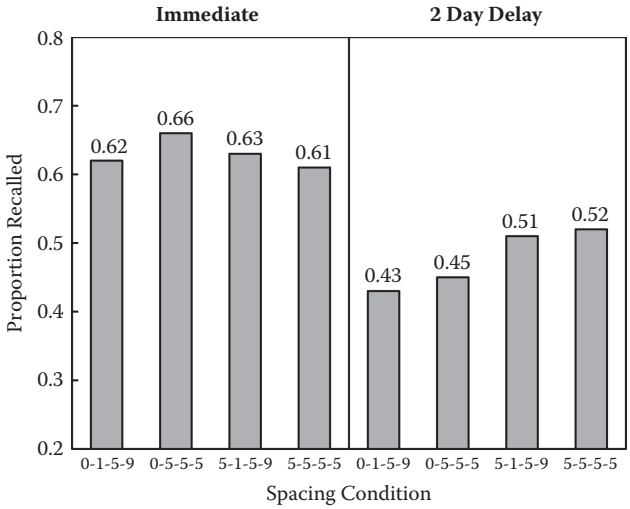

Figure 2.6 Final recall as a function of schedule of retrieval practice. The left panel shows final recall 10 minutes after the learning phase, and the right panel shows final recall 2 days after the learning phase. The four retrieval schedules factorially crossed the position of the first test (lags of 0 or 5) with the schedule of repeated tests (1-5-9 or 5-5-5). There was no effect of schedule on immediate final tests, but there was a main effect of delaying the first test on the delayed final tests. (Data adapted from Karpicke, J. D., & Roediger, H. L., *Journal of Memory and Language*, *57*, 151–162, 2007a, Experiment 3.)

The idea of expanding retrieval practice emerged from this study, and subsequently the argument was made that expanding retrieval was a general technique that could be applied broadly. The data reviewed here suggest that expanding retrieval might not represent the best retrieval schedule for promoting long-term retention, but as of yet there have been few tests of the idea that expanding retrieval might apply broadly to materials and contexts that are more educationally relevant than those used in paired-associate learning tasks. Perhaps when taken out of the context of paired-associate learning, an advantage of expanding retrieval would become apparent.

To address this question, we examined free recall of brief expository texts (Karpicke & Roediger, 2010). Subjects read brief texts and recalled them on free recall tests spaced according to different schedules. In both experiments we factorially crossed the position of the first test (immediate or delayed) and the spacing of repeated tests (expanding or equal interval). We examined the effects of the different retrieval practice schedules on a final criterial test one week after learning.

Figure 2.7 shows several important results. First, there is a testing effect: taking a single test after reading a text enhanced long-term retention more than reading the text and not testing. Second, repeated testing (in the spaced retrieval conditions) enhanced retention more than taking a single test. Third, testing with feedback (restudying the passages) produced better retention than testing without feedback. However, and most importantly for our purposes, there were no differences between expanding and equally spaced schedules of retrieval practice.

In sum, the body of evidence indicating that expanding retrieval practice is not beneficial (relative to equal interval practice) is growing. If anything, equal interval schedules seem to produce better retention on delayed tests, probably because the initial test is rendered more difficult when it does not occur immediately after study, as is the case in expanding schedules of retrieval. The difficulty of the initial retrieval seems to hold the key to performance in experiments of this kind. The subsequent schedule of retrieval practice seems to have little effect under conditions examined thus far.

PRACTICAL ADVICE

What advice might we give students about how to apply the research on testing reviewed in this chapter? We think the answer is straightforward: Students should determine the knowledge they want to retain, create a testing mechanism with feedback, and test themselves until they can retrieve the information on a much-delayed test (say, two days

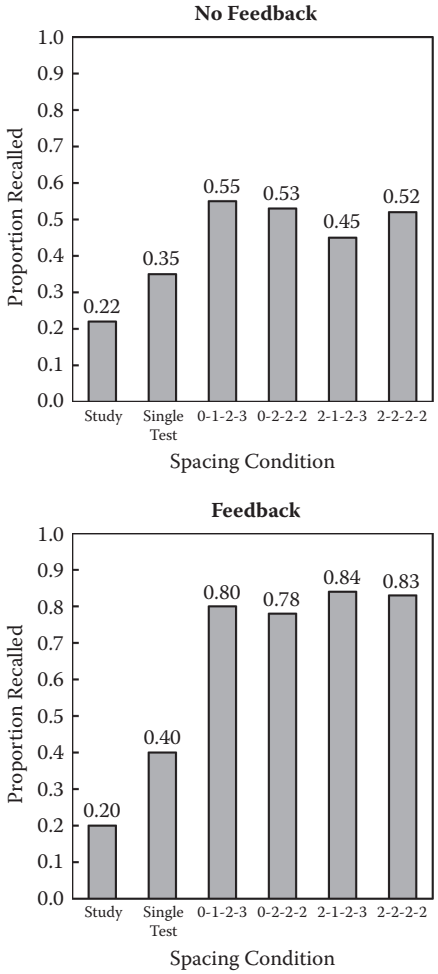

Figure 2.7 Final recall of expository texts as a function of initial retrieval practice schedule. The top panel shows performance without initial feedback, and the bottom panel shows performance with feedback (students reread the texts after each recall test). Taking a single test enhanced retention relative to reading once, and repeated testing produced even greater effects on retention. Feedback also enhanced long-term retention. However, the schedule of retrieval did not matter. (Data adapted from Karpicke, J. D., & Roediger, H. L., *Memory and Cognition*, *38*, 116–124, 2010, Experiment 2.)

since original study). The testing should not be done under massed or even closely spaced fashion; if the literature is clear on any point, it is that repeated testing under conditions in which retrieval is easy leads to poor long-term retention. (So much for the principle of errorless retrieval being a good way to study.) But what about the mechanism for spacing of retrieval? Our data reviewed above suggest that the critical ingredient is encouraging fairly difficult retrieval, especially on an initial test. Beyond that point, it probably does not matter whether students test themselves using expanding or equal interval conditions. What matters is repeated spaced retrieval (with feedback if an error is made).

Let us consider a practical example. A fifth grade student needs to learn the capitals of the 50 states. She creates flash cards for each state with, for example, Montana on one side and Helena on the other. The 50 flashcards would first be studied one at a time, perhaps employing some mnemonic (my aunt Helen was from Montana). After this initial study, the cards are shuffled and then ten minutes later the student gives herself a test, looking at the name of each state and trying to remember the capital. Whether or not she produces a name, she turns the card over to study the reverse side (see Butler, Karpicke, & Roediger, 2008). Any items missed are put at the end of the deck for further practice in the same session. She records the number correct on the first pass through and then returns to test herself again on the ones she missed, again with feedback. After this phase, the student puts the cards away and studies other material. Then, hours later, she returns to the cards and tests herself in the same way. This process would be repeated the next day and then sporadically thereafter, as needed. Each time the deck would be shuffled anew. With spacing between retrievals spread over days, the whole issue of schedule of individual state-capital pairs within a session would not need to be much considered. Of course, the spacing of entire testing/relearning sessions would then be of interest.

One critical point about the foregoing advice: students should not trust their own intuitions about what they know and quit testing themselves too soon. Just because Helena can be retrieved a time or two does not mean that it is in a "learned" state. Students need to practice retrieval even of learned information (Karpicke & Roediger, 2008).

The technique just described can be applied to nearly any sort of factual material—scientific concepts, the critical points of important journal articles, the presidents of the United States and their main accomplishments and events while they were in office, and so on. The title of one of our articles is "Repeated Retrieval During Learning Is the Key to Long-Term Retention" (Karpicke & Roediger, 2007a), and we believe more firmly than ever that this is the case.

CONCLUSION

We began the chapter by noting how Bob and Elizabeth Bjork's work had, over the years, pointed to several apparent paradoxes (or at least nonintuitive findings). We explored several paradoxes and applied their (and our) analyses to the issue of the best way of practicing retrieval over relatively short intervals, such that testing can be used to best advantage. All studies show that repeated massed retrieval is poor, despite its errorless nature. Bjork (1975) has argued this was true based on data then available. However, the mystery of whether expanding or equal interval retrieval leads to better long-term retention turns out to rest on a similar consideration. When retention is measured at a healthy delay (say two days or one week after learning), delayed recall is better following equal interval practice because (in the usual design) the first retrieval in the equal interval design occurs under more difficult retrieval conditions. Thus, expanding retrieval turns out to exemplify the Bjorkian principle of a desirable difficulty—although initial recall is poorer with equal interval schedules relative to expanding schedules, long-term retention is better.

Our results provide a resolution of claims in the literature: Landauer and Bjork's (1978) results can be replicated at short retention intervals (when testing occurs in the same experimental session as acquisition). However, after longer retention intervals (two days or a week in our experiments), the situation reverses: Equal interval schedules of retrieval practice in an initial learning session produce better retention than expanding schedules of retrieval practice. We suggest in the preceding section on practical applications that, so long as one uses sessions of spaced retrieval practice with feedback, the question of expanding or equal interval schedules within a session may well be moot. Spaced retrieval practice (with feedback) is the key to long-term retention.

Even though Landauer and Bjork's (1978) important claim about expanding retrieval turns out 30 years later to be limited (or even wrong), the reasons for this state of affairs are accounted for by Bjork's other research and theorizing (Bjork, 1975; Bjork & Bjork, 1994). Even when Bob Bjork seems to be wrong in one arena, he turns out to have been right all along.

AU: Only Bjork in references (one author).

REFERENCES

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs*, *11*, 159–177.

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

Balota, D. A., Duchek, J. M., & Logan, J. M. (2007). Is expanded retrieval practice a superior form of spaced retrieval? A critical review of the extant literature. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger, III* (pp. 83–105). New York, NY: Psychology Press.

Balota, D. A., Duchek, J. M., & Paullin, R. (1989). Age-related differences in the impact of spacing, lag, and retention interval. *Psychology and Aging*, *4*, 3–9.

Balota, D. A., Duchek, J. M., Sergent-Marshall, S. D., & Roediger, H. L. (2006). Does expanded retrieval produce benefits over equal interval spacing? Explorations in healthy aging and early stage Alzheimer's disease. *Psychology and Aging*, *21*, 19–31.

Bjork, R. A. (1970). Positive forgetting: The noninterference of items intentionally forgotten. *Journal of Verbal Learning and Verbal Behavior*, *9*, 255–268.

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

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., & Allen, T. W. (1970). The spacing effect: Consolidation or differential encoding? *Journal of Verbal Learning and Verbal Behavior*, *9*, 567–572.

Bjork, R. A., LaBerge, D., & LeGrande, R. (1968). The modification of short-term memory through instructions to forget. *Psychonomic Science*, *10*, 55–56.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 218–928.

Camp, C. J., & McKitrick, L. A. (1992). Memory interventions in Alzheimer's-type dementia populations: Methodological and theoretical issues. In R. L. West & J. D. Sinnott (Eds.), *Everyday memory and aging: Current research and methodology* (pp. 152–172). New York, NY: Springer-Verlag.

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, *19*, 619–636.

Craik, F. I. M., & Lockhart, R. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684.

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

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14*, 215–235.

Cull, W. L., Shaughnessy, J. J., & Zechmeister, E. B. (1996). Expanding understanding of the expanding-pattern-of-retrieval mnemonic: Toward confidence in applicability. *Journal of Experimental Psychology: Applied*, *2*, 365–378.

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

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior*, *15*, 1–16.

Golding, J. M., & MacLeod, C. M. (Eds.). (1998). *Intentional forgetting: Interdisciplinary approaches.* Mahwah, NJ: Lawrence Erlbaum.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, *17*, 649–667.

Jacoby, L. L., & Bartz, W. H. (1972). Rehearsal and transfer to LTM. *Journal of Verbal Learning and Verbal Behavior*, *11*, 561–565.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*, 469–486.

Karpicke, J. D., & Roediger, H. L. (2007a). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*, 151–162.

Karpicke, J. D., & Roediger, H. L. (2007b). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 704–719.

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

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

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London, England: Academic Press.

Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition*, *15*, 257–280.

Logan, J. M., Roediger, H. L., & McDermott, K. B. (2009). Using spaced retrieval practice to learn foreign language vocabulary: How does activity during the interval affect learning? Manuscript in preparation.

AU: Please update.

Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal of Verbal Learning and Verbal Behavior*, *8*, 828–835.

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior*, *9*, 596–606.

Morris, D. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533.

Morris, P. E., & Fritz, C. O. (2002). The improved name game: Better use of expanding retrieval practice. *Memory*, *10*, 259–266.

Peterson, L. R., Wampler, R., Kirkpatrick, M., & Saltzman, D. (1963). Effect of spacing presentations on retention of a paired associate over short intervals. *Journal of Experimental Psychology*, *66*, 206–209.

Rea, C. P., & Modigliani, V. (1985). The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. *Human Learning*, *4*, 11–18.

Robbins, D., & Wise, P. S. (1972). Encoding variability and imagery: Evidence for a spacing-type effect without spacing. *Journal of Experimental Psychology*, *95*, 229–230.

Roediger, H. L., & Crowder, R. G. (1975). Spacing of lists in free recall. *Journal of Verbal Learning and Verbal Behavior*, *14*, 590–602.

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

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

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*, 207–217.

Shaughnessy, J. J., & Zechmeister, E. B. (1992). Memory-monitoring accuracy as influenced by the distribution of retrieval practice. *Bulletin of the Psychonomic Society*, *30*, 125–128.

Skinner, B. F. (1953). *Science and human behavior.* Oxford, England: Macmillan.

AU: Please cite in text.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*, 641–656.

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2008). Accelerated relearning after retrieval-induced forgetting: The benefit of being forgotten. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 230–236.

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

Tzeng, O. J. (1973). Stimulus meaningfulness, encoding variability, and the spacing effect. *Journal of Experimental Psychology*, *99*, 162–166.

Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 465–478.