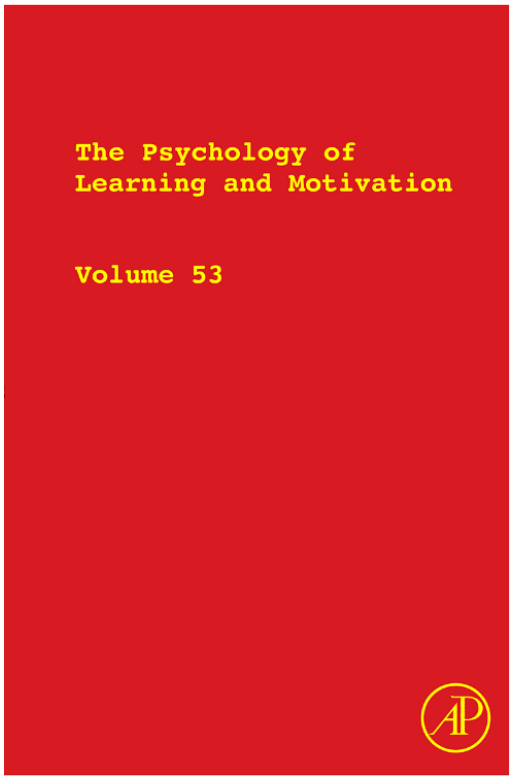

From: Peter F. Delaney, Peter P. J. L. Verkoeijen, and Arie Spirgel,
Spacing and Testing Effects: A Deeply Critical, Lengthy, and
At Times Discursive Review of the Literature
In Brian H. Ross editor:
The Psychology of Learning and Motivation, Vol. 53,
Burlington: Academic Press, 2010, pp.63-148.
ISBN: 978-0-12-380906-3

CHAPTER THREE

Spacing and Testing Effects: A Deeply Critical, Lengthy, and At Times Discursive Review of the Literature

Peter F. Delaney, Peter P. J. L. Verkoeijen, *and* Arie Spirgel

Contents

Psychology of Learning and Motivation, Volume 53
ISSN 0079-7421, DOI: 10.1016/S0079-7421(10)53003-2

Abstract

What appears to be a simple pattern of results—distributed-study opportunities usually produce better memory than massed-study opportunities—turns out to be quite complicated. Many "impostor" effects such as rehearsal borrowing, strategy changes during study, recency effects, and item skipping complicate the interpretation of spacing experiments. We suggest some best practices for future experiments that diverge from the typical spacing experiments in the literature. Next, we outline the major theories that have been advanced to account for spacing studies while highlighting the critical experimental evidence that a theory of spacing must explain. We then propose a tentative verbal theory based on the SAM/REM model that utilizes contextual variability and study-phase retrieval to explain the major findings, as well as predict some novel results. Next, we outline the major phenomena supporting testing as superior to restudy on long-term retention tests, and review theories of the testing phenomenon, along with some possible boundary conditions. Finally, we suggest some ways that spacing and testing can be integrated into the classroom, and ask to what extent educators already capitalize on these phenomena. Along the way, we present several new experiments that shed light on various facets of the spacing and testing effects.

1. Introduction

This chapter reflects our best attempt to review the state of theoretical and empirical knowledge on the family of memory effects that deal with the impact of studying the same thing several times—the distributed-practice family. Extra study opportunities produce better memory, but how we distribute those study opportunities is also important for memory.

The distributed-practice family of effects comprises a variety of phenomena, including the spacing effect, lag effect, and testing effect. Cognitive psychologists have produced hundreds of papers over the last century arguing that there is a *spacing effect*—that is, a memory advantage to restudying something with a delay between the repetitions compared to immediate restudy. The spacing effect is often viewed as an instance of the broader *lag effect*, in which longer spacing intervals are associated with changes in later recall. Specifically, the lag effect reveals that short spacing results in lower recall relative to moderate spacing, and very long spacing begins to show declines again. Finally, the spacing effect's first cousin is the *testing effect*, which refers to the advantage of testing an item relative to just studying it again. Thus, the distributed-practice family includes several of memory theory's favored children because of their obvious implications for improving education.

We intend this chapter to serve as a comprehensive review of the spacing and testing literature and their associated theories, circa 2010. We are due for a long narrative review of the spacing literature anyway. This review, like many others, culminates with a theoretical proposal that attempts to explain the vast range of empirical results in the spacing literature. We also present some new data and draw attention to the importance of some recent papers, whose importance might otherwise be missed.

What we think our review contributes beyond that is a careful experimental analysis of the task used in spacing experiments: verbal list learning. No one is inherently excited about word lists, but they have been used in the preponderance of studies on the spacing effect, and therefore understanding what people are doing in these experiments is critical. We will take the rather strange stance that there is a ''real'' spacing effect somewhere and that all of the other (e.g., rehearsal borrowing, strategy changes during study) phenomena are ''imposters'' that masquerade as the spacing effect. Just because many different phenomena have a similar observable outcome—namely, better memory for spaced repetitions than for massed repetitions—does not mean that all of these phenomena are the same. It would be like arguing that giving extra study time and asking people to process items for survival value are ''really the same thing'' because they both result in better memory for studied items. We cannot rely on similar outcomes in recall rates as the sole diagnostic criterion for identifying the spacing effect. For example, ''deficient processing'' accounts of spacing propose that when people encounter a massed repetition, they exert less encoding effort on the second presentation than they do for second spaced repetitions. Several studies have demonstrated that deficient processing does happen in some cases, and it produces a spacing effect. Furthermore, the deficient-processing effect can be discriminated from other spacing effects because it weakens the benefit of massed repetitions over single presentations rather than enhancing the recall of spaced items relative to massed repetitions. Therefore, although it is

phenomenonologically similar, the deficient-processing effect is not the ''real'' spacing effect—it is an impostor.

These impostors often produce effect sizes as large as or larger than the ''real'' spacing effect. Furthermore, they may operate in the same direction as the real spacing effect, thereby greatly exaggerating its impact, or they may operate in the opposite direction from the real spacing effect, canceling it out. Without a careful experimental analysis of participants' behavior during the verbal learning task, it is quite difficult to understand the circumstances under which the ''real'' spacing effect occurs and the circumstances under which it does not. This confusion has produced a bewildering thicket of experimental results that seemingly contradict one another. In this chapter, we do our very best to untangle the thicket on a briar by briar basis, identifying the impostor phenomena and providing guidelines for running future impostor-free spacing experiments.

A crucial part of this effort involves the analysis of the strategies that participants use when they study lists of words. For the past 10 years, our laboratories have worked to understand what people do when they encounter an instruction to ''study words for a later memory test'' and how the strategies they choose interact—often in surprising ways—with the number of lists people study, whether items are repeated in a massed or spaced fashion, and whether massed and spaced repetitions are mixed together on the list or kept on separate lists. We think that very few experimental studies meet rigorous standards for comparing theoretical views about the ''true'' cause of spacing effects, because human participants do not cooperate with researchers by ''just behaving normally'' during memory experiments. Instead, they devise a variety of clever strategies for memorizing lists of words, and these strategies interact in surprising ways with the structure of the lists to affect memory. For example, we will see that rote rehearsal strategies sometimes enhance and sometimes reduce the impact of spacing, depending on the structure of the list.

2. A Field Guide to the Spacing Literature: Spotting Impostors

There have been three major meta-analyses of the spacing literature conducted in the past decade (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Donovan & Radosevich, 1999; Janiszewski, Noel, & Sawyer, 2003), which produced conflicting results that depend on what studies were included. The most comprehensive meta-analysis of verbal learning was the most recent (Cepeda et al.), which identified confounds in some earlier studies and included the largest number of studies. For each study, they assessed the lag between repetitions (i.e., how much time passes between

each repetition), and the retention interval (i.e., how much time passes between the last repetition and the test). For any given retention interval, there is an optimal lag between repetitions that maximizes memory. Shorter-than-optimal and longer-than-optimal lags between repetitions produce suboptimal memory. Furthermore, as the retention interval increases, so does the optimal lag between repetitions. Therefore, memory is a function of both the retention interval and the lag between repetitions. Finally, they found that the same pattern held for both free recall and cued-recall tests.

Their analysis represents the best current conclusions regarding the spacing effect. However, their reliance on verbal learning data is problematic due to the large number of confounds present in existing spacing studies. Specifically, there are a variety of impostor spacing effects that deserve their own names, and should be carefully watched for in studies that attempt to measure the "true" spacing effect. Table 1 outlines the major phenomena we will review here that affect the conclusions of many spacing studies.

2.1. Recency Effects

It is fitting to begin with recency effects, an impostor that is so well known that it stars in virtually every introductory psychology textbook's discussion of memory. The problem with recency confounds in spacing studies is an old one in the literature, highlighted by Crowder's (1976) review. Specifically, because spaced items must occur in multiple locations on the list, their final presentation tends to be more recent than an equal number of massed items unless care is taken to equate the final positions. Because recent items are more easily recalled than older items, an artifactual spacing effect can be observed. One approach to solving this problem, whose discovery was attributed to Melton (1967) by Crowder, was to use primacy and recency "buffer" items that would not be tested, or just not counted for free recall. In fact, this approach was used earlier by Waugh (1962), but it is not terribly effective at controlling recency. Zimmerman (1975), for example, found an extended recency function that produced 20% higher recall for later-presented than earlier-presented items, even though he included primacy and recency buffers. He required participants to focus on only the current item, which eliminated the primacy effect, but resulted in an extended recency function.

Even in recent work, recency control has been a problem. Toppino and Bloom (2002), in their Experiment 1, replicated an experiment of Greene (1989) that compared free recall following incidental and intentional learning. The lists contained some massed and some spaced items, with spaced items of varying lag. Greene tried to control for recency biases by counterbalancing the assignment of words to quadrants of the list.

Table 1 Five Impostors: Spacing-Like Phenomena.

1. *The recency effect.* Even if we control rehearsal, there is an extended recency function. Failing to account for this can artificially enhance the memory of spaced items, because their last presentations are more recent and therefore stronger.
2. *Rehearsal-borrowing effects on mixed lists.* Mixed lists encourage rehearsal borrowing, which artificially inflates the spacing effect on mixed lists relative to pure lists. The degree of borrowing varies depending on list structure as well, so one can create some super-spaced items unintentionally. This effect is often wrongly discounted as unimportant because spacing effects emerge in incidental learning, and because people often change encoding strategies during study (see Delaney & Knowles, 2005).
3. *The zero-sum effect on pure lists.* Because people rehearse during study, there is no guarantee—particularly with pure-list designs—that the primacy items wo'nt receive differential practice on some types of lists compared to others. A spacing effect occurs on pure lists if you throw away the beginning of the list, but only because the beginning of the list benefits tremendously from displaced rehearsal on all-massed lists.
4. *Deficient-processing effects.* There are a family of deficient-processing effects, including the Deficient-processing effect, in which processing is reduced; the Rose effect, in which people choose to spend less time on massed items when they have control over study time; and the speed effect, in which too-fast presentation rates encourage people to mass items or to skip spaced items.
5. *List-strength effects.* In free recall, there are output effects at recall that favor spaced items over massed items. These effects appear only on mixed lists and vanish on pure lists.

The Toppino and Bloom study was virtually an exact replication of the experiment, except that it more carefully controlled recency by controlling the position of the second presentation of words instead of just the quadrant. Surprisingly, this subtle change eliminated the spacing effect for incidental learning observed by Greene. The study highlights the fact that seemingly minor recency biases can inflate or deflate the magnitude of spacing effects, altering our conclusions about the magnitude of the spacing effect—or even its presence or absence under varying conditions.

2.2. Intentional Learning and Mixed Lists: Rehearsal Effects and Strategy-Switching

Our second spacing impostor is the rehearsal-borrowing effect. Like recency, rehearsal is a well-known phenomenon, but it also provides a convincing impostor spacing effect when rehearsal favors spaced items over massed items. It is a serious problem for most spacing studies, because

most spacing studies use a *mixed-list design*, meaning that they have massed repetitions and spaced repetitions on the same list. Furthermore, they use nonspecific instructions to study the words on the list for a later memory test, and therefore they do not really control how long people study each item. Such designs encourage *rehearsal borrowing* that redistributes study time away from massed items and awards it to spaced items. The obvious result of spending a much longer time in studying the spaced items than the massed items is that the spaced items are better remembered on a test.

Hall (1992a) went so far as to revive the theory that rehearsal borrowing was the *only* mechanism necessary to explain the emergence of a spacing effect in most studies. The borrowing explanation was first advanced in the original Atkinson and Shiffrin (1968) "modal model" paper. Atkinson and Shiffrin argued that people comply with instructions to study a list of words by reading each item and then rehearsing earlier-presented items in a short-term memory buffer. Because the buffer had limited capacity, adding new items to the rehearsal buffer resulted in dropping some earlier words. The time in the buffer—equivalent to the number of rehearsals the item received—would then predict its later strength and hence probability of final recall. Such a mechanism would naturally produce a spacing effect and a lag effect, because spaced items (but not massed items) appear in multiple places on the list. The longer the lag between presentations, the more likely it was that the item had already received a "full run" through the buffer when it was next encountered. Upon being refreshed, it would get a new run through the buffer, receiving extra rehearsals. However, massed items appear in only one location on the list, and therefore get only one "full run" through the rehearsal buffer. The result is more rehearsals for spaced than for massed items. Rundus (1971) verified this prediction using rehearse-aloud protocols, and discovered that the probability of rehearsing an item was directly predictive of its probability of later recall. He further showed that spaced items received more rehearsal than did massed items, demonstrating rehearsal borrowing.

If borrowing is pervasive on mixed lists, we would expect that mixed lists greatly overestimate the true benefit of spacing. Furthermore, if Hall's (1992a) contention were correct and rehearsal borrowing were the only mechanism necessary to explain the spacing effect, then spacing would be virtually useless as a learning tool. The goal of spacing practice is to improve memory for all of the to-be-learned items, not to selectively improve memory for a few of the items at the expense of the rest! Because of this concern, Hall used *pure lists*—that is, lists composed of only spaced items or only massed items—to see if the spacing effect would disappear once people could no longer borrow time from massed items to help the spaced items. In three experiments, he showed that studying pure lists eliminated the spacing effect on a free recall test, using presentation times ranging from 1 to 4 s per item. Furthermore, compared to a mixed list, the pure lists resulted in lower

recall of spaced items and higher recall of massed items. The latter result is consistent with Hall's argument that for mixed lists, rehearsal borrowing awards extra study time to the spaced items at the expense of the massed items. In another study, Hall (1992b) compared pure lists of spaced items with pure lists of once-presented items that were presented for the same total duration. At 2, 4, and 6 s per item presentation rates, he obtained no spacing advantages with free recall tests. Taken together, the results suggested that rehearsal borrowing might be a serious problem for our conclusions about the spacing effect, since virtually all of the studies in the literature use mixed-list designs together with intentional learning.

Two later studies seemed to overturn Hall's (1992a) conclusions, however. An important paper by Toppino and Schneider (1999) demonstrated that you could still get spacing effects on pure lists, provided multiple study lists were employed (with a free recall test after each list). We will later see that the inclusion of multiple lists within the session is important because people change how they study throughout the course of an experiment. Toppino and Schneider also included a condition that used a mixed list, but where each half of the list was pure. That is, the first half of the list contained only spaced or only massed items, while the second half contained the opposite type of item. These "special" lists would presumably reduce the extent of rehearsal borrowing across item types (if that borrowing tended to come from recent items). Their most crucial evidence against the rehearsal-borrowing explanation was that the pure lists and the "special" mixed lists produced relatively similar spacing effects (8% for mixed lists and 7% for the pure lists). It is worth noting, however, that Hall found that "regular" mixed lists produced spacing effects roughly twice as large (14%).

A later paper by Kahana and Howard (2005) also obtained spacing effects in free recall using pure lists, and further demonstrated that the lag effect was present. Results such as these—especially when combined with earlier papers that obtained spacing effects using pure lists[1] (Underwood, 1969, 1970)—seemed to indicate that rehearsal was less important than Hall (1992a) had believed. However, more recent work has suggested that the story is more complicated, and we will discuss this more recent research next.

2.2.1. People Do Not All Rehearse, and They Change Strategies with Practice

Hall (1992a) assumed that most people comply with the instructions to study words for a later memory test by rehearsing. But do they really? Ironically, there are almost no studies that have asked the straightforward

[1] Underwood's (1969, 1970) studies were atypical, however, in that they used very long presentation rates (10 s per item) and often many repetitions, which would tend to produce deficient processing effects; see below for more on deficient processing.

behavioral question, "What do people do when you tell them to study words for a later memory test?" When we create cognitive models, we typically implicitly assume that people (a) all do pretty much the same thing, and (b) do pretty much the same thing from one trial to the next. As someone trained in the problem-solving tradition, these assumptions seemed rather flimsy to the first author. After all, rather ordinary people can obtain digit spans greater than 70 with a few months' practice (e.g., Chase & Ericsson, 1981), and they rapidly discover better strategies than rote rehearsal. At the extreme, memory experts like the memorist Rajan will discover new mnemonic strategies to deal with memory tasks deliberately created to interfere with his existing mnemonic techniques in just a few days of practice (Ericsson, Delaney, Weaver, & Mahadevan, 2004).

We therefore conducted a series of studies using methods typically reserved for the thinking literature. We asked participants to study lists of words, but afterwards asked them to tell us what they were thinking as they studied the words. We then coded these verbal reports into strategy groups (Delaney & Knowles, 2005; Sahakyan & Delaney, 2003). It turns out that on the first list of words that people study, about 70% use a rote rehearsal strategy in which they read each item as it appears and then rehearse earlier items. However, rote rehearsal is not a terribly effective memory strategy, and if people receive a test after each list, they will often abandon rote rehearsal for something else.

The second most frequent strategy after rehearsal was the *story mnemonic* (Bower & Clark, 1969; Drevenstedt & Bellezza, 1993; Reddy & Bellezza, 1983), in which people make up a story using all the words on the list. There are various other "deep" mnemonics that people use, like linking each word to their own personal experiences or making up sentences using each word. On the first list, about 16% of participants used a deep encoding strategy. However, by the fourth study list, about equal numbers of people (43–44%) were using a deep strategy and the rote rehearsal strategy. Thus, when people study multiple lists, they tend to abandon rote rehearsal in favor of more effective strategies.

Tests are one way to induce people to switch strategies. In fact, you do not even have to explicitly test people; metacognitive judgments or various disruptions of the rehearsal strategy between two lists also result in strategy changes (see Sahakyan & Delaney, 2003, 2005; Sahakyan, Delaney, & Kelley, 2004). Strategy changes favoring better encoding on later-studied lists may also work to ameliorate the deleterious effects of proactive interference build-up in cases when people are instructed to study word lists without any specific instructions on the strategy to use during study (Szpunar, McDermott, & Roediger, 2008).

We have summarized the impact of encoding strategy on the magnitude of the spacing effect in Table 2, based on several recent studies conducted in our laboratories. Delaney and Knowles (2005) explored the role of study strategy in the spacing effect on pure lists of words. In Experiment 1, they

Table 2 Magnitude of the Spacing Effect in Free Recall by Encoding Strategy and List Type.

Strategy	Mixed lists	Pure lists
Rehearse each item alone	Small	Small
Rehearse the items together	Large	Null
Story mnemonic	Large	Small

Note: Assuming a list of 32 items presented twice and free recall testing, a small effect is about a 6% spacing advantage, a large effect is around 15%, and a null effect is less than 2%. Mixed lists contain both spaced and massed items, while pure lists contain only spaced or massed items (but not both).

partitioned their data into participants who used rote rehearsal and those who used a "deep" encoding strategy like the story mnemonic. Replicating Hall (1992a), when people reported using rote rehearsal, there was no significant spacing effect on pure lists—at best, it was a small (1–2%) advantage. There was no spacing effect regardless of how many lists they had studied, provided they stuck with rote rehearsal throughout. However, for people who switched strategies to a deep encoding strategy, the spacing effect emerged on pure lists. Thus, Delaney and Knowles concluded that Hall's participants, who saw only a single list, were mostly using rote rehearsal, and thus showed no spacing effect. However, later papers like Toppino and Schneider's (1999) study had people study multiple lists, which caused people to abandon the rote rehearsal strategy. Consequently, they obtained a significant spacing effect even on pure lists.

In a second experiment, Delaney and Knowles (2005) controlled the study strategy their participants used by instructing them to either use a rote rehearsal strategy or to use the story mnemonic. They again found no reliable spacing effect in the rote rehearsal condition, but a significant spacing effect in the story mnemonic condition, confirming their earlier results.

A similar study by Paivio and Yuille (1969) had earlier shown similar strategy-switching for cued recall. They found that participants often start by using a rehearsal strategy, but switch to a mediation or imagery-based strategy. Thus, the concern that the number of lists employed in spacing experiments, and the particular mix of strategies used, is not limited to free recall and single-item recognition experiments—although no one has specifically repeated the Delaney and Knowles (2005) experiments using cued recall. Bahrick and Hall (2005) have argued for item-specific strategy changes in cued recall, such that when people see a pair again, if they retrieve their earlier association they will strengthen it. However, if they fail to retrieve that association, then they generate a new one. In a Darwinian selection/retention process, successful mediators are retained while unsuccessful ones are replaced, resulting in better memory following long spacing of items.

2.2.2. Rote Rehearsal and the Borrowing Hypothesis Revisited

In a recent paper, we examined the rote rehearsal strategy in order to learn how rehearsal interacts with list structure (Delaney & Verkoeijen, 2009). Specifically, we asked our participants to rehearse using the rote rehearsal strategy as described to us by people who used it in our earlier laboratory studies. Our participants described a process we called the *rehearse-together strategy*, in which they would read each word as it appeared on screen and then use any remaining time to rehearse earlier items. Consistent with the Delaney and Knowles (2005) studies, we found that the rehearse-together strategy resulted in a null spacing effect on pure lists. However, it resulted in a large spacing effect on mixed lists. The same results were obtained with both free recall and recognition tests.

In order to understand how rehearsing groups of items affected memory, we compared the rehearse-together conditions to a *rehearse-alone condition*, in which participants read each word and then repeated only that item until the next item appeared (see also Wright & Brelsford, 1978; Zimmerman, 1975). In several experiments, we found identical small spacing effects on pure and mixed lists using the rehearse-alone condition. The experiments are particularly dramatic because they show that the "real" spacing effect—as manifest in the rehearse-alone condition—can be *doubled* in magnitude on mixed lists and eliminated on pure lists simply by changing how people study the lists. Another way of saying this is that the rehearsal confounds in a typical spacing experiment are larger than the spacing effect that the experiments are designed to study.

An earlier study by Wright and Brelsford (1978) also compared rehearse-alone and rehearse-together instructions although they used only the mixed-list conditions. In Experiment 1, they compared rehearse-alone and rehearse-together using overt rehearsal, and obtained no spacing effect with rehearse-alone instructions, but a significant spacing effect with rehearse-together instructions. However, their results were vulnerable to a floor effect interpretation (see p. 637), and we found that a spacing effect does emerge on mixed lists with rehearse-alone—it is just smaller than in the rehearse-together condition (Delaney & Verkoeijen, 2009). In their Experiment 2, they let people rehearse covertly, which may have allowed some of them to violate the instructions. However, they found results more similar to ours in that they obtained a larger spacing effect for rehearse-together than for rehearse-alone. Furthermore, in their rehearse-together condition, the massed items were recalled at a rate similar to singletons, consistent with displaced rehearsal.

Why does the rehearse-together strategy affect memory so differently on pure and mixed lists? One part of the story is that rehearse-together strategies manipulate recency effects in interesting ways. In the rehearse-alone condition, we obtained an extended recency effect (better memory for the end of the list) and no primacy effect (better memory for the

beginning on the list). This is similar to what one observes when people do not expect a test and do not try to study the words at all. When people rehearse items together, we obtained both primacy and recency effects. People tend to rehearse early items on the list throughout the entire duration of the list, making them artificially recent (cf. Tan & Ward, 2000). Another effect is that rehearsing earlier-studied items turns them into a kind of spaced item. When we asked people to rehearse out loud, we found that they tended to rehearse spaced items more frequently than massed items on the mixed lists (see also Rundus, 1971). In contrast, on pure lists, massed items benefit because they are more likely to receive distributed rehearsal than they would if people focused only on the current item, making them functionally similar to spaced words.

2.2.3. Summary

In summary, research often fails to control encoding strategy in spacing experiments, which results in participants adopting increasingly better study strategies across lists. Because different encoding strategies result in different magnitudes of the spacing effect, averaging across multiple lists, even when the order is counterbalanced, can produce misleading estimates of the true effect size.

Encoding strategies that encourage rehearsal borrowing tend to result in much larger spacing effects on mixed lists than on pure lists. In the typical studies conducted in the past, people have used mixed lists and intentional rehearsal, which encourage borrowing. Since the borrowing effect is as large as or larger than the actual spacing effect, such studies cannot provide accurate estimates of the true magnitude of the spacing effect.

2.3. Primacy and Recency Buffers: The Zero-Sum Effect

Our third impostor is also related to rehearsal borrowing, and we call it the "zero-sum effect" (Verkoeijen & Delaney, 2008). The zero-sum effect is a consequence of the common experimental practice of throwing away some of the items on the list and measuring recall of the rest. Waugh (1962) introduced the practice of including items at the beginning and end of the list—called primacy and recency buffers, respectively—that were not counted and served only to reduce the impact of primacy and recency biases on massed versus spaced comparisons. This practice has apparently been enforced by generations of spacing researchers, as it is used in the majority of studies. One of the unusual features of the Delaney and Knowles (2005) and Delaney and Verkoeijen (2009) studies is that they do not include any primacy or recency buffers. Consistent with our general position that everything spacing researchers think is good is really bad, we think primacy and recency buffers are problematic—especially if they are used on pure lists.

To understand why, it is important to first note that Toppino and Schneider (1999) showed that the serial position function of pure-spaced and pure-massed lists differ in interesting ways. Specifically, the pure-massed lists show an enhanced primacy effect compared to the pure-spaced lists, resulting in a crossover interaction such that massing produced better memory for the beginning of the list while spacing produced better memory in the rest of the list. Toppino and Schneider termed this the enhanced primacy effect.

However, we have already proposed that the spacing effect observed in Toppino and Schneider's (1999) study reflected a mixture of strategies. When one plots the serial position function for the rehearse-together strategy, one obtains an enhanced primacy effect, but no overall spacing effect (Delaney & Knowles, 2005; Delaney & Verkoeijen, 2009). The serial position function for the story mnemonic produces no primacy and a weak recency effect, with a spacing effect throughout the list. If one mixes together some rehearse-together participants and some story mnemonic participants—as we think Toppino and Schneider's study naturally did—one would obtain a function that displays enhanced primacy, but also has a spacing effect. As that is exactly the pattern they obtained, the strategy mixing seems quite plausible.

Just because one uses pure lists does not mean rehearsal-borrowing stops; it just means participants cannot borrow from massed items to help spaced items. One can still rehearse some items more often than others. There are well-established rehearsal frequency differences that depend on serial position, such as the primacy effect, which results from extensive rehearsal of the early items on the list (e.g., Tan & Ward, 2000). On pure-massed lists, the extra rehearsal for primacy items is likely to be greater than for pure-spaced lists, because each of those primacy items is presented right away for twice as long. On the pure-spaced lists, in contrast, primacy items are already being replaced with new items soon after they are introduced. According to this logic, the enhanced primacy effect on massed lists is a result of rehearsal patterns that strengthen items at the start of the list. A corollary of this argument is that the apparent spacing effect in the rest of the list might be due to rehearsal borrowing, such that the strong primacy-region items steal rehearsal time away from the rest of the list on the massed lists.

To test this idea, Verkoeijen and Delaney (2008) recently conducted a series of pure-list spacing experiments in which we required participants to use the rehearse-together strategy. As in our earlier studies, the spacing effect was small and nonsignificant. Our next step was to plot the serial position functions and to ask whether people who showed a bigger enhanced primacy effect—that is, a bigger massing advantage in the first quadrant—were the same people who showed a bigger spacing effect throughout the list. To illustrate, Figure 1 shows two participants, A and B. Participant A shows a large enhanced primacy effect, because she focuses on rehearsing the beginning of the massed list to a greater extent than

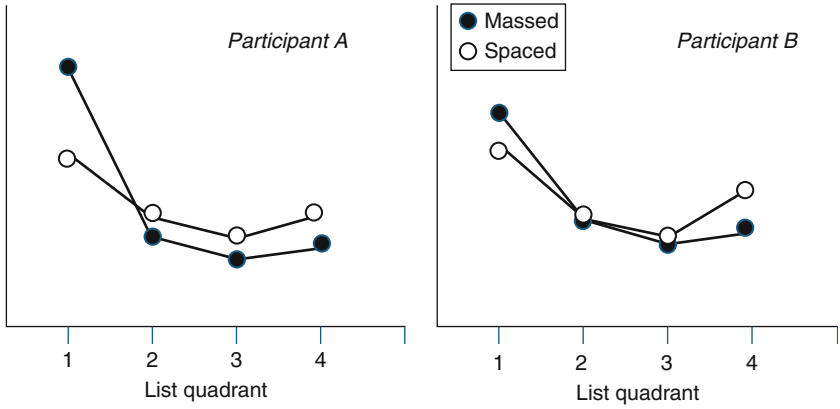

Figure 1 The zero-sum hypothesis proposes that if you show a bigger enhanced primacy advantage (Quadrant 1 is better recalled on massed than spaced lists), then you will show a smaller spacing advantage throughout Quadrants 2–4. Participant A rehearsed the beginning of the spaced list quite a lot, resulting in lower recall of the rest of the spaced list. Participant B showed a smaller primacy effect, and hence better recall of the rest of the list.

Participant B. However, this extra rehearsal of the beginning of the massed list comes at a cost; compared to Participant B, she shows less memory for the rest of the massed list, resulting in a spacing advantage throughout the rest of the list. Verkoeijen and Delaney called this the *zero-sum hypothesis*, as it suggests that the better you do on one part of the list, the worse you are likely to do on the rest of the list. Indeed, this is exactly the pattern we found: people who showed larger enhanced primacy effects were the same people who showed larger spacing effects in the rest of the list. People who showed little or no enhanced primacy effect also showed little or no spacing effect in the rest of the list, suggesting trade-offs in memory.

Turning back to the issue of primacy and recency buffers, it should be clear that they are part of the list to be studied from the perspective of the participants. Before we throw those parts of the list away, we should check whether the list structure affects the recall of the primacy and recency buffer items. Pure lists cease to be pure if they have primacy and recency buffers, because those items then receive rehearsal. Primacy buffer items, for example, are likely to be rehearsed throughout the list. This effect can be magnified if they are followed by a large number of massed repetitions, during which people will continue to rehearse the primacy buffer items. At this time, we also have no way of knowing whether on mixed lists the primacy buffer items receive more rehearsal during massed than during spaced repetitions. Hence, we do not favor the inclusion of primacy and recency buffer items—which is unfortunately a feature of the majority of spacing studies.

In summary, even designs that throw away the primacy and recency regions can result in rehearsal-borrowing effects that differ between spaced and massed lists. This is because people may not distribute practice to the primacy and recency regions equally in spaced and massed lists. Thus, rehearsal-borrowing problems persist even with pure lists. Many of these problems can be ameliorated by controlling encoding strategy and by measuring recall rates for the entire list, and not just a portion of it.

2.4. Deficient-Processing Effects

One of the earliest proposed explanations for the spacing effect involved deficient processing, which is our fourth impostor. The idea behind deficient-processing explanations was that the second time an item is encountered, processing the item is somehow easier than it was the first time. In verbal learning studies where people study individual words, there is not usually very much "processing" that people need to do; they read the word and activate its meaning. Deficient processing makes more sense when people need to generate something on each repetition. For example, if we ask people to rate a twice-presented word for pleasantness, they have no need to think about their answer on the second occurrence unless they have forgotten their original answer. An even clearer example of deficient processing was demonstrated by Jacoby (1978), who asked people to solve word puzzles that consisted of two words. The first word was a cue that helped participants to solve the puzzle, and the second word had some missing letters. For example, he might present shoe—F _ _ T, and the answer would be FOOT. Jacoby found that when people had recently seen the word FOOT, these puzzles became trivial, and later memory for the word was much lower on a surprise cued-recall test compared to puzzles they had solved themselves (see also Cuddy & Jacoby, 1982).

A classic demonstration of deficient processing was a study by Thios (1972), who presented participants with sentences whose subject and object were sometimes repeated in a later sentence. Repetitions either used the same "sense" of the subject and object, or a homographic "sense" of the subject and object. For example, if participants read, "The electric *drill* cut into the cinder *block*," then a same-sense repetition might be, "The hi-powered *drill* entered the masonry *block*." A homographic repetition might be, "The fire *drill* emptied the city *block*." After 80 sentences, they were cued with the subject words and had to recall the object words. The major result of the study was that there was a spacing effect in both conditions, but by comparing the massed repetitions to once-presented sentences, they determined that massed homographic repetitions improved memory more than did massed same-sense repetitions. The results suggest that sentences that were more dissimilar reduced the massed-item

processing deficit. (In contrast, for spaced items, the lag effect was larger with same-sense repetitions.)

Similar results were reported by Dellarosa and Bourne (1985). In Experiment 1, they either repeated sentences verbatim or paraphrased them. They further varied the lag, using massed repetitions and spaced repetitions with lags out to eight sentences. Changing the surface form of the sentence improved memory for massed repetitions, but had small and inconsistent impact on spaced repetitions. In Experiment 2, sentences were repeated using either the same-gender voice or a different gender voice. Switching the gender of the speaker improved memory for the massed sentences substantially, but improved memory for spaced sentences only slightly. Both of these results are consistent with a deficient-processing explanation whereby identical or nearly identical repetitions provide little benefit to memory when they are repeated without any lag.

Another source of deficient processing can be participants' own choices about how long to study. Zimmerman (1975) gave participants the option to control the rate at which items appeared on screen for study. By hitting the space bar, they could terminate the presentation and move to the next item. He found that people would terminate study of massed items more quickly than they would spaced items, suggesting that people would intentionally induce deficient processing on the massed items. Furthermore, people terminated study of short-lag items sooner than long-lag items, producing a lag effect. A study conducted by Shaughnessy, Zimmerman, and Underwood (1972, Experiment 3) produced similar results.

A recent study by Toppino, Cohen, Davis, and Moors (2009) raises another possibility for deficient processing—though in this case, for spaced repetitions. Toppino et al. manipulated the difficulty of study items, and showed that for more difficult items, participants often failed to fully perceive them at rapid presentation rates. Under these circumstances, they showed better memory for massed than spaced repetitions. The Toppino et al. study suggests that if the presentation rates in a typical spacing study are too fast, people may have no choice but to skip some of the items to cope with the fast pace. If so, they might favor massed items, which they feel they have time to process fully, and skip many of the spaced items. The item-skipping approach predicts that if the presentation time is very fast, you might observe a reverse spacing effect (i.e., better memory for massed items). It turns out that is exactly what one finds. Metcalfe and Kornell (2003) used Spanish–English word pairs to demonstrate that at a 0.5-s presentation rate, the spacing effect reverses itself, and at a 1-s presentation rate, it is a null effect (for further null spacing effects at 1-s presentation rates, see Waugh, 1963, 1967, 1970).

In sum, there are several conditions under which people will show marked deficient processing of massed items (e.g., the deficient-processing effect), and a few cases when they will show deficient processing of spaced

items (e.g., fast presentation). These results obviously complicate the interpretation of spacing effects observed in many experiments; just as with rehearsal effects, they can sometimes magnify and sometimes diminish the effects of spacing on learning.

2.5. Incidental Learning and Mixed Lists: List-Strength Effects

We would like to raise one final issue that is often important in considering spacing effects, and that is the presence of list-strength effects in free recall. However, far from being a negative feature of spacing experiments, we think list-strength effects provide important evidence regarding the source of spacing benefits. Therefore, while the list-strength effect makes the impostors list, we think it may be a consequence of the "true" spacing effect rather than a confound (more on that later, in Section 4).

The list-strength effect was first demonstrated by Tulving and Hastie (1972), who showed that items presented multiple times on a study list reduced recall of the once-presented items. This inhibitory effect was consistent with global memory models like SAM that assumed that repeated items accumulate context strength and that stronger items are therefore sampled more frequently when the context is used as a cue to retrieve them (Ratcliff, Clark, & Shiffrin, 1990). However, subsequent studies posed a problem for global memory models because they demonstrated convincingly that once rehearsal was controlled, recognition memory did not show global competition effects (e.g., Hirshman, 1995; Yonelinas, Hockley, & Murdock, 1992). A more general conclusion is that more difficult tasks that invoke recollective processes tend to show the list-strength effect (Diana & Reder, 2005; Murnane & Shiffrin, 1991; Norman, 2002). However, simple cued-recall or recognition tests are unlikely to show a list-strength effect (see also Bäuml, 1997).

The signature list-strength pattern is obtained by comparing recall on pure lists (i.e., all-spaced or all-massed lists) to mixed lists (i.e., lists with some spaced and some massed items). The list-strength effect consists of two effects when switching from pure to mixed lists. First, the spaced items show better recall on mixed than on pure lists. Second, the massed items show poorer recall on mixed than on pure lists. If this sounds familiar by now, it is because it is exactly the pattern obtained by Delaney and Verkoeijen (2009) in our studies on rehearsal. The concern that covert rehearsal was responsible for earlier list-strength effects led to extreme attempts to control encoding, but the final resolution of this work seems to be that list-strength effects emerge in incidental learning (Sahakyan, Delaney, & Waldum, 2008; Yonelinas et al., 1992). We also obtained a list-strength effect for free recall but not for recognition when we forced participants to use a rehearse-alone strategy (Delaney & Verkoeijen).

A further twist to this story is that the list-strength effect has been observed *only* with spaced repetitions (Malmberg & Shiffrin, 2005;

Sahakyan et al., 2008). Other methods of strengthening items, such as extra presentation time or deeper orienting tasks, increase recall of the stronger items, but do not produce the list-strength pattern; they produce only a main effect of strength, such that the strong items are recalled better than the weak items on both pure and mixed lists. Therefore, the list-strength effect can be *equated* with the spacing effect, and it directly predicts a larger spacing effect in free recall on mixed than on pure lists.

Perhaps the list-strength effect, like the other encoding effects mentioned in this section, is a confound that must be eliminated to understand the "real" spacing effect. However, another possibility is that the list-strength effect is an indicator as to the true source of the spacing effect. Specifically, we will argue later that a theory that incorporates some assumptions about how context is stored with a trace and how different types of tests use context can provide a viable explanation of the spacing effect, once the encoding confounds described in this guide are taken into account.

2.6. Summary: The Impostor Effects and Confounds in Spacing Designs

Our review of the impostor phenomena provides a bleak view of the spacing literature as a whole. Based on the above review, the "ideal" study should use presentation rates slow enough that people do not skip items. It should control recency very carefully, as even small biases in favor of spaced items can inflate estimates of the magnitude of the spacing effect. It should use pure-list designs (and perhaps compare those designs to mixed lists), and preferably have no primacy and recency buffers. Furthermore, it should carefully control the strategies participants use to study, preferably by using incidental-learning procedures.

How many of the hundreds of spacing studies have used a design of this type? The answer is vanishingly few. As we then consider the theories of spacing and the evidence against each of those theories, it may be worth keeping in mind that we are using flawed data to reject most of these theories—albeit lots of flawed data collected in multiple laboratories using multiple methods.

3. The Failure of Existing Spacing Theories

Before indicating what theoretical position we favor, we will examine the successes and failures of earlier theories. We cannot explore every theoretical perspective ever advanced in our limited space, so we will focus on theories that have been seriously considered by at least one

researcher in the past 20 years. Furthermore, we will mostly restrict our review to accounts of what we termed the "real" spacing effect, trying to ignore the various impostor effects that produce benefits of spacing over massing, but that apply in limited circumstances. To evaluate the theories, we will lay out what we see as the most important phenomena that spacing theories need to explain. Table 3 lists these major phenomena. In some cases, we will note that a phenomenon, although important, may need to be replicated under controlled circumstances in order to be sure that it is real. By the end, we will be poised to offer our thrilling alternative.

3.1. Intention Invariance

We already outlined (in Section 2) the rehearsal-borrowing effect. However, one of the earliest theories of spacing effects was that there was no "true" spacing effect, and it was all due to rehearsal borrowing (Atkinson & Shiffrin, 1968). While we agree that rehearsal borrowing is an important problem when interpreting the spacing literature, it cannot be the full explanation of the spacing effect because spacing effects still emerge robustly in incidental learning (Braun & Rubin, 1998; Challis, 1993; Glenberg & Smith, 1981; Greene, 1989; Paivio, 1974; Rose & Rowe, 1976; Sahakyan et al., 2008; Shaughnessy, 1976; Toppino & Bloom, 2002; Verkoeijen,

Table 3 Major Spacing Phenomena.

1. *Intention invariance*. Spacing effects emerge with both incidental and intentional learning, using a wide range of materials.
2. *Age invariance*. Children (including infants), young adults, and older adults all show the spacing effect.
3. *Species invariance*. Everything from marine mollusks (Carew, Pinsker, & Kandel, 1972) to honeybees (Menzel et al., 2001) to mice (Scharf et al., 2002) shows spacing effects of some sort.
4. *The Glenberg surface*. The effect of lag is jointly determined by retention interval and type of test. Typically, the relationship between memory and lag is U-shaped, with the peak of the U-curve moving further to the right as the retention interval increases.
5. *Manipulating contextual variability seldom helps recall*. There are numerous failures to get multiple retrieval routes to help recall compared to a single repeated retrieval route.
6. *Recognition is required*. Items people fail to recognize on later repetitions show little or no spacing benefit.
7. *Perceptual priming effects*. A priming account might handle material that is not semantically coded, like faces and nonwords, but it can't handle semantic information.

Rikers, & Schmidt, 2005). All of these experiments used mixed lists, so it is not clear from the literature whether spacing effects emerge following incidental learning on pure lists or not. Each of them is therefore vulnerable to a list-strength effect critique.

Additionally, some of the experiments demonstrating incidental-learning effects may be vulnerable to deficient-processing explanations. For example, Greene (1989) wrote that "... asking subjects to make the same response to an item every time it occurs... may lead the subject to base the response to a second occurrence on memory for the response to the first occurrence." Indeed, Jensen and Freund (1981) conducted two experiments in which they compared incidentally-learned lists containing either a single semantic judgment (done twice) or two different semantic judgments (done once each). The lists were mixed with respect to spacing and massing, and also included once-presented items. In both studies, mixing encoding strategies lowered subsequent free recall of once-presented and spaced items relative to using only a single dimension. However, mixing encoding strategies actually *helped* massed items. Very similar results were obtained with children in the first, third, and sixth grades by Toppino and DeMesquita (1984). Such results suggest a possible switch cost for using two different encoding strategies, but that massed items likely suffered from a processing deficit when rated twice on the same dimension. In other words, there was a deficient-processing effect for incidentally processed items when they were rated twice on the same dimension.

There have been some attempts to argue that some incidental-learning instructions might encourage rehearsal-like processes that favor spaced items by forcing retrieval of earlier items. If people make ratings by comparing the current item to previously encountered items, for example, they would have to retrieve earlier-presented items. Because spaced items occur in more places on the list, they are more likely to be recent at any given time, and therefore may be differentially often used as the basis of comparisons, thus strengthening them. Of course, there is absolutely no empirical evidence to support this, but the study is easy enough to conduct—simply ask participants to do a rating task and ask them to report whether they make their judgment by comparing the item to another word (and if so, which one), or if they are rating it without reference to any other items. Having tried the task out on ourselves, we suspect the latter is more common, but it may vary depending on the difficulty of the rating task such that more sensitive scales (e.g., 1–9) may result in more covert retrieval than less sensitive scales (e.g., yes/no). Incidental-learning tasks that encourage people to look for a rule in a sequence may be the most likely to show covert rehearsal effects (e.g., Greene, 1989; Paivio, 1974), as the task requires comparison across items.

In sum, it would be nice if there were a clearer demonstration of incidental-learning effects that could not be attributed to any of the impostors outlined in Section 2. While the balance of evidence seems to suggest

there is a ''true'' spacing effect in incidental learning, there is as yet no convincing demonstration using pure lists. However, in an unpublished study, Delaney and Verkoeijen asked 85 participants to view two lists of 32 medium-frequency nouns. Each word was repeated twice for 2 s on each presentation, with a 1-s interstimulus interval. The design of the study was 3 Lag (massed, spaced lag 2, and spaced lag 12) × 2 Intentionality (incidental vs. intentional) design, with intentionality manipulated within-subjects and lag manipulated between subjects. That is, every participant saw two pure lists, one of which was learned incidentally and the other intentionally. The order of these lists was counterbalanced so that half of the people received intentional instructions first and the other half received incidental instructions first.

The incidental-learning instructions told participants to indicate for each word either (a) whether it was man-made or not, if they saw an ''mm'' symbol; or (b) whether it was pleasant or not, if they saw a '';-)'' symbol. They always received one of the instructions on the first presentation and the other on the second presentation. The intentional-learning instructions told them to rehearse the words aloud in order to learn the list. At the end of each list, there was a free recall test. Participants gave no indication that they expected the test, but of course it is always possible that they expected it. To summarize the results, there was a spacing effect in the incidental condition, but not in the intentional condition. Figure 2 shows the pattern of recall. Consistent with our other studies (Delaney & Knowles, 2005; Delaney &

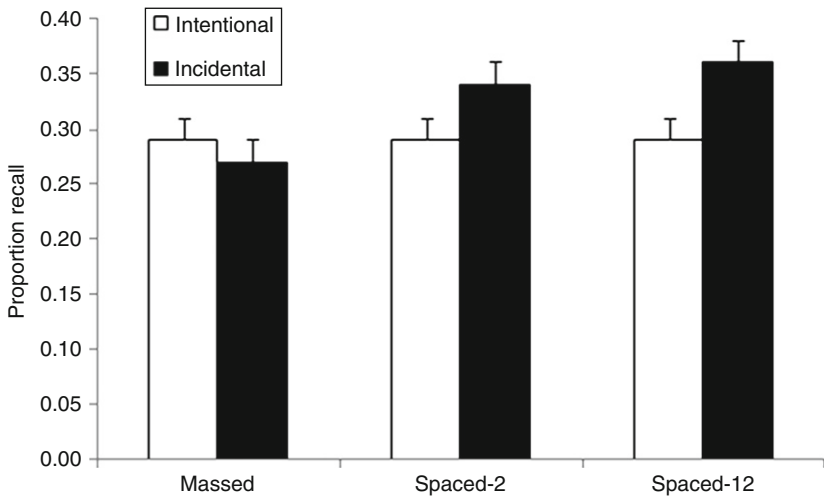

Figure 2 Proportion recall as a function of the lag between repetitions on pure lists for lists learned either via rote rehearsal (intentional) or incidentally. From an unpublished study by Delaney and Verkoeijen.

Verkoeijen, 2009), pure lists with instructions to study via rehearsal produced no overall spacing effects. In contrast, incidental learning produced a significant spacing effect, although the lag effect was not significant ($F < 1$). Therefore, it seems that spacing effects do emerge on pure lists when studied incidentally, though we did not obtain a significant lag effect.

3.2. Age-Invariance

A second clue that rehearsal cannot fully explain the spacing effect is that it occurs throughout the lifespan, even in children too young to rehearse. There are now numerous studies showing that the spacing effect emerges in children, using both recognition (Cahill & Toppino, 1993; Toppino, Kasserman, & Mracek, 1991; Vlach, Sandhofer, & Kornell, 2008) and free recall (Seabrook, Brown, & Solity, 2005; Toppino, 1993; Toppino & DeMesquita, 1984; Toppino & DiGeorge, 1984; Wilson, 1976). It persists over 48 h, at least in recognition (Cahill & Toppino). Although one study failed to obtain the effect with preschoolers (Toppino & DiGeorge, 1984), many later studies obtained it with preschool-age children (e.g., Rea & Modigliani, 1987; Toppino, 1991, 1993; Toppino et al.). Furthermore, the effect occurs with spacing lags up to 1 day for autobiographical events (Price, Connolly, & Gordon, 2006). These studies are important in part because preschool children are too young to implement a rehearsal strategy, and therefore the results cannot be attributed to rehearsal biases.

Even infants show the spacing effect. Using habituation, Cornell (1980) showed babies a photo four times, with the repeated exposures spaced either "massed-like" with 3 s between viewings, or "spaced-like" with 60 s between viewings. The baby would then see the same photo again, along with a novel photo. Because babies usually like to look at novel things, they would be expected to spend less time looking at the previously seen photo if they remembered it better. In fact, babies looked longer at the massed-like photos than they did at the spaced-like photos, suggesting they had better memory for the spaced-like photos. This was true when the delay until the test was 1 min, 5 min, or 1 h. (An added advantage of the infants design is that it is not vulnerable to a list-strength effect interpretation.) Habituation is probably mediated by a kind of perceptual priming, suggesting that perceptual priming may be important for the spacing effect, especially with nonsemantic materials—a point we will return to later.

Another infant study used operant conditioning of a foot kick in response to a toy mobile in 8-week-old infants (Vander Linde, Morrongiello, & Rovee-Collier, 1985). On a final test two weeks later, the response was retained better when 18 min of training were split into three sessions separated by 1 or 2 days compared to 18 min on a single day. The effect was quite large, with 48-h spacing resulting in an average of 25 kicks on the final test as compared to only 15 for massed study. As operant

conditioning relies on motor responses, it is unlikely to be due only to perceptual priming.

What about older adults? Perhaps unsurprisingly, older adults show spacing effects roughly comparable to those of young adults (Balota, Duchek, & Paullin, 1989; Kausler, Wiley, & Phillips, 1990). Benjamin and Craik (2001) found that for both older and younger adults, spacing made it easier to discriminate studied from unstudied items than massing did. However, two lists were studied and the task was to respond only to the items from one of the lists—that is, when a source judgment was required—older adults were more likely to mistakenly endorse items from the wrong list. Younger adults showed no such trend. The study suggests that while item memory is improved with spacing in both older and younger adults, older adults do not show a spacing effect for source memory.

In sum, the spacing effect seems to emerge throughout the lifespan and with many types of materials, which suggests that simple strategic explanations are insufficient to account for the results. Results like these suggest that very basic neural phenomena could be involved in producing the spacing effect.

3.3. Species Invariance

A further piece of evidence that spacing effects might arise from basic memory processes comes from comparative psychological studies. One interesting study by Menzel, Manz, Menzel, and Greggers (2001), for example, used classical conditioning procedures to condition honeybees to extend the proboscis (in response to various stimuli such as carnations, propionic acid, and hexanol). They varied the spacing between acquisition trials to produce massed trials (< 30 s between trials) and spaced trials (3, 10, 20, or 30 min between trials). Spacing sped acquisition of the conditioned response in the honeybees, as opposed to the slowing of acquisition that spacing produces during learning in human beings (e.g., Schmidt & Bjork, 1992). They further varied the retention interval, with intervals ranging from relatively short (30 min) out to several days. The results suggested that spacing advantages on a final test were absent at very short retention intervals, but after 3 days the advantage of spacing was pronounced, with massed trials mostly forgotten and spaced trials showing memory rates similar to the end of the acquisition period.

An advantage of using the honeybees is that memory based on protein synthesis in the honeybee develops rather slowly, and the time-course is well known. Blocking protein synthesis did not affect acquisition, but it prevented the spacing effect from emerging on the final test after a delay. Specifically, after 1–2 days' retention, blocking protein synthesis harmed spaced but not massed retention, dropping spaced recall to the level of massed recall. After 3–4 days, it harmed both spaced and massed retention, with both spaced and massed recall dropping to a low level. The results

suggest that for honeybees, the spacing advantage during acquisition is independent of protein synthesis, but that to display a spacing advantage over longer periods of time requires consolidation processes. Similar results have been obtained with other organisms, including other insects like drosophila (flies) and mice—the latter of which, like humans, have a hippocampus (DeZazzo & Tully, 1995; Scharf et al., 2002).

Results such as these suggest that neural consolidation processes might be involved in the advantage of spaced memories, and that these consolidation effects might occur over relatively long periods of time. Forms of consolidation theories were proposed early, notably by Landauer (1969), who proposed that when an item is repeated, the second repetition needs to be delayed sufficiently to allow for consolidation of the first response before additional learning benefits can be seen. These early consolidation theories were rejected based mainly on a celebrated study by Bjork and Allen (1970), who presented a word triplet, then repeated it either following an ''easy'' distractor task or a ''hard'' distractor task. After the second presentation, a filler task was followed by a recall prompt. Contrary to consolidation accounts, the harder task did not impair the consolidation of the first trace; in fact, harder tasks *improved* the benefit of restudy.

However, consolidation theories need not assume that the second presentation disrupts consolidation of the first. The honeybee studies, for example, suggest that protein synthesis in the brain occurs because the spaced presentations show superior consolidation when the same neurons are repeatedly activated, suggesting the second and subsequent presentations would show slower forgetting over time—a result consistent with several mathematical models of the spacing effect (Pavlik & Anderson, 2005; Reed, 1977).

3.4. The Glenberg Surface

One of the most important discoveries in the spacing effect was that spacing is not an all-or-none proposition. Melton (1967) showed that there is also a lag effect such that the longer the spacing is between repetitions, the more memory is helped. Peterson, Wampler, Kirkpatrick, and Saltzman (1963) modified Melton's conclusions, showing that the relationship between lag and memory was actually an inverted U-shape, such that longer lags initially improve recall up to a local maximum, with longer lags than that resulting in lower recall. The U-curve was replicated by a number of early spacing researchers, including Atkinson and Brelsford (cited in Atkinson & Shiffrin, 1968) and Young (1971). Glenberg (1976, 1977, 1979) subsequently demonstrated that the retention interval between the final study episode and the test was also important. Specifically, he demonstrated that the peak point of the U-curve is proportional to the retention interval, with longer retention intervals yielding longer optimal spacings.

One further wrinkle to this story is that Glenberg (1976) argued that the U-curve occurs only for cued recall, not for free recall. However, he may just have used too short of a lag. Verkoeijen et al. (2005) showed the U-curve in free recall rather clearly, using both intentional and incidental learning. Toppino and Bloom (2002) used free recall and found that the peak of the U-curve depended not on the number of items that intervened between repetitions, but rather on the *time* between the two repetitions. Using different presentation rates and different lags between repetitions, they were able to show that some manipulations that apparently eliminated the spacing effect were just far enough along the U-curve that they produced negligible spacing benefits. Thus, it seems that the U-curve holds for both free recall and for cued recall. A major meta-analytic review by Cepeda et al. (2006) concluded that the optimal spacing typically was around 10–20% of the retention interval (see also Pashler, Rohrer, Cepeda, & Carpenter, 2007). The optimal spacing calculations, however, depend on studies where rehearsal borrowing was not controlled, and therefore may be inaccurate in some circumstances.

3.5. Deliberate Contextual Variability at the Item Level Doesn't Help

To be considered successful, any theory of spacing needs to be able to reproduce the Glenberg surface, which therefore places a major constraint on theory. One of the earliest attempts to produce the Glenberg surface was contextual variability theory—in fact, it is also the theory Glenberg himself championed. Melton (1967) was the first to propose that contextual variability could account for the lag effect. The basic idea behind contextual variability accounts is that spaced items occur in multiple contexts, and therefore have more retrieval routes by which they can be later accessed than massed items do. (The assumption that events are encoded with respect to some background context, and that the context at test is used as one of the retrieval cues to help aid recall has a long history in psychology, and underlies many modern memory models.) At short lags, context does not vary much between repetitions, producing overlapping contexts at study (and hence less resistance to forgetting). However, compared to massed items, even the contextual variability between repetitions in close proximity would still provide additional retrieval routes, favoring retrieval of spaced items.

Producing the downward slope of the U-curve is more challenging. Glenberg (1976) favored an explanation in terms of the match between the test context and the study contexts. He reasoned that the study context of the first repetition would, for very long lags, mismatch the test context too

much, resulting in complete reliance on the second repetition at recall. Hence, the spacing effect should become small at long retention intervals.

There were a number of reasons why people abandoned the original contextual variability accounts, but chief among them was a paper by Ross and Landauer (1978) which reasoned that if contextual variability increases the likelihood of retrieving two repetitions of the same item, it should also help with recalling at least one of two different items. That is, if two words appear at a longer distance from one another, then recall of at least one of those words should be higher than if they are nearby one another, because there are multiple contextual routes to retrieve the items. This turns out not to be the case, which is a problem for the classic version of contextual variability.

In addition, attempts to deliberately induce contextual variability had mixed—but generally negative—results. Early attempts to test encoding variability were based on early notions that the semantic connotation of words was biased by its neighbors, and that therefore it was encoded in different ways at different places on the list (Madigan, 1969). Therefore, researchers quite reasonably used homographs, which are words with multiple meanings, to create maximally-different connotations for words. As an example, Johnston, Coots, and Flickinger (1972) presented homographs twice together with a cue word that was supposedly to help them remember the words. For some of these, the word was deliberately chosen to bias one or the other meaning of the word (e.g., river-BANK, money-BANK) or was seemingly unrelated to the meaning (e.g., dog-BANK, spoon-BANK). They also manipulated whether people saw the same cue on both repetitions or a different biasing cue, and the lag between the repetitions. Encoding variability theory should predict that at wider spacings, different cues should result in better memory than the same cues, but that is not what Johnston et al. found. Figure 3 shows their results, which represent a fairly common pattern in the spacing literature. For biasing cues, changing the cue helped on massed items but had no impact on spaced items. For neutral cues, massed items were unaffected, but spaced items were better recalled when the same cue was present on each occasion than when different cues were presented. These results are clearly problematic for classic encoding variability theories. Other similar studies using homographs produced similar results (Bobrow, 1970; D'Agostino & DeRemer, 1973; Hintzman, Summers, & Block, 1975; Madigan; Thios, 1972), although there is one aberrant paper that might be worth trying to replicate (Gartman & Johnson, 1972).

As a quick aside, Johnston et al. (1972) also included pure lists of once-presented items as a control. They made a big deal about failing to obtain a list-strength effect on their once-presented items. However, extracting means from their data, once-presented items were recalled numerically less often on mixed lists containing some twice-presented items (27%)

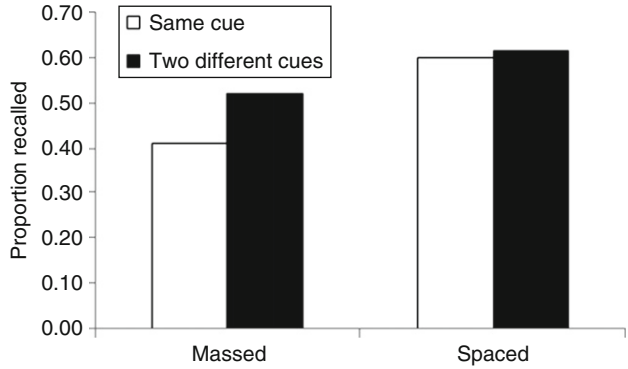

Figure 3 Proportion free recall of capitalized words after viewing the words twice, either with the same biased cue each time (sports-FAN, sports-FAN) or different biasing cues each time (sports-FAN, electric-FAN) for massed or spaced (lag 3 or 7) repetitions. Adapted from Johnston et al. (1972).

than on the pure lists containing only other once-presented items (33%). This is a 6% list-strength effect, which is as large as the spacing effect in experiments where encoding strategy is controlled (e.g., Delaney & Knowles, 2005). It may not be reliable statistically, but hardly provides a strong failure to replicate Tulving and Hastie's (1972) original demonstration of the list-strength effect.

Later authors argued that homographs provided a poor test of the encoding variability hypothesis because the two meanings of the homographs were, in many ways, two different words entirely (e.g., Hintzman, 1974; Maskarinec & Thompson, 1976). Therefore, a second round of encoding variability tests tried to keep the items and their semantic meanings the same, but varied whether the cues presented together with those items were the same or different. For example, in Experiment 1 of their oft-cited paper, Postman and Knecht (1983) presented people with a list of words embedded in sentences. Participants were told that they were supposed to remember the words for a later memory test. The lists were presented three times, so that each sentence was shown three times for a total presentation time of 15 s. Participants either saw three different short sentences using the word, or the same sentence using the word three times. Finally, participants either had a free recall test on the words, or they were tested using the cue sentences they had studied. Contrary to encoding variability accounts, the different "contexts" created by the multiple sentences did not produce better free recall, and actually produced *worse* memory on cued-recall tests. If having multiple retrieval routes should help recall, then why did it not?

Their subsequent experiments were similar. Experiment 2 used incidental learning and included a test either immediately or after a 24-h delay.

The results were essentially identical to Experiment 1, with variable encoding producing either equivalent or worse memory. Experiment 3 attempted to use three different instances of a noun (e.g., wine bottle, medicine bottle, and thermos bottle) or a single repeated instance of the noun and obtained results identical to Experiment 2.

So much for deliberate attempts to manipulate the retrieval cues. Another approach was to vary intrinsic properties of the stimulus. One attempt was to vary the language used in the experiment. For example, an early study by Kolers (1966) used French–English bilinguals and presented words either once or several times in a spaced fashion. On some lists, words were repeated in the same language each time, while on other lists, they were repeated in each language (French and English). There were also some lists that had some items repeated in the same and some in a different language. The results suggested that there was no difference between repeating an item in the same language or in a different language, which is contrary to the predictions of encoding variability theory. A later study by Glanzer and Duarte (1971) tried something similar using Spanish–English bilinguals; they presented word pairs twice either in English, Spanish, or once in each language. Statistically, the results supported the simple interpretation that changing the language helped massed items, but had less and less effect at longer lags. However, their data in fact show that the probability of recall for a pair presented in each language (0.56) was roughly equivalent to pairs presented twice in Spanish (0.57). However, two repetitions in English showed very poor recall (0.44). Hence, their result may be an artifact of poor recollection of English-only pairs, or a result of output interference driven by recalling English-only pairs last. To our knowledge, only the massed portion of the Glanzer and Duarte (1971) study has ever been replicated (by Durgunoğlu & Roediger, 1987, who further showed similar results with yes/no recognition—that is, better recognition of mixed-language pairs than same-language pairs). It would be interesting to see whether the spaced pattern emerged in both free recall and recognition under more controlled conditions, or on unmixed lists where output interference could be ruled out as an explanation. Nonetheless, these studies suggest that encoding variability helped massed items much more than it helped spaced items, if the results are in fact replicable.

In sum, the encoding variability theory was originally proposed to account for the Glenberg surface. However, by the mid-1970s, the evidence for encoding variability was looking pretty grim, at least for within-list contextual variation. Many studies attempted to vary the local context around an item and found that contextual variation either did not help—or even hurt—memory. Furthermore, Ross and Landauer (1978) had earlier argued that most versions of the encoding variability account implied that two different items at two places on the list should also show a memory benefit, which they do not. We will later see that a version of encoding

variability theory can be proposed that is generally consistent with the results outlined here, but not without introducing some different assumptions about the nature of context.

3.6. Recognition Required for Spacing Benefits

The next class of theories to survive into the modern day involves study-phase retrieval, a fancy term for "recognizing that something is repeated when you see it." Hintzman and Block (1973) first proposed the study-phase retrieval account, which assumes that in order to obtain the spacing effect, people must retrieve the prior presentation. Study-phase retrieval provides a straightforward explanation for the lag U-curve, because as the lag between repetitions increases, so does the difficulty of the retrieval. Harder retrievals are thought to result in more strengthening of the original trace. Furthermore, as the retrieval becomes too difficult, the probability of successful retrieval on the second presentation will drop. If the previous repetition is not retrieved, then there is no benefit of spacing, producing the downward slope of the U-curve at long lags.

The earliest evidence supporting study-phase retrieval was that when people are asked to indicate how far apart two repetitions of an item occurred, their estimates track the actual distance between presentations (Hintzman & Block, 1973; Hintzman et al., 1975). The same turns out to be true for related items and homographs (Hintzman et al.), even though for two unrelated items, people are essentially at chance. This suggests that people noticed the related items during study, and tagged how far apart they were.

An important study by Johnston and Uhl (1976, Experiment 2) tested some of the predictions of study-phase retrieval theory. They used a continuous recognition paradigm, whereby participants had to attempt to recognize whether a word had been seen earlier in the list. As the lag between repetitions increased, the probability of successful recognition of an item as "old" went down slightly (though it was still 91% at a lag of 13). Furthermore, the spacing benefit was observed on a final free recall test only for items that were successfully recognized as "old" during the initial study phase. There are interpretations of this study that do not require study-phase retrieval in ordinary spacing studies, however. The introduction of the continuous recognition procedure makes this a testing effect study rather than a spacing study, and we will later see that testing only benefits items that are successfully retrieved during the test (e.g., Carpenter & DeLosh, 2005). Therefore, the study may not tell us much about "normal" study-phase retrieval, which may not happen spontaneously.

A clever experiment in the same vein by Braun and Rubin (1998, Experiment 3) noted that a strict version of the study-phase retrieval account suggests that it is the first presentation of the item that benefits

from study-phase retrieval, not the second presentation. Therefore, if there were a way to differentiate the first and second presentation, one might be able to demonstrate study-phase retrieval effects directly. To get at this issue, they used the same continuous recognition paradigm as Johnston and Uhl (1976), but with a twist—instead of exact repetitions, people looked for words that had the same first three letters (e.g., BURden, BURlap). As with Johnson and Uhl's study, they found that the longer the lag, the lower the probability of successfully recognizing that a word was a repetition during the study phase. On a final recall test, participants received the stem (e.g., BUR) and two blanks. They were instructed to write whatever words they remembered and to fill both blanks if they recalled two words with that stem. This procedure allowed Braun and Rubin to distinguish whether it was the first or second presentation that benefitted from spacing, or both. On the one hand, they found that the first word with a given stem was better recalled than the second, consistent with study-phase retrieval explanations. On the other hand, they also found that both the first and second presentation showed a spacing benefit. This latter result could be due to output facilitation, such that recalling the first member of the pair at test facilitates recall of the second member of the pair. However, if so, one would expect that if the cued-recall test were replaced with a recognition test, then the first presentation would lose its advantage. Instead, the same pattern emerged with recognition testing—spacing effects on both repetitions, with better recognition of the first than second presentation. Therefore, their results are puzzling if one assumes that spacing effects are entirely due to study-phase retrieval. Furthermore, even for massed items they obtained better memory for the first than for the second item in the pair. It would have been nice to know whether this result was due to their continuous recognition procedure (which makes this study a testing effect study), or if the same results would hold if people were not explicitly asked to watch for repetitions.

A particularly nice study by Sahakyan and Goodmon (2007) took advantage of the extensive research on what words remind people of what other words (e.g., Nelson, McKinney, Gee, & Janczura, 1998). In their studies, they created lists of words that were unidirectionally-related—that is, one of the words automatically reminds people of the other, but not vice-versa. They then created lists where List 2 and List 1 deliberately had this unidirectional relationship. When List 2 words reminded participants of List 1 words, the List 1 words benefitted on a later free recall test, consistent with a study-phase retrieval effect. However, when the List 1 words reminded people of List 2, memory was no different than in control conditions where List 1 and List 2 words were unrelated. There was one exception to the latter rule—sometimes, they observed output facilitation such that at the time of recall, the List 1 words were output first and then reminded people of the semantically related List 2 words. When output

order was controlled to force output of List 2 first, this effect disappeared. Output facilitation cannot happen with "real" spaced repetitions except on frequency judgment tests, since once you recall a word, it does not matter if it reminds you of its other presentation; you already output that word. These results are consistent with a study-phase retrieval account, and one of the nice things about this study is that there was no instruction to retrieve, showing that people did it spontaneously.

The study-phase retrieval account provides some other nice predictions regarding the difficulty of recognizing items. Items that are harder to recognize should have a shorter optimal lag than items that are easier to recognize. It can therefore explain rather complicated patterns of data on the interaction between different types of items, lag, and retention interval. A celebrated study by Paivio (1974) used words and easily named pictures, the latter of which typically show a memory advantage over the former. In Experiment 1, the lists contained items presented once and items presented twice with a rather long lag (48 items), with all combinations of pictures and their corresponding words (i.e., repeated word, repeated picture, word then the corresponding picture, and picture then the corresponding word). Once-presented items were mixed with twice-presented items, but spaced and massed items did not appear on the same list. Furthermore, Paivio included an incidental-learning condition in which participants were unaware they would be tested and had to predict whether the next item would be a picture or a word (there was no true rhyme or reason to the order). For repeated pictures and repeated words, he obtained a spacing effect, but for repetitions where the type of the item changed, he obtained no spacing effects. In Experiment 2, he compared once-presented items with twice-presented items at varying lags (massed, spaced/24, and spaced/48). In this experiment, he obtained different results for intentional and incidental learning. For twice-presented words, he obtained no spacing effect on incidental but obtained it with intentional. For twice-presented pictures, he obtained spacing effects for both intentional and incidental learning. Finally, for the mixed-type items, he received no spacing effects with incidental learning and a U-shaped curve for intentional learning. If we were to summarize these results using the study-phase retrieval account, they show that the more difficult it is to recognize the previous presentation, the shorter the lag should be to obtain a significant spacing effect. Pictures are better recalled than words, so they should be recognized at a longer lag than words are, producing spacing effects even at long lags. Intentional learning leads to rehearsal, which also strengthens memorability of the items and leads to spacing effects at longer lags. Finally, mixing the type of item on the repetition tends to reduce the chance that people detect a repetition as well.

Another set of results that only the study-phase retrieval account can handle involve inhibited items. For example, in retrieval-induced

forgetting, people study a list of category–exemplar pairs, such as fruit–lemon or profession–scientist. Subsequently, some of the items receive retrieval practice, while others do not. The result is that on a final test, memory for the practiced items improves, while memory for the unpracticed items from the same category suffers (e.g., Anderson, Bjork, & Bjork, 1994). This inhibitory process provides a way to weaken some items below a once-studied baseline. An important recent study by Storm, Bjork, and Bjork (2008) used retrieval-induced forgetting to weaken some studied items. When the inhibited items were then presented for restudy, they subsequently showed better memory than comparable items that had been studied one time but not inhibited. This result is difficult to explain until one realizes that the difficulty of the retrieval predicts the benefit of a restudy trial in study-phase retrieval accounts. Therefore, weakening items below once-presented items can nonetheless allow for more difficult retrieval later on, thus strengthening them.

In sum, the study-phase retrieval account had a number of advantages over earlier accounts. It could explain why recognition was required for a spacing benefit to emerge, why inhibited items show bigger spacing benefits than noninhibited items, and why difficult-to-learn material might result in shorter optimal lags. It therefore persists as one of the major theories of spacing to this day.

3.7. Semantic and Perceptual Priming Accounts for Cued-Memory Tasks

In the 1990s, a family of priming-based accounts of spacing in cued recall and recognition emerged. They are technically deficient-processing explanations (similar to the impostors seen in Section 2), but are assumed to provide the explanation of the ''real'' spacing effect, at least for cued recall and recognition. Our view is that important findings from the spacing effect literature are difficult to reconcile with the priming account, but it is worth reviewing in detail anyway for two reasons. First, intriguing findings with nonverbal materials provide constraints on theories of spacing effects. Second, no one has yet done a major narrative review of this literature, as it is relatively new.

Challis (1993) proposed the priming account, which suggests that semantic processing is critical to obtaining spacing effects. According to his account, the semantic representation of an item is activated by its first occurrence, and it remains active for a short period. During massed repetitions, the second repetition of the item occurs while the semantic representation of the item's first presentation is still activated. During spaced repetitions, enough time has passed that the semantic representation has partially deactivated. Consequently, less total semantic processing will be devoted to later occurrences of the massed items than later occurrences of

spaced items, producing spacing effects. This explanation is based on the finding that semantic-associative priming is a short-lasting phenomenon, usually obtained when the prime immediately precedes the target, but not when more than one item intervenes between the prime and target words (Bentin & Feldman, 1990; Dannenbring & Briand, 1982; Kirsner, Smith, Lockhart, King, & Jain, 1984; McNamara, 1992).

To test the semantic priming account, Challis (1993) designed two experiments using mixed lists of massed and spaced words. Participants were either instructed to use various orienting tasks that encouraged processing the words at a semantic level (e.g., rate them for pleasantness) or at a graphemic level (e.g., count the number of descending letters such as *g* and *y*). After viewing the list, they received a frequency-judgment test (Experiment 1) or a cued-recall test (Experiment 2). Both experiments confirmed Challis' predictions—the spacing effect was present only with semantic and not with graphemic encoding manipulations.

The semantic priming account's central prediction is that priming should be higher for massed repetitions than for spaced repetitions, and consequently massed repetitions should be less accessible than spaced repetitions on a subsequent cued-memory test. To address this issue, Rose (1984, Experiment 2) instructed participants to answer semantic questions about words repeated at various lags. Half of the participants received the same question on all three repetitions, while the rest received a different question on each repetition. After the list, they received a surprise old/new recognition test (as well as other tests). Crucially, when a different question was presented on each repetition, answer time was unrelated to lag. In contrast, when the same question was used on each repetition, answer times were faster at shorter lags than at longer lags. Massed items were sped up the most by moving from different questions to the same question. Furthermore, the recognition data showed a spacing effect in the same-question condition, but not in the different-question condition. Although measures of semantic priming were not calculated in Rose's second experiment, the combination of reaction times data and recognition data are consistent with the semantic priming account. Specifically, the semantic priming account predicts—in line with the experiment—that when semantic priming is reduced by asking different semantic questions about the occurrences of a repetition, the magnitude of the spacing effect should be reduced.

In another study, Wagner, Maril, and Schacter (2000) tested two central predictions of the semantic priming account. Participants in an fMRI scanner did incidental semantic processing of massed and spaced words, followed by a final old/new recognition test. One prediction of the semantic priming account is that memory should be better for spaced repetitions than for massed repetitions (i.e., a spacing effect), whereas semantic priming should be lower for spaced repetitions than for massed repetitions—a result that Wagner et al. obtained. The semantic priming account further predicts

that within massed and spaced repetition, the magnitude of a participant's semantic priming effect should be inversely correlated with their memory performance. For spaced repetitions, Wagner and colleagues obtained the predicted negative correlation between mean priming level and memory performance. By contrast, within massed repetitions, no reliable correlation between priming and memory performance was found—a finding that is at variance with the semantic priming account.

However, some aspects of the Wagner et al. (2000) study were suboptimal, and therefore it is difficult to draw firm conclusions about the semantic priming account on the basis of the reported outcomes. To begin, and as acknowledged by Wagner and colleagues, the inconsistent findings regarding the within-repetition type correlations may have been due to the small sample size ($n = 12$). Furthermore, priming scores are difference scores and such scores are known to be less reliable than the constituent scores. Consequently, the power of statistical analyses that involve difference scores is often lower than analyses of the single constituent scores. Considering both of these factors, their analyses could have missed even medium-sized correlations.

Furthermore, Wagner and colleagues did not attempt to control or assess baseline memory performance in their sample, which may affect priming scores. Baseline reaction times may either be positively (higher reaction times indicate a better task focus) or negatively (higher reaction times indicate lack of motivation) related to memory performance. In the former situation, the resulting positive bivariate correlation between priming and memory performance will be incorrectly interpreted as evidence against the semantic priming account, whereas in the latter situation the resulting negative bivariate correlation will be incorrectly interpreted as evidence in favor of the semantic priming account. Therefore, when assessing the relationship between priming and memory, it is important to control for baseline effects. In sum, the correlations between priming and memory performance within repetition type are difficult to interpret.

3.7.1. Difficulties for the Semantic Priming Account: Related Items

A central prediction of the semantic priming account is that spacing effects should emerge not only for repetitions but also for semantically related words. There appears to be no reason why the influence of semantic priming on memory performance should differ between repetitions and associated word pairs. However, a study by Hintzman et al. (1975) suggested that spacing effects for repetition pairs and associated pairs differ tremendously (see Greene, 1990 and Stern & Hintzman, 1979, for similar studies using synonyms as stimulus materials). Hintzman et al. did not design their experiment to test the semantic priming account; the target items in the study list were organized such that pairs were associated in a *backward* direction. That is, the second presented word of a pair evoked the first

word, but the first word did not evoke the second word. Given that semantic priming (for a review see Neely, 1991) is assumed to operate in a forward direction (i.e., from the first word to the second word), the demonstrated reversed spacing effect cannot be taken as evidence against the semantic priming explanation.

Verkoeijen, Rikers, Pecher, Zeelenberg, and Schmidt (2010) conducted a study to address this issue. In their Experiment 2, lists containing either massed and spaced word pairs with forward associations (e.g., fork–knife) or massed and spaced repetitions (e.g., knife–knife) were shown. They were encoded incidentally using a semantic yes/no judgment as to whether they would fit into a big box (the experiment was conducted in Dutch and "big box" refers to a specific size). At the end, a yes/no recognition test was administered. Semantic priming, as assessed by reaction times during the study phase, was larger for massed pairs than for spaced pairs both in the associated-pairs condition and in the repetition condition. In addition, the recognition data showed a *reversed* spacing effect in the associated-pairs condition, whereas a standard spacing effect was revealed in the repetition condition. The finding—observed in the associated-pairs condition—that a larger priming effect was associated with a better memory performance was interpreted as evidence against the semantic priming account of spacing effects in cued-memory tasks.

Some researchers (e.g., Mammarella, Russo, & Avons, 2002) have argued that Challis (1993) was referring to semantic *repetition* priming rather than to semantic *associative* priming, which would render results using associated pairs irrelevant. To us, however, Challis' account seems more consistent with an associative-semantic priming interpretation than with a semantic repetition priming interpretation. After all, associative priming is a short-lived phenomenon and repetition priming is not (e.g., Dannenbring & Briand, 1982; Kirsner et al., 1984; Zeelenberg & Pecher, 2002). Repetition priming can endure a retention interval of several days (e.g., Jacoby, 1983), whereas associative priming disappears when one or two items intervene between the prime and the target (e.g., Dannenbring & Briand; McNamara, 1992). In addition, several of the papers referred to by Challis suggest he had an associative priming explanation in mind when proposing his priming account as the experiments reported in these papers used semantically associated word pairs (e.g., Neely, 1977; Smith, Theodor, & Franklin, 1983). Thus, clearly Challis was not *excluding* the possibility that spacing effects were due to semantic-associative priming.

Arguably, even a semantic repetition account has trouble accounting for the Verkoeijen et al. (2010) study's results. According to the semantic repetition priming account, the spacing effect emerges because priming of semantic features is stronger for massed repetitions than for spaced repetitions. When associated pairs are used, priming of semantic features will also take place, albeit to a lesser extent than with repetitions. So, if priming of semantic features is

indeed underlying the spacing effect, a straightforward prediction seems to be that the spacing effect is smaller for associated pairs than for repetitions. Verkoeijen et al.'s inverse spacing effect for associated pairs runs counter to this prediction. The question that remains then is why priming of semantic features produces a spacing effect in repetitions and an inverse spacing effect in associated pairs.

Furthermore, the semantic priming account has difficulties with the results of a study by Peterson, Hillner, and Saltzman (1962; see also Peterson et al., 1963, Experiment 3), who presented a list of paired associates containing some massed repetitions and some spaced repetitions, with the latter 8 s apart. Retention interval was varied from 2 to 16 s. The semantic priming account dictates that the spacing effect emerges during study due to deficient processing of massed repetitions as compared to spaced. Furthermore, according to the semantic priming account there is no reason to expect that the spacing effect interacts with the retention interval. However, Peterson and colleagues obtained a reverse spacing effect at retention intervals of 2–4 s, and a regular spacing effect at retention intervals of 8–16 s—a pattern of results that clearly contradicts the semantic priming account.

The semantic priming account also applies only to a restricted range of lags and retention intervals, given the relatively short duration of priming. It cannot predict the nonmonotonic relationships between spacing interval and retention interval which are observed in cued-recall tasks at very long delays (see Cepeda et al., 2009; 2006; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008), because it predicts that memory performance should increase with spacing until the interval at which priming no longer occurs, after which memory performance will remain at a constant, asymptotic level. Also, this predicted relationship between spacing and memory performance should be independent of the retention interval. (Of course, these longer-term spacing effects could rely on a completely different mechanism.)

3.7.2. Mounting Evidence for a Structural-Perceptual Priming Mechanism

Another serious problem with the semantic priming account is that it incorrectly predicts null spacing effects for complex nonverbal materials that are processed perceptually and not semantically (Russo, Parkin, Taylor, & Wilks, 1998). However, spacing effects are obtained with nonsense shapes (Cornoldi & Longoni, 1977) and unfamiliar faces (Parkin, Gardiner, & Rosser, 1995; Russo et al.). In defense of the semantic priming account, one might argue that participants use some kind of semantic processing mode to encode the pictorial targets. Russo et al. sought to rule out this alternative explanation. Russo et al. reasoned that semantic processing of faces will most probably not occur when participants perform

orienting tasks that focus on perceptual features of the faces. In their Experiment 3, participants studied massed and spaced unfamiliar faces while judging each on symmetry and length—the equivalent of a "graphemic" manipulation for faces. However, a robust spacing effect emerged, which is at odds with the predictions of Challis' (1993) theory.

To account for the results with complex perceptual materials, Russo and colleagues proposed a theoretical framework that distinguished two qualitatively different priming mechanisms. For semantically processed materials, the spacing effect in cued-memory tests was thought to be produced by Challis' (1993) semantic priming mechanism. Alternatively, for stimulus materials unlikely to be processed semantically, a structural–perceptual priming mechanism is assumed to underlie the spacing effect. The structural–perceptual priming explanation is conceptually analogous to the semantic priming mechanism, with the exception that the structural–perceptual mechanism operates at an item's orthographic, rather than at its semantic level of representation.

Several recent studies provide support for the structural–perceptual priming explanation of the spacing effect for nonsemantic materials. Russo, Mammarella, and Avons (2002) presented participants with mixed lists containing massed and spaced nonwords. On each repetition, participants performed two graphemic orienting tasks. On half of the trials, both repetitions were in the same font, whereas for the rest each occurrence was in a different font. This was expected to reduce perceptual priming, and hence reduce the spacing effect. Consistent with their predictions, there was a spacing effect for the same-font nonwords, but no spacing effect for the different-font nonwords. Furthermore, supporting their claim that structural–perceptual priming is a mechanism specific to the spacing effect for unfamiliar stimuli, when words were substituted for the nonwords in another experiment, a spacing effect emerged with both same and different fonts.

Their Experiment 3 provided the most convincing corroboration of the structural–perceptual priming mechanism. Participants performed a lexical decision task on words and nonwords that were presented once, repeated twice in a massed fashion, or repeated twice with lags of three or six items between the repetitions. As in their other experiments, the font was changed between the repetitions for half of the items and kept the same for the rest. Consistent with the priming account, repetition priming for words decreased as a function of inter-repetition lag, with a comparable decrease for words repeated in the same font as for words repeated in a different font. In contrast, for nonwords presented in the same font, repetition priming sharply declined as a function of lag. However, nonwords repeated in a different font showed much less priming decrement. The memory data closely reproduced the data from their earlier studies, and could be mapped on to the priming data. Taken together, the Russo and

colleagues experiments provide converging evidence that a structural–perceptual priming mechanism causes the spacing effect in cued-memory tests for unfamiliar materials.

Other studies partially replicated or extended the results of Russo et al. (2002). Russo and Mammarella (2002) asked participants to evaluate perceptual features of either words or nonwords that were repeated in a massed or a spaced fashion (see also Mammarella, Avons, & Russo, 2004). Subsequently, participants were given a yes–no recognition test on the studied items. Similar to the findings obtained by Russo and colleagues in the 2002 study, Russo and Mammarella found a spacing effect for nonwords but not for words.

Comparable studies with faces and ''nonfaces'' was conducted by Mammarella et al. (2002) with similar results. Their Experiment 1 provided, in our view, the strongest test of the structural-priming mechanisms. Faces and nonfaces were shown repeated in the same pose or in a different pose at three different inter-repetition lags: massed, lag 2, and lag 4. For each item, participants indicated whether it was a face or not. At the end of the list, there was an old/new recognition test. As in Russo et al.'s (2002) Experiment 3, changing the pose between two occurrences of a repetition reduced perceptual priming, and this decrease was larger for massed items than for spaced items. Also, the pose change eliminated the spacing effect in yes-no recognition.

In sum, with both nonwords (Russo et al., 2002), and with unfamiliar faces (Mammarella et al., 2002) manipulations that reduced structural repetition priming also eliminated the spacing effect. However, caution should be exerted when interpreting this correlation, as an unknown third variable might explain both the reduction of structural repetition priming and the reduction of the spacing effect.

3.7.3. Conclusions about Priming Mechanisms

The empirical evidence seems stacked against the notion that semantic priming explains spacing effects with cued-memory tests. Among other failures, the semantic priming account cannot explain why words that semantically prime later-seen words produce *better* final-test memory of the primed words, when the priming account predicts the opposite. The evidence against the perceptual priming explanation for meaningless materials is currently weaker. Indeed, there are some reasons to think it may be a real phenomenon, though the jury is still out. Even if the latter holds, it may be more properly classified as an ''impostor'' phenomenon that applies in very restricted circumstances, as it does not explain spacing effects with meaningful materials. Finally, there are competing accounts that may be able to explain the results from the meaningless materials more parsimoniously—a point we will consider in the next section of our review.

3.8. Hybrid Accounts

The modern trend has been to suggest that spacing effects cannot be explained by a unitary mechanism—a trend that we wholeheartedly endorsed in our lengthy exposé of impostor spacing effects in Section 2. An oft-cited paper by Greene (1989) laid out a major theory that proposed two separate processes are involved in spacing effects. He was one of the first to note that the spacing effect depended on the type of test used. Specifically, he distinguished between cued tests—which include cued recall, frequency judgments, and recognition (in recognition, the "cue" is the item itself)—and free recall. Greene never tested cued recall, so his "cued" tests were recognition and its cousin frequency judgment. For cued tests, he argued that rehearsal was extremely important because the spacing effect and the lag effect both occur under intentional-learning conditions, but not under incidental-learning conditions. Therefore, he reasoned, rehearsal processes were likely responsible for the spacing and lag effects in such cases. However, for free recall, he obtained spacing and lag effects regardless of whether intentional or incidental learning were used. He believed this represented study-phase retrieval effects.

Our work has shown that rehearsal is important in both recognition and free recall (Delaney & Verkoeijen, 2009) and contributes to both similarly. It is curious that Greene (1989) failed to obtain a spacing effect with recognition following incidental learning, given that spacing effects emerge when participants are encouraged to rehearse only the current item (e.g., Delaney & Verkoeijen). Later research suggested that it was the low level of semantic processing encouraged by his encoding instructions that may have eliminated the spacing effect, which is sometimes observed with incidental learning, provided semantic orienting tasks are employed (Challis, 1993; Greene & Stillwell, 1995). It may also be that the long 10-s presentation rate and instructions to look for a rule resulted in unusual rehearsal patterns; for example, perhaps people focused heavily on the reasons why some items were massed, resulting in unusual attention to massed repetitions. Finally, the anomalous result may simply reflect that with a very long presentation time—most of which was not used for studying—and incidental learning, items were very weakly encoded and so the optimal lag was quite short.

A more recent hybrid model by Raaijmakers (2005) combined encoding variability theories and study-phase retrieval theories. Raaijmakers only sought to explain cued-recall paradigms, but we will see later that extending the model to free recall is not terribly difficult with some additional assumptions. The model—which is based on Raaijmakers and Shiffrin's (1980, 1981) SAM model of recall—proposes that when an item is repeated, if people recognize the item as old, then they strengthen the memory for that item. If not, then they store the presentation as a new trace.

It is worth digressing into the SAM model of recall, because our own approach is grounded in SAM as well, and differs only a little from the Raaijmakers (2005) model. In SAM, memory traces contain three types of information: item content information, item context information, and associative information. *Item content information* refers to features of the item itself, such as its semantic properties, phonemic properties, and so on. *Item context information* refers to the link between the item and its context, and context is usually list membership in memory experiments. Finally, *associative information* stores relationships between traces in memory. (In fact, item context information is treated as a special case of associative information where the associate is the background context.)

A useful thing about discriminating item content from item context information is that they may interact with the type of test in interesting ways. The SAM model assumes that retrieval involves two types of processes, which are called sampling and recovery. When people attempt to retrieve, they start by sampling a memory trace from the sea of all memory traces. The sampling process begins with the cues present on the test (including the test context). Therefore, for cued recall and recognition, the associative information is usually the most relevant to sampling, with items that are more strongly linked to the test cues being most frequently retrieved. However, context information also plays a role, as the test context is always implicitly part of the test cues. In free recall, there is usually no test cue present, and so people rely almost exclusively on the context cues (at least initially). Importantly, sampling is a function not only of the strength of the "correct" item but of its strength *relative to all the other items in memory*. Thus, the chance of sampling the "correct" answer goes up as a function of the strength of the link between the test cues and the item, but it also depends critically on the strength of all the other items in memory to the same test cues.

Only once an item's image in memory is sampled do people attempt to recover the image. An item that cannot be successfully recovered will not be remembered. Recovery also depends on strength, but absolute strength (not relative strength). That is, for recovery it is not important how strong other competing items are; it only matters how strongly the cues activate the target item.

In SAM, all of the strengths are usually incremented whenever study is happening. Longer study results in stronger links between the cues and stronger item strength. (We will later propose that for free recall, this rule may be different, but Raaijmakers did not alter these default assumptions.)

Raaijmakers introduced additional assumptions in order to account for the lag effect. Specifically, he assumes that contextual elements change over time, and that when an item is encountered a second (or third) time, a retrieval attempt is made. If the item is still active in short-term

memory, then there is no need to make a retrieval attempt, and so no further information is stored. However, if the item has already dropped out of short-term memory, then the usual sampling and recovery process is engaged. If the retrieval of the prior presentation is successful, then additional contextual elements are stored with the image. If retrieval fails, a new image is generated. This mechanism captures the study-phase retrieval mechanism because a successful retrieval is required to store additional information about an item. At long lags, the context information mismatches, and so the probability of a successful retrieval drops. It also captures the basic lag effect, because the more time has passed since the first presentation, the larger is the contextual change between the two repetitions, resulting in a stronger link to the context. Finally, Ross and Landauer's critique that once-presented items in multiple places on the list should benefit from context variability (they do not) does not apply to the model, since additional contextual information is only stored following a successful study-phase retrieval. It does not predict better memory for two unrelated items when they are spaced apart, because context information is only incremented when a successful retrieval occurs. Hence, the Raaijmakers model incorporates both study-phase retrieval mechanisms and contextual fluctuation mechanisms.

3.9. Summary: Theories and Key Phenomena

Our review proposed that rehearsal explanations were insufficient to fully explain the spacing effect. Among other findings that argue against the rehearsal explanation is our demonstration that pure lists with incidental-learning instructions still show a spacing effect. Another set of findings that suggest rehearsal is not the end of the story is that spacing effects are obtained with infants and honeybees, who do not rehearse. Classical versions of encoding variability theories, which propose that spaced items have more varied neighbors during study, were proposed to explain the Glenberg surface, which related optimal lag and retention interval. However, they also failed in a long series of studies that showed that deliberate attempts to vary the neighbors of items did not help memory. Another theory that largely failed is the semantic priming account, which attributed spacing effects to priming-related processing deficits. Although a perceptual priming version of the account survives for "meaningless" materials, the semantic priming account failed on numerous counts.

The study-phase retrieval theory, which attributes spacing effects to the benefits of covert retrieval of previously seen items when they are re-exposed, fares better and is able to explain many important phenomena, including why difficult-to-learn items have shorter optimal lags and why inhibited items show a larger spacing effect than noninhibited items. More recent versions of the encoding variability account, built

around the SAM model, are likewise able to handle many more of the phenomena outlined here. It seems likely that hybrid accounts incorporating some aspects of study-phase retrieval and encoding variability theory are likely to provide the best-developed theoretical accounts of spacing.

4. Extending a Context Plus Study-Phase Retrieval Account of Spacing Effects

Our goal in this section is not to propose a full-fledged theory of spacing, but rather to add support to accounts similar to the SAM account proposed by Raaijmakers (2005). We will propose some likely extensions to his approach while providing additional experimental evidence that is consistent with the general account. We remind the reader that testing these accounts requires carefully controlling for all of the impostor spacing phenomena outlined in Section 2, and that many of the puzzling results in the literature that seem contrary to its predictions can be handled by noting that those results are vulnerable to one of the many criticisms raised in that section.

The Raaijmakers model as written has some shortcomings. First, quantitative tests of the model have been restricted to cued recall and intentional learning. Predictions regarding free recall seem less clear. Second, rehearsal-based confounds are possible in all of the data that Raaijmakers modeled. The data may therefore be vulnerable to critiques based on some of the impostor spacing phenomena in Section 2. Third, the short-term memory mechanism in the Raaijmakers model does not allow for enhanced context storage when an item is encountered again while it is still in short-term memory—a mechanism essential for capturing the deficient-processing impostor phenomenon in the model. However, it may be too generous, at least for cued recall, because it overestimates the probability that short spacings produce no memory benefits (Pavlik & Anderson, 2005).

Fortunately, a possible solution to the third problem has been published by Malmberg and Shiffrin (2005) as part of their research on the list-strength effect. We will therefore consider their solution next.

4.1. An Account of the List-Strength Effect Using SAM

Malmberg and Shiffrin (2005) have shown that a version of SAM/REM[2] can elegantly handle the list-strength effect results outlined in Section 2.5 by incorporating some assumptions about how item content and item context information accrue during study. To remind the reader, the list-strength effect occurs with mixed lists that contain some "strong" and some "weak"

[2] REM is Retrieving Effectively from Memory, and is a revision of the original SAM theory.

items. Recall of the strong and weak items on the mixed list is compared to recall on pure lists; pure-strong lists contain only strong items, and pure-weak lists contain only weak items. Compared to the pure-strong list items, strong items on mixed lists are better recalled. Compared to the pure-weak list items, weak items on mixed lists are more poorly recalled. Thus, on mixed lists, the strong get stronger and the weak get weaker. Furthermore, the effect happens robustly in free recall, but not frequently in cued recall or recognition, indicating that at least in free recall, one will often observe dissociations such that mixed lists show larger spacing effects than pure lists.

Malmberg and Shiffrin (2005) demonstrated that spacing is a special sort of strengthening in that it produces list-strength effects. Other kinds of strengthening manipulations such as giving each item extra massed-study time increased recall, but they did not produce a list-strength effect. Therefore, they proposed that spaced repetitions strengthen the link between items and their contexts, whereas other strengthening manipulations affect only the item strength. Specifically, Malmberg and Shiffrin proposed that each time an item is encountered, people store one "shot" of context, provided sufficient study time is allowed. During additional study time beyond that minimum—or during massed repetitions, which amount to the same thing as extra study time—no additional "shots" of context are stored; only the item strength is increased by massed study. However, if sufficient lag occurs between two repetitions, then both the item *and* context information are incremented. Their computational model based on these assumptions was able to successfully model important aspects of their experiments.

The Malmberg and Shiffrin version of SAM/REM preserves the distinction between two retrieval processes—sampling, which depends on the strength of items relative to all others in memory, and recovery, which depends only on that item (and its associative links to the probe cues). However, unlike in some other versions of SAM, the context strength plays a role mainly during sampling, and not during recovery.

According to their model, the list-strength effect is a relative strength phenomenon produced by the sampling process. That is, it depends on the competition between strong and weak items at test, not on the absolute strength of the items. In free recall, people use the context to sample items. For pure lists, there is no competition between strong and weak items (because they are not on the same list), so there is no sampling advantage for strong over weak items. However, on mixed lists, the strong items are sampled more frequently than the weak items. Consequently, the strong items benefit (relative to pure lists) and the weak items suffer (relative to pure lists).

Comparing the Malmberg and Shiffrin model to the Raaijmakers model, Malmberg and Shiffrin make no assumptions about contextual drift over time. The Malmberg and Shiffrin model therefore does not predict a lag

effect (which is wrong, at least in some circumstances). The Malmberg and Shiffrin model also does not make the assumption that items in short-term memory are not strengthened; in fact, it rather explicitly makes the opposite prediction, that items receive item content information as long as they are still being processed. Finally, the Malmberg and Shiffrin model works for free recall and has not been extended to cued recall, whereas the Raaijmakers model works only for cued recall. A productive activity for modeling researchers would be to merge the models into a more general SAM/REM account that can make accurate predictions on a broader range of data.

4.2. A Modified One-Shot Account of Spacing?

There are many possible ways to unify the two models, and actual modeling efforts will be needed in order to identify which is correct. However, we will take a stab at proposing a "verbal theory" that incorporates some quantitative assumptions in order to demonstrate the plausibility of a model based on these principles. (At the very least, future cognitive models will be able to use our list of phenomena to compare competing models; at best, our proposed theory can form the basis of a rigorous quantitative model that captures our intuitions.) The basic theory is similar to the Malmberg and Shiffrin (2005) theory, except that we add additional assumptions to account for cued-recall and recognition tests.

Specifically, we assume that the first time an item is studied stores context, associative, and item information. Item content information continues to accumulate as long as the item is seen; context information rises to a maximum value ("one shot") and then stops. This assumption is lifted directly from the Malmberg and Shiffrin (2005) model. Our possibly controversial addition to their account is that extra study time should not strengthen item-to-item associative information either; it also rises to a maximum value and then stops, unless study-phase retrieval occurs. We need this assumption in order to explain cued-recall spacing effects, because cued recall is less reliant on background context than it is on the association between the test cue (which is present both at study and at test) and the item.[3]

To explain the spacing effect, we assume that when an item is seen, people automatically initiate a search of memory for the identical or highly related items. This reminding process is the study-phase retrieval part of the account and follows the usual SAM/REM conventions; memory images are sampled using the current item and the background context as the cue and

[3] Actually, a more plausible assumption may be that there are two kinds of associative information, one that is used mainly in sampling and the other that is used mainly in recovery. It would be the former kind that discriminates spacing from massing.

followed by an attempt to recover sampled images. If an item is recovered—usually the same item, but highly related items may also be found—then the accumulation process begins again for the recovered item, strengthening item content, context, and associative information. It is worth stressing that the recovered trace is strengthened, and a new trace is not stored. If, however, an item has been seen before but the person fails to retrieve it, then a new image is stored instead. This differential strengthening produces the recovery advantage of items studied for a long time without increasing their sampling advantage.

Next, to explain the lag effect, we must introduce a mechanism that strengthens context and associative information when a successful retrieval occurs. The magnitude of this strengthening must be proportional to the difficulty of the retrieval, such that more difficult retrievals (along some dimension of difficulty) result in more strengthening (a *closed loop* phenomenon; see Murdock, 2003). We favor an account similar to the Malmberg and Shiffrin (2005) account, but that is extended to provide increasing context and associative information storage as a function of the lag between repetitions. Specifically, we assume that the increase in context and associative strength is proportional to the difference between the maximum possible strengths and the current strengths at the time of retrieval. Thus, for a short-lag item, the current strength will be relatively strong when its second and subsequent presentations come along, resulting in a "shot" of context and associative information that is relatively small. However, for long-lag items, the current strength will be weaker when its second and subsequent presentations appear, producing a "shot" of context and associative information that is relatively large. However, we note that there are many alternative formulations of this general rule that could be explored, so we will leave it at the level of a "verbal theory" for now.

At the test, the sampling and recovery implications of the "modified" one-shot hypothesis produce different effects depending on the type of test. For cued tests like recognition and cued recall, the impact of contextual information is minimized because there are associative cues present at both study and test. The list-strength effect is absent because at test people rely mainly on associative strength for sampling, and not on context strength. However, there is still a spacing benefit, because associative strength is increased during spaced restudy. For free recall, people rely less on associative cues and more on context strength. This produces a list-strength effect according to the same types of mechanisms described in Section 5.1.

An account like this can readily handle the phenomena in Table 3. The account produces intention invariance, as the reminding process is obligatory. It can be interpreted to produce age invariance, as there is nothing strategic about the process. The Glenberg surface emerges because the gain in associative and contextual strength is bounded, producing a local maximum gain at a certain retrieval distance. Furthermore, as the distance

between an item and its repetition increases, the probability of successful reminding goes down (a recency effect). Together, these mechanisms produce the U-shaped curve. Manipulating contextual variability will seldom help recall, because changing the cue between restudy opportunities means that the associative strength does not go up. Finally, a study-phase retrieval mechanism is used, so recognition is required to get a spacing benefit.

Last, we note that this account turns the list-strength effect from an impostor effect to a signature effect that is directly predicted by the model. Rather than viewing the list-strength effect as a problem, we view it as fully consistent with our context-based account.

4.3. Some Experiments Linking Context and Spacing

The previous account assumes that context strengthening is critical to producing spacing effects. By "context" we mean neither the semantic connotation of a word nor the other words that surround an item on a long list; rather, we are referring to *incidental background stimuli* that are present during encoding. For example, when the physical environment during test mismatches the physical environment during study, memory is reduced (Godden & Baddeley, 1975; Smith, 1979, 1984; Smith, Glenberg, & Bjork, 1978). Internal states can also be part of the context, such as mood (e.g., Eich, 1980).

While Raaijmakers' version of SAM assumes that list context fluctuates as people move through the list, it may well be that within a short list, the context does not change very much between the items. In such cases, the "list" context is relatively stable, and can be approximated without modeling the small drift in context during the list. Anderson and Bower (1972), for example, successfully modeled many list learning paradigms with a model called FRAN that linked studied items to a global "list" node. However, between lists or over time, the background context is likely to fluctuate more dramatically. Hence, context may change at different rates depending on what people are doing.

One piece of evidence that context fluctuates at a different rate within-list and between-list is that if you ask people to tell you how far apart two unrelated words were within a list, they are not very calibrated (Hintzman & Block, 1973; Hintzman et al., 1975). However, if you ask them to tell you which list an item came from, they are often quite accurate. Intrusion errors across lists are surprisingly rare following relational encoding like that in typical spacing studies, and people are capable of recalling items from a previously studied list even when a subsequent list intervenes, suggesting that people have good memory for list membership (e.g., Shiffrin, 1970).

An important study by Smith et al. (1978) demonstrated that studying a list in two distinct environmental contexts reduced the forgetting when

tested in a third context relative to studying a list twice in the same environmental context. This result is important because it shows that encoding variability at the list level is quite predictive of later recall, even if at the item level it is not. In general, studying in multiple contexts protects against forgetting caused by mismatch of the test and study contexts. One reason this might be true is that the test context has a better chance of matching some components of multiple contexts than it does of matching components of a single study context—essentially, an encoding variability phenomenon. Contexts may be more likely than individual cues (e.g., *a* in an *a–b* pair) to contain components that repeat across different contexts, as they are composed of a large number of features compared to the relatively few features present in a single cue.

Turning now to within-list context, a recent paper from one of our laboratories (Verkoeijen, Rikers, & Schmidt, 2004) explicitly manipulated the background against which words were studied. Experiment 1 used solid background colors, while Experiment 2 used a cityscape and a forest image. As with many studies that varied properties of the stimulus, changing the background improved memory for massed items compared to keeping the background the same. This is probably because of one of the impostors outlined in Section 2, deficient processing. However, for spaced items, changing the background reduced the probability that people would retrieve the earlier presentation, thereby lowering subsequent recall. These results are very similar to the results of the homograph studies and the change-of-encoding studies reviewed earlier (see Section 3.5).

4.4. Directed Forgetting as a List-Strength Phenomenon

Another piece of evidence that context is important to the spacing effect comes from a recent study from one of our laboratories (Sahakyan et al., 2008). We asked participants to rate mixed lists of spaced and massed words for pleasantness and animacy by giving a simple yes/no judgment on each dimension. To reduce deficient-processing explanations of our results, we used a different rating dimension on each presentation. After the first list, half of our participants were instructed to try to forget (i.e., directed forgetting) the previously studied list, as it was just for practice, and that the real list would be coming next. The other participants were told that it was just the first half of the list, and that they should keep rating words on the second list. After the second list, there was a distractor task followed by a free recall test on List 1.

Directed forgetting is often explained as a context effect (Sahakyan & Kelley, 2002). Sahakyan and Kelley argued that people comply with a forget instruction by "thinking of something else" which results in a new mental context being set up for the second list. At the time of the test, the test context better matches the second list than the first list, producing impaired

recall of List 1 items, but enhanced recall of List 2. Consistent with their explanation, Sahakyan and Kelley found that when the original context was reinstated using guided retrieval, memory for List 1 items recovered (and List 2 items suffered). Furthermore, when participants are told to keep remembering but were instructed to engage in a distracting thought—such as imagining themselves invisible or imagining their parents' house—they still show impaired List 1 recall (Sahakyan & Delaney, 2003; Sahakyan & Kelley).

From the perspective of SAM/REM models, directed forgetting is conceptually identical to a list-strength effect. List 1 items are associated with a context that mismatches the test context, and so List 2 items are sampled more often than List 1 items are. Once sampled, their recovery is unaffected. This is why it is difficult to obtain forgetting using recognition tests (e.g., Sahakyan & Delaney, 2005): just as in the list-strength effect, the primary cue that people rely on during recognition tests is the item shown, not the context. One would expect that with very difficult recognition tests that rely on retrieving specific stimulus information and not on familiarity alone, people would tend to rely more on the context. Indeed, it has been shown that asking people to discriminate whether items were presented in their singular or plural form (e.g., baker vs. bakers) and other recognition tests that rely heavily on recollection produce both list-strength phenomena (Diana & Reder, 2005; Norman, 2002) and directed forgetting of List 1 items (Sahakyan, Waldum, Benjamin, & Bickett, 2009).

Another directed forgetting finding that falls directly out of SAM/REM is that List 2 must be studied to produce forgetting. Because SAM/REM assumes that directed forgetting is a sampling phenomenon, without competing items to preferentially sample, there should be no directed forgetting. This prediction has also been confirmed in the directed forgetting literature (Bjork, 1989) and using Sahakyan and Kelley's (2002) context-change paradigm (Pastötter & Bäuml, 2007). Furthermore, one should expect that the more List 2 encoding occurs, the larger is the competition, as there are more items that could potentially be sampled—a prediction confirmed by Pastötter and Bäuml (2010).

The Sahakyan et al. (2008) study effectively combined a list-strength paradigm on List 1 with the two-list directed forgetting paradigm. A direct prediction is that since spaced items on List 1 are more strongly linked to the context than the massed items on List 1, then spaced items should also suffer more forgetting than massed items. Indeed, this is exactly what we found in our study, whose main results are reproduced as Figure 4. The greater forgetting of spaced than massed items is a direct—if counterintuitive—prediction of the SAM/REM account because spaced items are more closely linked to List 1 context. When the test context changes, their usual sampling advantage is lost, and they drop to near the level of massed items.

One detail of Figure 4 that is often puzzling is that massed items apparently show no directed forgetting at all. However, this is exactly what the

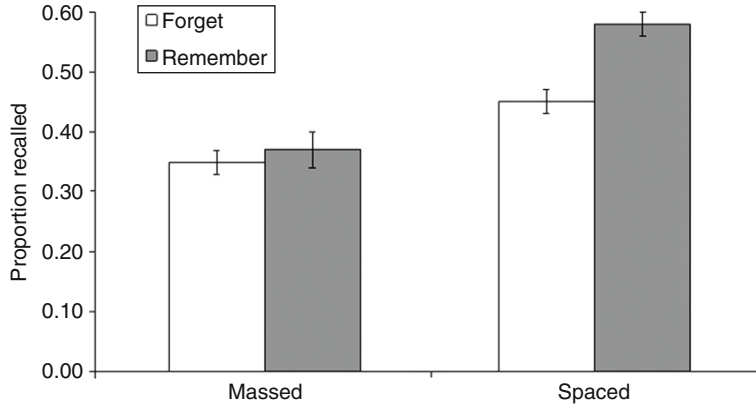

Figure 4 Proportion free recall of words as a function of spacing and intentional forgetting cue. Adapted from Sahakyan et al. (2008).

model predicts! Specifically, as the test context is no longer such a good match for List 1 spaced items following intentional forgetting, List 1 spaced items no longer enjoy such a large sampling advantage over massed items. As with the list-strength effect, strengthening the link between items and their context not only enhances their memory relative to weak competitors but also drives down recall of the weak competitors. Consequently, when the spaced items lose that advantage, massed items show spontaneous recovery. Therefore, while massed items are being "forgotten" because List 2 items are now competing with them more effectively, they are also simultaneously receiving less competition from the weakened List 1 spaced items, and hence the List 1 massed items become more memorable compared to the remember condition. These countervailing effects largely cancel one another out, producing no forgetting of massed items.

In sum, the Sahakyan et al. (2008) experiment is an important contribution because it confirms counterintuitive predictions made by the SAM/REM model of the list-strength effect. The full pattern of results of the study are difficult to explain without understanding what the model would predict, and are fully consistent with the context-based theory of spacing outlined here.

4.5. Summary and Untested Predictions of the Account

To summarize Section 4, we proposed that the Raaijmakers (2005) account of the spacing effect is largely correct. However, we proposed some extensions based on the Malmberg and Shiffrin (2005) SAM/REM account of the list-strength effect, which refers to the larger spacing effect observed with mixed lists than with pure lists. The hybrid account is able to explain

why the list-strength effect emerges in free recall but not in recognition or cued recall, and correctly predicts greater list-method directed forgetting from spaced items than from massed items on mixed lists. It also correctly predicts the pattern of recall in experiments that manipulate background incidental context. Taken together, this tentative account provides a number of correct qualitative predictions on which a computational model could be constructed.

The tentative theory described herein also makes a number of as-yet untested predictions. We are in the process of testing some of these, but consider it worth outlining them now to set others thinking along the same direction. The first and perhaps most amazing omission is that no one has demonstrated that longer lags between spaced repetitions produce list-strength effects. One should generally predict that mixing long-lag and short-lag items in a free recall study should produce a list-strength effect. This is not predicted by the original Malmberg and Shiffrin (2005) account, but it would be predicted from our expanded version of their model. Long-lag items would store more context than short-lag items, resulting in a sampling advantage at the test. They should then be output sooner and with higher frequency than the short-lag items. A failure to obtain a lag list-strength effect would falsify the prediction of greater context storage following more difficult retrievals.

One potentially surprising prediction is that the spacing effect should interact with retention interval differently depending on the type of test. As the retention interval gets longer, there is a greater mismatch between the test context and the study context. In free recall, spacing results in a stronger link between the item and the study context. Hence, if the test context mismatches the study context, then spacing will confer relatively little advantage over massing (for evidence that this pattern is observed in directed forgetting, see Sahakyan et al., 2008). Therefore, we anticipate that in free recall, increasing the retention interval will tend to reduce the advantage of spaced items. In contrast, in cued tests like recognition and cued recall, target items are associated during study with the cue. This cue will be presented again at the test, suggesting that for such tests the spacing effect should get stronger and stronger over time. This latter prediction has been confirmed in a number of studies, as reviewed in a recent meta-analysis (Cepeda et al., 2006). However, the former prediction has never been tested.

5. The Testing Effect

Up until this point, we have been discussing how distributing practice grants memorial advantages over massing practice. The studies described above mostly repeat items by granting additional study opportunities.

However, when an item is tested, it also constitutes a kind of restudy, and so most of the same things that take place in spacing studies also take place when the additional study opportunity is replaced with a test. Furthermore, it is now well known that inserting tests into a learning sequence produces better memory for the presented material than a similar amount of time spent studying. Researchers have extensively investigated this beneficial influence of testing on memory in recent years (see Roediger & Karpicke, 2006a, for an excellent review), and frequently make pleas for employing testing as a learning aid in educational practice.

Empirical work on the testing phenomenon—as well as calls for its practical application in the classroom—has a long tradition. For example, in an early review of the results from experimental psychology, Offner (1911) wrote, "Witasek hat, was Ebbinghaus und Pilzecker schon berührten, an Silbenreihen umständlich gezeigt, daß das bloße Lesen einen geringeren Einprägungswert hat als das Rezitieren." Loosely translated, this means Witasek demonstrated, consistent with what Ebbinghaus and Pilzecker had already suspected, that the memory trace of nonsense syllables is stronger when participants regularly recited (i.e., tested themselves) during learning than when they read the syllables multiple times. Furthermore, Offner, who was a preparatory-school principal in Munich, noted that the recitation method was gaining ground in school practice. He advised students to attend to the meaning of the material and to try to retrieve it from memory instead of re-reading it, at least for material like verse. Thus, it appears that in the beginning of the twentieth century, some German students were already being advised to use self-tests instead of re-reading to learn.[4]

In this section, we will follow a format similar to Section 3. We will simultaneously identify some phenomena that need to be explained, which we summarize in Table 4, and describe the development and periodic rejection of theories that have been proposed to account for the testing effect. As there are far fewer theories to explain testing than to explain spacing, this is a much briefer endeavor.

5.1. Early Research: Tests Slow Forgetting

One classic study on the testing effect was conducted by Gates (1917), who conducted a large-scale study aimed at comparing memory after restudy versus self-testing. Children in first, fourth, sixth, and eighth grade studied lists of nonsense syllables and brief biographies taken from *Who's Who in*

[4] Offner (1911, p. 52): "Vielfach gibt man den Schülern den Rat, beim Memorieren eines Gedichtes, einer Regel u. dgl. (. . .), wenn das Hersagen oder Vortragen nicht glatt von statten gehen will, ins Buch oder ins Konzept zu blicken, sondern sich aufs Folgende oder auf den Zusammenhang zunächst noch zu besinnen und erst wenn dies ohne Erfolg bleibt, nachzusehen."

Table 4 Some Possible Testing Phenomena.

1. *Testing effects grow over time.* While restudy often produces better memory than testing after a short delay (at least without feedback on the test), testing tends to produce better memory after a long delay. Compared to restudy conditions, tests result in slower forgetting over time.
2. *Test type invariance.* Tests benefit memory on other types of tests, not just the original type of test. Furthermore, all types of test benefit memory.
3. *Asymmetry.* Testing usually produces asymmetric recall benefits, whereas restudy results in symmetric recall. Specifically, if a–b pairs are studied, then a-? tests are provided, a-? tests benefit more than ?-b tests.
4. *Difficulty enhances testing effects.* More difficult retrievals typically result in bigger testing benefits. Both weakening the cues and increasing the lag between study and test result in bigger testing benefits.
5. *Testing reduces proactive interference.* Inserting a test after a study event seems to reduce or eliminate build-up of proactive interference on subsequent material.
6. *Integration weakens testing effects.* There is some preliminary evidence that integrated materials may weaken the testing effect.

Note that these effects are generally less well-established empirically than the effects in Table 3.

America. The children read the material, and at some point were told to stop reading and look away from the material in order to mentally retrieve whatever they could from their reading. The amount of time children spent on self-testing was varied (from 0% to 90% of the total study time). Immediately after learning, the children received a written free recall test. After 3–4 h, they were tested again. If the children were unable to recall, they could look back at the material during their self-tests—a feature that provides high ecological validity, as glancing back at the material during self-testing is probably what students do during self-testing. Obviously, however, this aspect of the procedure also loosened experimental control.

The general conclusion from Gates' study was that a combination of study and self-testing produces better memory than studying the same material over and over again. This result held for both nonsense syllables and biographies, and for most of the grade levels tested, with greater percentages of self-test resulting in better final test performance. However, the positive effect of recitation over restudying seemed to be moderated by a number of factors. First, age interacted with the effect of self-testing, as for nonsense syllables the effect did not occur for the youngest children (Grade 1). Second, the effect of self-testing was stronger for meaningful materials (biographical facts) than for meaningless nonsense syllables—a point we will return to later. Third, at least for meaningful materials tested after a 3–4 h delay, when self-testing took more than 60% of the time there was a

downturn in effectiveness. This suggests that a certain amount of study is required before self-testing can facilitate learning.

Another classic study involved over 3600 students—the entire population of 91 Iowa elementary schools (Spitzer, 1939). Spitzer's participants read an approximately 600-word instructional text on bamboo. Each student then received two or three tests within the next 63 days. The general result was that on a final test, students who received intervening tests performed better than those who had merely studied the text (but not received a test). Furthermore, it appeared from his data that the tests greatly slowed the forgetting rate over time.

Around 1940, the interest in the effect of testing on learning waned, only to emerge again in the 1960s. Hanawalt and Tarr (1961) compared the effect of an intermediate free recall test on final recognition performance. In their experiment, participants studied 23 statements that each contained a subject, copula, and final predicate adjective (i.e., Brown eggs are expensive). Following the study phase, participants in the intermediate-test group had to recall as many of the adjectives as possible, while participants in a study-once group engaged in an unrelated activity. After either 8 min or 48 h (note that there was also a condition that received the final test after 52 h; however, the results of this group were similar to those found in the 48-h delay group) participants received a final five-choice recognition test on the previously studied adjectives. After 8 min, the intermediate-test and study-once groups had similar recognition accuracy, but after 48 h the intermediate-test group had substantially better recognition accuracy than the study-once group did.

Taken together, the early studies showed that intermediate tests improved memory compared to study alone. However, the benefits of testing typically grow larger over time, perhaps because the forgetting rate is slower following a test.

5.2. The Importance of Retention Interval

Subsequent studies focused heavily on the effect of retention interval. For example, Allen, Mahler, and Estes (1969) gave participants lists of three-letter English nouns paired with two-digit numbers. Participants studied 27 paired associates, each consisting of a three-letter English noun and a two-digit number. On Day 1, the session began with ten cycles of 18 study trials; nine paired associates appeared in all ten cycles (training condition 10), nine appeared in the first five cycles (training condition 5F), and nine appeared in the last five cycles (training condition 5L). During each study trial, a paired associate was presented on a projector screen for approximately 2 s. Participants were instructed to repeat the item appearing on screen as often as possible. Immediately after the training phase, one-third of the paired associates in each training condition received five test trials, one-third

received one test trial, and one-third were not tested. At a test trial, a noun was presented on screen, and a participant was required to enter the two-digit number associated with this noun. Participants received no corrective feedback on their responses. After the test trial had been administered, participants were dismissed and they were instructed to return 24 h for the second session. In the second session, participants were tested on the paired associates they had learned on the previous day. The test format was identical to that used on Day 1 in the intermediate-test conditions, with each Day-1 item being tested four times. The dependent variable was the percentage of errors on the cued-recall test. For the present purposes, the most interesting outcome pertains to the comparison of two conditions, namely the condition in which items were studied ten times without an intermediate test (henceforth termed the repeated-study condition), and the condition in which items were studied five times (the combined training conditions 5F and 5L) and received five intermediate tests (henceforth termed the study-test condition). Allen, Mahler, and Estes showed that after 24 h, the average test performance in the study-test condition was better than in the repeated-study condition. This finding suggests that a relatively long retention interval is required before the memory benefit of intermediate testing over restudying emerges.

Hogan and Kintsch (1971) compared intermediate testing with restudying at multiple retention intervals. In their first experiment, there were seven conditions that differed in terms of the training schedule and the final test. Most important for now is to make a distinction between repeated-study conditions and study-test conditions. Participants in the repeated-study conditions studied a list of 40 words three times with short breaks between presentations. Alternatively, participants in the study-test conditions studied the list once and then received two intermediate tests. These tests were either two free recall tests or two two-alternative force choice recognition tests. In both conditions participants took a final test (a free recall or recognition test) immediately after the last study trial or after the last test trial. In addition, a second test (again a free recall test or a recognition test) was administered after 2 days.

The study produced some interesting findings. First, when free recall was used in the intermediate tests and in the final tests, the repeated-study group recalled more words than the study-test group on the immediate test. However, after a 2-day retention interval, mean free recall did not differ between the study-test group and the repeated-study group. Second, when recognition was used in the intermediate tests and free recall in the final tests, mean free recall performance after a 2-day retention interval was higher for the study-test group than for the repeated-study group. Third, when recognition was used in the intermediate tests and the final tests, it turned out that the repeated-study group performed as good as the study-test group. To strengthen the experimental manipulation from their first

experiment, Hogan and Kintsch conducted a second experiment. Participants studied the same 40 words as in Experiment 1 four times (restudy condition) or they studied the words once and took three consecutive free recall tests (study-test condition). A final test (either free recall or recognition) was administered after 2 days. The results showed that participants in the study-test condition outperformed those in the repeated-study condition on the final free recall test, whereas the reversed pattern was observed for the performance on the final recognition test.

On the basis of the results reported by Hogan and Kintsch (1971) we can provide a definition of the testing effect: *the testing effect refers to the finding that an intervening test leads to a better memory performance on a delayed test than restudying the material for the same amount of time*. Furthermore, this positive effect of testing on long-term retention emerges because more forgetting occurs following restudying than following intervening testing. The testing effect has proven to be a very robust phenomenon as it has been demonstrated in laboratory studies with simple stimulus materials such as word lists or paired associates, and with a variety of memory tests (e.g., Cull, 2000; Karpicke & Roediger, 2007; Wheeler, Ewers, & Buonanno, 2003), as well as in laboratory studies with relatively complex stimulus materials, for example, prose materials or short papers (e.g., Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Kang, McDermott, & Roediger, 2007; Nungester & Duchastel, 1982; Roediger & Karpicke, 2006b), visuospatial maps (Carpenter & Pashler, 2007), and obscure facts (Carpenter, Pashler, Wixted, & Vul, 2008).

5.3. The Return of Deficient-Processing Accounts

A problematic aspect of many early testing studies—at least when it comes to drawing conclusions about the effect of intermediate testing on retention—is that the observed mnemonic benefits of intermediate testing may simply be due to re-presentation of (some of) the studied material during a test rather to the testing *per se*. In other words, testing may introduce additional processing of items compared to restudy. If this sounds eerily familiar, it is because it is the deficient-processing theory of spacing applied to the testing effect.

To rule out this somewhat trivial explanation of the testing effect, an intermediate-testing condition ought to be pitted against a repeated-study condition. Such comparison was made in a study by Tulving (1967). In his Experiment 2, participants learned lists of 36 nouns in several different ways. One group studied the list, then was tested using oral free recall, then studied the list again, then repeated the test (the STST condition). This pattern was repeated six more times. In the study condition, participants received three study trials followed by a test trial (again six times). In the repeated-test condition, they studied the list once then received three test trials (again six times). Interestingly, Tulving found that the learning curves were almost identical in the three conditions. Thus, it seems fair to conclude

that when the total presentation time is controlled for, a study trial produces as much learning as a test trial. This seems to rule out the simplest form of the deficient-processing account.

However, some researchers have argued that the testing effect should be attributed to overlearning of the successfully tested items (e.g., Slamecka & Katsaiti, 1988; Thompson, Wenger, & Bartling, 1978). Overlearning occurs when an item that is already well known continues to receive practice. Hence, these overlearned items are at ceiling-level recall during the test, and even after their strength drops, they remain so well learned that they show no apparent forgetting. If tested items were overlearned relative to restudied items, then when forgetting happens with a delay, the restudied items will drop off the ceiling and show forgetting. The tested items, however, will weaken but still be at ceiling-level recall for some time, producing a slower forgetting rate. Only once they drop off the ceiling will they show the same forgetting rate as other items.

One way to overcome overlearning is to ensure that participants achieve a largely errorless intermediate-test performance. In Thompson and colleagues' third experiment this was done by testing and re-presenting series of five item sublists. The intermediate test on each sublist consisted of writing the five previously studied items down thrice with a short distractor task between the three free recall tests. Although maximum retrieval was not attained (participants recalled on average four of the five sublist items), overlearning was greatly diminished. Inconsistent with the overlearning account, a small advantage of repeated testing over restudying was found after a 48-h retention interval; no difference was observed after a short 20-min delay (see Kuo & Hirshman, 1996 for a similar short-delay result with a similar procedure).

Furthermore, other studies (e.g., Carrier & Pashler, 1992; Toppino & Cohen, 2009) have demonstrated strong testing effects using experimental procedures which seemingly prevent overlearning. In addition, the overlearning account predicts that intermediate testing will result in a superior final memory test performance at all retention intervals. However, most studies in the literature show that the testing effect only emerges after a relatively long delay; in fact, at short retention intervals restudying often produces a better memory performance than intermediate testing. This pattern is clearly at variance with the overlearning explanation of the testing effect. Lastly, when highly integrated texts are used as stimulus materials, it has been demonstrated that the beneficial effect of testing can "spill over" to information not tested during the intermediate test (e.g., Chan, 2009; Chan, McDermott, & Roediger, 2006). It is unclear how the overlearning account can accommodate such findings.

In sum, in view of the above-presented empirical evidence, we feel (in line with other researchers; see for instance Roediger & Karpicke, 2006a) that the testing effect cannot be attributed to either additional exposure or overlearning. That said, both additional exposure and

overlearning might be important "impostor" testing effects in certain circumstances. It remains to be seen whether clear evidence for either will emerge in experimental tests, but both are quite plausible problems.

5.4. Transfer-Appropriate Processing Accounts

A second mechanism that has been frequently proposed to explain the testing effect is transfer-appropriate processing (e.g., Roediger & Karpicke, 2006a). According to this approach, the testing effect emerges because the mental processes that enhance performance on a final memory test are more closely reflected in the processes occurring during an intermediate test than in those occurring during study. The basic tenet of transfer-appropriate processing—that transfer from the learning phase to a final test is optimal when there is a close match between learning and test—is very useful. For instance, if a teacher encourages learning strategies that lead to conceptual understanding of the class material, then a final examination should ideally also emphasize conceptual understanding. While transfer-appropriate processing is therefore useful to educators, there are a number of empirical findings militating against accepting the transfer-appropriate processing account as a theoretical explanation of the testing effect.

Like the overlearning account, the transfer-appropriate processing account has difficulties explaining why the testing effect is found after a long retention interval, but typically not after a short retention interval, and it cannot accommodate the finding that testing can also enhance memory for untested items (e.g., Chan, 2009; Chan et al., 2006).

Furthermore, an important prediction of the transfer-appropriate processing account is that memory performance is best when there is a close match between the intermediate test and the final test. However, studies aimed at assessing this prediction failed to obtain strong support for it. For instance, Glover (1989) in his Experiments 4A–C instructed participants to study a 300-words essay describing the fictitious state of Mala. Subsequently, in the control condition participants were dismissed, whereas in the other three conditions participants received a free recall test, a cued-recall test, or a recognition test on the previously studied essay. Two days after the first session, participants returned to the laboratory for a final test, which was either a free recall test (Experiment 4A), a cued-recall test (Experiment 4B) or a recognition test (Experiment 4C). Contrary to the transfer-appropriate processing account, final test performance was not best when the intermediate test and the final test were identical. Instead, in each of the three experiments, final test performance was always best when participants had received an intermediate free recall test (see also Carpenter & DeLosh, 2006).

Results reported by Carpenter, Pashler, and Vul (2006) also argue against the transfer-appropriate processing account. Their participants studied 40 weakly bidirectionally-associated noun-pairs (e.g., coffee-morning). After

the entire list, they received a restudy opportunity for 20 of the pairs, while the other 20 pairs were tested by presenting the cue and asking for the target. On tested items, the pair was further re-presented briefly after each test. A final test was administered 18–48 h after the study session. The most interesting test conditions were the *forward* condition, when participants had to provide the target in response to the cue (e.g., coffee-?), and the *backward* condition, when participants had to retrieve cue given the target (e.g., ?-morning). According to the transfer-appropriate processing account, the advantage of testing over restudying should be largest when the final test and the intermediate test are identical (i.e., forward). However, Carpenter and colleagues found that the magnitude of testing effect was comparable across the four conditions.

Interestingly, though, slight changes to the experimental procedure produce outcomes consistent with the transfer-appropriate processing approach. A new study by Carpenter, Pashler, and Jones (2008) used a procedure similar to the Carpenter et al. (2006) study, but this time using semantically unrelated pairs. As before, participants studied the pairs and then received either a restudy or a test opportunity on the pairs, with feedback provided after the test pairs. Testing benefitted both forward and backward conditions compared to restudy. Restudy produced symmetric recall in the forward and backward direction, but testing produced an asymmetry such that the direction that was used during the test resulted in superior recall compared to the reverse direction. Similar results have been reported by Zeelenberg, Pecher, and Tabbers (2008), whose participants studied a list of 24 unrelated word pairs five times in a row. A different list of 24 unrelated word pairs was studied thrice and subsequently tested twice by providing the cue and asking participants to generate the target. No feedback was given. After the study phase, a final test was administered on half of the items after 5 min and the rest after 1 week. Consistent with the Carpenter et al. (2008) results, after 1 week, they found that the testing effect was larger when the final test was in the same direction as the intermediate test than when the final test was in the opposite direction.

There is now a substantial amount of evidence that when two items X and Y are studied together, they form a bidirectional representation such that X serves as a retrieval cue for Y, and Y also serves as a retrieval cue for X. Anderson and Lebiere (1998), in their ACT-R models, have always assumed that study produces bidirectional relationships. However, when people repeatedly retrieve one member of the pair using the other, people may learn specific procedural rules that create asymmetric recall. That is, if I repeatedly use X to retrieve Y, then eventually I will bypass the usual declarative memory mechanisms and create a procedural memory of the form ''If X, then retrieve Y.'' As such rules are inherently asymmetrical, the ACT-R theory predicts that testing should result in asymmetric benefits for the direction of the test. Thus, the recent results demonstrating that testing benefits the direction of the test may not be evidence for transfer-appropriate

processing so much as they are evidence for a transition in the type of memory being formed by tests as compared to restudy. (Then again, perhaps it is the asymmetric production-based memory that is "transfer appropriate," as it is the type of memory that will be accessed at test.)

5.5. Retrieval Effort and Desirable Difficulty

If additional exposure, overlearning, and transfer-appropriate processing cannot adequately explain the testing effect, then what mechanism can? We will argue here that the desirable-difficulties framework and the related concept of effortful retrieval are very useful in understanding the testing effect. Bjork (1994) suggested that long-term retention is promoted when techniques that encourage students to engage in more effortful encoding operations during learning are used. Examples of these desirable-difficulties techniques are spaced practice, delayed feedback, and testing. Relative to restudying, taking an additional test requires more effort, and this may even slow initial learning. However, in the long run, testing will lead to better retention than restudying.

A straightforward prediction of the desirable-difficulties framework is that the beneficial effect of testing on a final test increases when the retrieval effort during an intervening test is greater, at least as long as an item is successfully retrieved. Several studies have provided support for this prediction using a variety of experimental manipulations. For example, both Glover (1989) and Carpenter and DeLosh (2006) compared the effect of three different types of intervening tests (free recall, cued recall, and recognition), on memory on a subsequent final free recall, cued recall, or recognition test. If retrieval effort is an important factor in the emergence of the testing effect, and if we assume that free recall requires more retrieval effort than cued recall and recognition, then an intervening free recall test should produce the largest testing effect regardless of the final-test type. The outcomes of Glover's experiments and of Carpenter and Delosh's experiment confirmed this prediction.

Other studies have substantiated the idea that information is better retained when it is harder to retrieve initially. Karpicke and Roediger (2007) showed that long-term retention is better with longer time-intervals between the presentation of information and the initial than with a shorter time-interval. Furthermore, Carpenter and DeLosh (2006, Experiments 2 and 3) examined final retention as a function of the number of cues participants needed to retrieve an item during an intervening test. They found fewer cues, and hence a greater retrieval effort, led to a better final test performance after a 5-min retention interval. Also, Pyc and Rawson (2009) provided more direct evidence for the retrieval-difficulty hypothesis (which is inherently related to the desirable-difficulties framework) by showing that difficult but successful retrievals produce better memory than easier successful retrievals both after a short and a long retention interval.

Thus, the results of the above-presented studies are clearly in line with the desirable-difficulties framework. However, it is not clear what exactly is enhanced by more difficult retrievals. There are many possible ways that memory could be enhanced by more difficult retrievals, all of which would be broadly consistent with the desirable-difficulties framework. It is therefore worth asking whether more detailed psychological mechanisms could be specified that produce results consistent with the desirable-difficulties framework.

5.6. Why Does Testing Help More Than Restudy?

One possible explanation for the testing effect is that testing enhances encoding variability. McDaniel and Masson (1985), in their Experiment 1, had participants encode words using either semantic or phonemic cues. A control group left after encoding, while the other participants received an immediate cued-recall test. After a 1-day delay, everyone returned and received a final cued-recall test using either semantic cues (category cues) or phonemic cues (rhyming cues). When the original encoding and the type of final test matched, there was no advantage of having received a test (21% recall) compared to not (22% recall). However, when the final test mismatched the original encoding, a test produced better recall (18%) than no test (11%). Hence, it seems that a test primarily helped make judgments that differed from the original type of encoding. In Experiment 3, the single-study condition from Experiment 1 was compared against a restudy condition. The restudy condition was almost identical to the intervening-test condition from Experiment 1 except that the intervening cued-recall test was replaced with a restudy opportunity. As before, restudying yielded a larger final-test advantage when final-test cues mismatched the original encoding than when the final-test cues matched the original encoding.

McDaniel and Masson interpreted these findings in terms of an encoding variability mechanism. If an intervening test or an additional study opportunity somehow adds new information elements to the existing memory trace, then the number of retrieval cues increases. Furthermore, and entirely consistent with the results found by McDaniel and Masson, these additional retrieval cues will particularly facilitate final-test performance when the final-test cue is different from the original encoding. Although McDaniel and Masson did not directly compare restudying with taking intervening tests, the combined findings from Experiments 1 and 3 suggest that encoding variability may underlie the testing effect. Specifically, an intervening test, but not, or to a more limited extent an extra study opportunity, serves to increase the number of retrieval cues encoded with an item's memory trace and this will provide tested items with a memory advantage over restudied items on a delayed final memory test.

Recently, it has been demonstrated that testing insulates against the build-up of proactive interference (Szpunar, McDermott, & Roediger, 2007; Szpunar et al., 2008). This finding can also be interpreted as being consistent with the encoding variability explanation of the testing effect. Consider the third experiment in the study of Szpunar et al. (2008), in which three groups of participants had to study five 18-words lists. In one group, participants received a free recall test after each list. In a second group, participants restudied each previous list, and in the third group, participants studied each list only once. Thirty minutes after the fifth list had been (re)studied or tested, a final free recall test was administered to all participants. The critical comparison between the three groups pertained to List-5 performance because proactive interference should be strongest for the last list in the initial study sequence. It was demonstrated that the intervening-test group outperformed both the restudy group and the study-once group on List-5 memory. In addition, the last two groups did not differ in terms of List-5 memory. These findings provide a strong argument for the idea that relative to restudying lists, or studying lists once, intervening tests protect against proactive interference. To explain this phenomenon, Szpunar and colleagues proposed an encoding variability mechanism. They suggest that ''testing adds contextual elements to a memory trace—over and above those added by a restudy episode—that enhance subsequent discriminability of recalled materials'' (p. 1397).

Given the presented empirical evidence, encoding variability may be the mechanism underlying the testing effect. However, Carpenter (2009) put forward the elaborative retrieval hypothesis as an alternative explanation of the testing effect. In this view, which is heavily based on spreading activation theories of memory, retrieval involves searching memory for a specific target, which activates a network of related concepts. The generation of this elaborative structure becomes helpful on a delayed final-test because it provides multiple retrieval routes to an item. By contrast, during a restudy trial, it is less likely that a participant will generate such elaborative structure, because an item is directly available. Therefore, on a delayed final test, tested items will be better remembered than restudied items.

Carpenter provided support for the elaborative retrieval hypothesis by comparing memory performance on previously studied paired associates. In Experiment 1, participants studied cue-target pairs that were either weakly associated, such as basket–bread, or strongly associated, such as dentist–teeth. During the initial encoding phase, all pairs were presented one by one on the computer screen, and participants had to rate the degree of relatedness between the words. In all pairs, the target appeared in bold, underline font. Following the initial study phase, half of the weakly associated and strongly associated pairs were rated again on their relatedness (restudied pairs), whereas the other half were tested. During a test trial, the cue was presented and participants had to enter the studied word.

No corrective feedback was given after the response. Five minutes after the pairs had been restudied or tested, a final free recall test was administered asking participants to recall as many of the underlined targets as possible. The final test results demonstrated that tested items were better recalled than restudied items. This outcome is quite interesting because it runs counter to the frequently observed finding in the testing literature that memory performance for restudied items surpasses that of tested items after a short retention interval (but see Carrier & Pashler, 1992). Furthermore, it was demonstrated that weakly associated pairs were better retained than strongly associated pairs; that is, the difference between the proportion correct at the intervening cued-recall test and the proportion correct free recall at the final test was smaller for weakly associated pairs than for strongly associated pairs. These findings were replicated with a slightly different procedure in Experiment 2.

Carpenter took the results of her study as evidence in favor of the elaborative retrieval hypothesis. Under the assumption that target retrieval requires more elaboration with a weak cue than with a strong cue, if follows that the number of pathways to a target is larger for targets from weakly associated cues than from targets from strongly associated cues. Consequently, targets from weakly associated cues will suffer less from forgetting than targets from strongly associated cues.

An interesting prediction that follows from the elaborative retrieval hypothesis is that an initial test should benefit memory not only of the tested information but also of related, but untested information. Chan (2009) and Chan et al. (2006) have corroborated this prediction.

5.7. Testing Effects for Integrated Stimuli

In the vast majority of testing effect studies, the stimulus materials are lists of items with a low level of integration; individual list items, such as words or paired associates, are not in any way connected to each other. However, there is both an empirical and a theoretical argument that the testing effect may be smaller for integrated than for nonintegrated materials. The empirical argument is based on a finding reported in a study by Chan et al. (2006; Experiment 1). They demonstrated with an integrated text about the toucan bird that restudying led to a better recall of text information than intermediate testing after a retention interval of 24 h. However, and as pointed out by Chan and colleagues, this finding should be interpreted with caution because the retention interval was short compared with other testing studies. Therefore, it may be possible that the restudy superiority would disappear and reverse, i.e., turn into a testing advantage, with a longer delay. Alternatively, there is also a theoretical argument for the idea that the testing effect will be weaker for integrated than for nonintegrated materials. Specifically, when participants study materials that are integrated either as

a result of material characteristics (e.g., all items in a word list are from the same category) or due to instruction characteristics (e.g., participants have to make a story of initially unrelated items, cf. Delaney & Knowles, 2005), they can construct a gist-feature that binds the studied items together. Furthermore, this gist-feature may serve as such a strong retrieval cue on a final memory test that it reduces the beneficial effect of intermediate testing over restudying.

Recently, we conducted two experiments that provide information about the role of integration in the testing effect. In one experiment (Verkoeijen & Delaney, 2010a), participants studied a list containing unrelated word using either a continuous rehearsal strategy (i.e., keep rehearsing as many words from the list as possible) or a story strategy (i.e., make a story of the words in the list). Apart from the learning strategy, we manipulated study type (restudy vs. intervening test) and the retention interval from the last study episode to the final test (5 min vs. 7 days) as between-subjects factors. Furthermore, free recall was used at both the intervening and the final test. Remarkably, the analysis of the final-test free recall performances revealed a three-way interaction. For the rehearsal strategy, a classic testing pattern emerged with more forgetting occurring for restudied words than for tested words. By contrast, in the story-strategy condition, we found a main effect of study type and length of the retention interval, without a trace of a study type by retention interval interaction. On average, restudying led to a better final-test performance than testing after 5 min and after 7 days. In addition, the forgetting rates were nearly identical for both study types.

In the other experiment (Verkoeijen & Delaney, 2010b), we asked participants to learn four categorized lists and manipulated study type (restudy vs. intervening test) and retention interval (5 min vs. 7 days) within-subjects. At both the intervening and the final test, a free recall test was administered to the participants. The final-test results showed that average performance was the same for restudied and for tested items after 5 min and after 7 days (also, overall performance was worse after 7 days than after 5 min).

The above-presented results suggest that the testing effect may be smaller for integrated than for nonintegrated materials. However, this preliminary finding needs to be corroborated by other empirical evidence. The experiment with the categorized lists also seems to indicate that the testing effect is absent with integrated materials. Yet, to strengthen our position, an extra experiment needs to be run in which categorized and noncategorized lists are compared with respect to the testing effect.

5.8. Summary: The Testing Effect

Early results from the testing literature indicated that tests seem to slow forgetting. Often on an initial test there is little advantage of testing over restudying, but due to the differences in forgetting rates, testing ultimately

results in better retention. As with spacing, deficient processing might produce some of the apparent testing benefits, and we should be mindful of this possibility. One way that could happen is through overlearning, with some items showing ceiling-level performance and then slower forgetting over time. However, studies seem to show that testing effects occur even with items that are not at the ceiling, so deficient processing is unlikely to provide the whole story.

One explanation for the testing effect focused on match between study and test processes (transfer-appropriate processing), but several studies show that tests help even when study and test processes mismatch. Later accounts focused mainly on the difficulty of retrieval, suggesting that more effortful retrievals produce more resilient memory traces. The latter account is quite similar to the study-phase retrieval account of spacing, and can explain most of the critical phenomena in the testing literature. Among the pieces of evidence for the retrieval difficulty account was that increasing the lag between study and test and reducing the specificity of the cue during the test both increase retrieval difficulty and enhance the impact of a test. Furthermore, unlike restudy, testing sometimes creates an asymmetric memory benefit such that the portion of the material that is retrieved benefits to a greater degree than the cue used for retrieval. Finally, we presented some new data which indicate that integrated materials may show smaller testing effects than nonintegrated materials because the former rely less on contextual information and more on item-to-item associations formed during study.

At this time, there is no formal computational model of the testing effect, although the ACT-R model has some successes in this direction. Future research should be directed at creating an integrated computational model of spacing and testing.

6. Spacing and Testing in Educational Contexts

Historically, studies on spacing and testing have been conducted in tightly controlled laboratory settings in which competing theories have been developed and tested. However, extending laboratory findings to educational settings is equally important. Applied studies, where longer delays and educationally relevant materials are used, have yielded results that are analogous to basic findings. For example, in a 6-week web-based Brain and Behavior course, being quizzed relative to rereading course material produced superior subsequent recall on a final exam (McDaniel, Anderson, Derbish, & Morrisette, 2007)—the standard testing effect. Additionally, as retrieval difficulty increased on the initial quiz, so did performance on the final exam, a finding that is consistent with laboratory research

(e.g., Carpenter & DeLosh, 2006). Other research has found spacing and testing effects to extend to a variety of educationally relevant materials including scientific prose (e.g., Roediger & Karpicke, 2006b), maps (e.g., Carpenter & Pashler, 2007), foreign languages (e.g., Bahrick, Bahrick, Bahrick, & Bahrick, 1993), history facts (e.g., Carpenter, Pashler, & Cepeda, 2009), and math learning (e.g., Rohrer & Taylor, 2006).

Recognizing that spacing and testing are excellent candidates for improving memory for factual knowledge, researchers have strongly recommended that educators include spaced practice and frequent testing in schools as ways to improve educational outcomes (e.g., Pashler et al., 2007; Roediger, Agarwal, Kang, & Marsh, 2010). Advocacy for the inclusion of spacing and testing in schools stems from the fact that they are empirically supported methods for improving memory.

As applicable to education as spacing and testing are, we argue that there are at least four unaddressed questions that prevent spacing and testing from having a greater impact on learning. The thesis of our argument is that cognitive psychologists have been successful at identifying how spacing and testing improve memory, but that there remain unaddressed concepts central to improving education. What follows are descriptions of those four questions and how they can be addressed. Because the application of spacing and testing have been recently reviewed elsewhere (e.g., Cepeda et al., 2006; Pashler et al., 2007; Roediger & Karpicke, 2006a; Roediger et al., 2010), we will only review prior research as it relates to our commentary. The questions we pose, our criticisms, and recommendations, apply only to research that seeks to make direct contributions to education.

6.1. Do Spacing and Testing Improve Learning or Just Memory?

Applied research on spacing and testing typically asks participants to study novel information (e.g., foreign vocabulary) and examine how some treatment (e.g., testing or spacing) impacts memory relative to a control group (e.g., restudying or massing). These studies have taught us a great deal about how spacing between study opportunities (e.g., Bahrick et al., 1993), retrieval difficulty (e.g., McDaniel et al., 2007), feedback (e.g., Butler, Karpicke, & Roediger, 2007), and retention interval (e.g., Cepeda et al., 2008) can be optimized to produce superior memory.

However, the typical focus of spacing and testing research is on memory, not on other kinds of learning. Although some work has shown that spacing benefits skill learning (e.g., Rickard, Lau, & Pashler, 2008; Rohrer & Taylor, 2006), rote memory is the usual dependent variable investigated in spacing and testing experiments. Kintsch (1994) drew a distinction between remembering and learning, where remembering involves being able to recall or identify a set of previously seen items. Learning, according

to Kintsch, implies deeper understanding of a subject where knowledge can be used flexibly. Thus, despite sometimes impressive spacing and testing effects, it is unclear whether these manipulations enhance memory alone, or *both* memory and learning.

In terms of making recommendations to educators, this is an important distinction. In schools, memory often is the primary outcome measure (e.g., most multiple choice exams), but one job of schools is to prepare people for employment where success depends on applying knowledge to novel situations. For example, remembering the historical causes of a societal collapse would allow one to perform well on an exam in school, but making a contribution outside of a school setting would require inferring what downfalls of past societies can tell us about prevention of our own societal failure. Learning, in other words, allows one to use prior knowledge to make novel connections and aid in solving an array of problems.

Several studies have demonstrated that remembering and learning (in the sense described here) are independent constructs. For example, before reading a technical article about microbes, Mannes and Kintsch (1987) gave participants background material that was presented either in the same order or a different order than the article. Although participants in the same-order condition outperformed the participants in the different-order condition on later free recall and sentence verification tasks about the article, participants in the different-order condition outperformed the same-order participants on inference and problem-solving tasks. Kintsch (1994) explained those results by attributing the difficulty associated with deriving coherence between the background text and the target text with forcing people to create a richly interconnected mental representation of the two. When background material matches the target text, there is little interference or need to develop a new mental model to integrate the two. Although this match facilitates rote memory, it is not as conducive to problem solving or inference making abilities. Kintsch's ideas are consistent with fuzzy trace theory (Brainerd & Reyna, 1990), where it has been found that studying material verbatim leads to relatively better memory, but activities that promote more gist-like encoding produce a deeper understanding of the material (Wolfe, Reyna, & Brainerd, 2005). At this point, it is unclear if spacing and testing have any effects on problem solving beyond contributing increased knowledge, or if they facilitate more sophisticated mental models.

One recent study has explicitly evaluated learning instead of merely memory (Kornell & Bjork, 2008). Motivated by Rothkopf's quote that "spacing is the friend of recall, but the enemy induction" and by research showing a massing effect in inductive learning, the authors set out to investigate if massing is in fact more conducive to inductive learning than spacing. In Experiment 1a, participants were shown six different paintings from each of 12 different artists. Six of the artists' works were presented in

spaced format, and the other six artists' works were presented in massed format. Experiment 1b was exactly the same, except spacing and massing were manipulated between subjects. At test, participants were shown new paintings one at a time from the previously seen artists and indicated which artist they thought painted the piece. In both experiments, participants were better able to infer new artists' paintings when they learned that artist's work through spaced presentation. Given that the results could be explained by participants simply being able to better remember which artist painted which painting in spaced conditions (a finding that would say nothing about inference), a second experiment was conducted that was almost identical to 1a (the only difference was that the test required participants to discriminate between familiar and unfamiliar artists). The results again revealed an advantage for spaced presentations. A similar study using children was recently published. As with undergraduates learning artists, it turns out that spacing instances of categories improves children's ability to induce whether a new item is a member of the category or not (Vlach et al., 2008).

A recent study by Johnson and Mayer (2009) explored whether testing benefits comprehension relative to restudy. In this study, participants learned a narrated animation about how lightning works. This animation, which was 140 s long, was presented on a computer screen. Afterwards, some participants had to study the same animation again (restudy condition), other participants were given a retention test, and the third group of participants received a transfer test consisting of four questions. Subsequently, half of the participants in each of the three conditions received a final retention test and a final transfer test after 5 min, whereas the other half of the participants received these tests after 7 days. It should be noted that the final retention test was identical to the intervening test; the final transfer test consisted of two questions from the intervening test and two new questions. For the present purpose, the most important finding was that at the 7-day delay, Johnson and Mayer found a testing effect on new transfer questions. That is, participants who had received an intervening transfer test, scored better on the new transfer question of the final test than the participants in the restudy condition. However, one peculiar aspect of Johnson and Mayer's study was the animated computer lesson, which was presented without any learner control. This type of material and the presentation format are not typically used in educational settings. Furthermore, the new transfer score was based on only two items, which is problematic in terms of the reliability and the validity of the test scores. Hence, it remains to be seen whether Johnson and Mayer's results can be substantiated in upcoming research.

Outside of the spacing and testing literature, researchers have spent many years studying distant transfer—that is, applying knowledge from one domain to solve problems in a relatively unrelated domain. An example of distant transfer is an army general applying his knowledge of chess to

battlefield tactics. The conditions under which people are able to execute distant transfer are not well understood (Barnett & Ceci, 2002), but it does remain a construct worth studying with respect to spacing and testing. When business and education leaders call for graduates with complex thinking skills, they are often speaking of distant transfer. In other words, they believe school should give students the knowledge and the skills to take what they learned in the classroom to generate ideas and solve problems in the real world. Using spacing and testing to develop students with such far-reaching abilities would require that cognitive researchers move beyond memory performance as the primary dependent variable in their research. Taking a cue from education researchers, cognitive psychologists might aim to better understand how spacing and testing impact skills such as critical thinking (e.g., Quitadamo & Kurtz, 2007), comprehension (e.g., Konopak, Martin, & Martin, 1990), and interpretation (e.g., Beins, 1993).

6.2. How Prevalent Are Spacing and Testing in Classroom Settings?

Many researchers point out that spacing and testing are rare in classrooms and that expanding their use would benefit education (e.g., Dempster, 1996; Pashler et al., 2007). Based on how effective spacing and testing are at improving memory, this is a logical conclusion. Anecdotally, high school teachers and college professors seem to teach in a linear fashion without repetition and give three or four noncumulative exams. Rohrer (2009) alleges that mathematics textbooks usually present blocked practice on a given topic, and only more rarely present review problems that would constitute spaced tests (see also Stigler, Fuson, Ham, & Kim, 1986).

Such structural problems seemingly preclude the possibility of frequent spaced- and retrieval-practice. However, measuring the prevalence of spaced- and retrieval-practice solely on the nonrepetition of lesson plans and the paucity of tests might underestimate their true frequency. Spaced practice is implicit in many domains. In statistics, ANOVA might be learned early in a semester and regression late in the semester, but many teachers likely review ANOVA when presenting regression for the first time. Even if instructors do not review information verbatim, there is evidence suggesting that when the second presentation of an item is a gist version of the second, massed items may be remembered just as well as spaced items (Dellarosa & Bourne, 1985; Glover & Corkill, 1987). Like in statistics, units of information in other domains do not exist in isolation but are integrated with other units of information. Learning newer information often requires restudying and retrieval of older information. The interconnected nature of knowledge might therefore inherently encourage spacing and testing, even if instructors do not deliberately try to build spacing and testing into their courses.

Prevalence estimates of spacing and testing in classrooms may also shortchange the value of in-class discussions. Discussion with classmates about a topic includes listening to what other people say (a form of restudying or spaced practice) and retrieving prior knowledge (a form of a test); research on writing shows that retrieving information in order to form arguments improves memory compared to rote retrieval (Wiley & Voss, 1996). Furthermore, instructors who pose questions to the class, even if they are rhetorical, might initiate students to covertly retrieve information. In sum, the real amount of spacing and testing in classrooms may need to be assessed by observing real classrooms.

Broadly speaking, it is important that researchers in the field develop a better understanding of spacing and testing in classrooms. Rohrer and Taylor (2006) have provided estimates indicating that spacing of problems in mathematics textbooks is the exception rather than the rule. Beyond that, we do not know much about how common spacing and testing are in classrooms. More specifically, we think that researchers have paid insufficient attention to what defines spacing and testing. For example, in many studies on the spacing effect spaced practice is compared to massed practice and in studies on the testing effect testing it is compared to restudying. If spacing and testing can in fact be something other than formal opportunities for restudy or tests, then future research might aim to uncover what constructivist activities that could be implemented in the classroom encompass spacing and testing. For example, a study might compare students who are tested versus students who engage in class discussions. This design would allow researchers to uncover informal instances of spacing and testing in the classroom. If class discussions prove to be just as effective mnemonic devices as traditional spacing and testing, research might be doing students a disservice by trading class discussion time for traditional restudy and retrieval practice.

6.3. How Can One Improve Learners' Use of Spacing and Testing?

Recent calls to use more spacing and testing have generally focused on classroom instructors. However, given that much of the learning we do happens outside of the classroom, one wonders how much more could be achieved by helping learners to space their own practice and to effectively test themselves. Given that when students study, they usually have control over which items they will study, it is important to know whether they even think spacing and testing are helpful. It is also important to know if, given the choice, they will space their own practice or not. If students are already doing substantial spaced practice on their own, then teachers' attempts to encourage spaced practice in the classroom may help very little (if at all).

One way to find out whether people are aware of the benefits of spaced practice is to compare perceptions of training regimens after people have experienced them. An important study by Baddeley and Longman (1978) involved teaching postal workers to type. The authors varied how much the training on subskills was spaced (interleaved) or massed (blocked). At the end of training, they asked the postal workers to indicate how satisfied they were with the training, and found that the objectively most effective training method, which involved the most spacing, was the least liked and that many postal workers would even refuse to participate if asked to train like that again. In contrast, the objectively least-effective regimen, which involved the most blocking and massing, was the most liked. Our interpretation of these results is that people find spaced practice effortful and unrewarding, at least when there is a lot of task-switching involved as well. Consistent with these results, Simon and Bjork (2001) gave people massed or spaced practice on a motor learning task and found that while massed practice resulted in faster acquisition of each response, spaced practice resulted in far better retention. Nonetheless, when people were asked which they preferred, they thought massing was better and that it promoted learning to a greater degree than spacing. Similar results were found in the Kornell and Bjork (2008) study in which participants learned painters' styles (see Section 6.1): more than 80% of participants classified more paintings correctly with spaced repetitions, but right after study, an approximately equal percentage believed massed presentation was at least as effective as spaced presentation. Taken together, these results suggest that people have little insight as to whether massed or spaced presentation promote learning, and may be tempted to mistakenly attribute the fluency of performance during study for effectiveness of training in the long run.[5]

Furthermore, it appears that students left to their own devices rarely space their study. A recent anonymous survey of over 200 University of North Carolina at Greensboro introductory psychology students conducted by the first author found that most indicated that they did not space their study; instead, they would study a single chapter straight through, and then move to the next, without ever revisiting the earlier one. Additionally, the majority of students indicated that they study only the night before an exam, although there was a sizeable minority who indicated that they study "a little every day." (There was also a not inconsiderable minority who indicated doing neither; they reported that they "rarely or ever" study at all.) In contrast, Karpicke, Butler, and Roediger (2009) reported that students' most favored study strategy is rereading.

[5] It seems strange to us that nobody to our knowledge has conducted the identical study for vocabulary memory, where there is no task switching. Task switching is generally effortful and unpleasant, but spacing is perfectly possible in vocabulary learning without any task switching at all.

Given that students do not space their study sessions, perhaps they nonetheless spaced their practice within a given session. If so, then given the choice, they should prefer to space items rather than to mass them. To find out, Ciccone and Brelsford (1976) allowed participants to choose the order of presentation of CVC paired associates (e.g., MAQ–TOJ) by pushing a button during the first presentation of the pair. Their goal was to learn all of the pairs in a set of 16. Participants chose lags of 2–5 approximately 70% of the time, suggesting that they favored short-lag spacing, and studied items on average between 11 and 14 times. After a 24-h break, they recalled an impressive 88% of the responses correctly on a surprise test. However, the most important aspect of the study was that having control of one's own study lag improved recall tremendously compared to a yoked control who received the same schedule. Apparently, people avoided lags that were too short or too long, and probably studied the items they personally found difficult more times. Aside from the fact that having control over one's study improved learning and retention, the study suggests that people are smart enough to avoid the massed item deficit.

More recent studies have examined whether students are sensitive to item difficulty when making decisions about spacing or massing items. An ingenious study by Son (2004) presented participants with a list of words to study, some of which were more difficult to learn than others. They then could choose whether they wanted to see the same item again immediately, or whether they wanted to "save" it for a spaced presentation. She found that people tended to mass the harder items and space easier items. Benjamin and Bird (2006), however, forced participants to space exactly half of the items. Under these conditions, participants preferred to space the harder items. These divergent results were reconciled neatly in a recent study by Toppino et al. (2010), who noted that presentation rate was a major difference between the earlier studies, with Son using a faster pace than Benjamin and Bird. Toppino et al. showed that for difficult items, participants could not fully encode the item in the time given, so they elected to mass the items. However, for easier items, they were more often able to fully encode the item, and so they spaced them. At slower presentation rates like those used in the Benjamin and Bird study, participants always elected to space the items. In a second experiment, they showed that participants often reported not perceiving the words if they were difficult and passing by too quickly, consistent with their argument that participants massed items to avoid skipping them entirely.

In sum, learners are fairly savvy when it comes to making item-by-item decisions about spacing. However, they are easily fooled by the fluency induced by massed practice into thinking that massed learning is superior to spaced learning. Speed of learning is not always a good indicator of effective retention in the future. Finally, students are not usually good about spacing their study sessions, even if many are normatively aware of the long-term

benefits of spacing study sessions. This problem may be exacerbated by the fact that massed study sometimes yields good scores on tests that occur shortly after the massed study; cramming may work if you know that the test will never be repeated again. Educators may want to consider clever methods for encouraging students to space their study at home.

6.4. Are There Individual Differences in Spacing and Testing?

Over recent decades, researchers have investigated numerous variables that can be manipulated to optimize spacing and testing effects. Unfortunately, almost none of these studies have assessed individual differences. The drawback of sweeping advocacy for spacing and testing in schools is that a learning schedule that benefits one student might have neutral effects for another student, or even come at a cost to more effective study strategies for others.

With the benefits of spacing and testing potentially emerging as a result of memory retrieval of previously seen items (Greene, 1989), baseline memory abilities might be a source of individual differences in the spacing and testing effects. The inverted U-shape of memory performance as a function of lag between items in spacing studies is evidence for this (e.g., Verkoeijen et al., 2005). Based on this, it is presumed that optimal lag differs depending on baseline memory abilities, but there is only indirect evidence for this hypothesis. For example, Sperber (1974) showed that in children who are mentally retarded, spacing practice is sometimes detrimental to those with lower IQs compared to those with higher IQs. In addition, Verkoeijen and Bouwmeester (2008), using a latent class regression analysis technique, demonstrated that under certain conditions, the spacing effect is smaller for college students with an overall lower memory-performance level than for students with an overall higher memory-performance level.

This same general relationship may also exist between testing and working memory. One factor that helps to optimize the testing effect is successfully retrieving an item once it has been cleared from working memory (Karpicke & Roediger, 2007). However, if an item has been cleared from working memory and it can no longer be retrieved, the benefits of testing are likely to be smaller (Baddeley, 1990). Based on this premise, we presume that a person with lower memory abilities would benefit from retrieval practice at some time sooner than a person with higher memory abilities. This hypothesis is yet to be tested (although Latasha Holden in the Delaney lab is currently conducting a study on this topic).

It is probably unrealistic to think that we could assess every student's memory ability and use that estimate to create personalized spacing and testing practice schedules. However, even without a personalized profile for every student, different students might benefit from metacognitive

techniques to optimize learning. For example, if students could be taught to recognize the time when restudying or testing themselves on a particular item is difficult enough that it improves memory, but not too difficult that the item cannot be retrieved, this would allow for learning improvement for a wider range of students regardless of individual differences. Prior research supports the notion that memory in educational settings can be improved through metacognitive training (Metcalfe, Kornell, & Son, 2007) and monitoring for retrieval failures (Bahrick & Hall, 2005).

7. Conclusions

Our review sought to make sense of the conflicting and frequently bewildering results in the spacing literature. We began by pointing out that spacing-like effects can probably be produced by many different means, and that likewise there are things that people choose to do in our experiments that may obscure the "true" spacing effect. Before we can make sense of the spacing literature from the perspective of retrieval processes, we must understand how people act in our studies and how their strategic decisions affect memory. Often, peoples' decisions about how to study are obscured by the procedures we use in laboratory memory studies, and yet we have demonstrated repeatedly that the effects of these strategies may be larger than the "true" spacing effect itself. Most studies of spacing use word lists and ask people to study those words for a later memory test. Not only can rehearsal interact in unexpected ways with list order to produce larger or smaller spacing effects (Delaney & Verkoeijen, 2009), but people often change their study strategies as they encounter several lists (Delaney & Knowles, 2005). These encoding strategy differences can often obscure whether spacing effects "should" be present or not. For example, there is no reason why spacing effects should be absent on pure lists when people use rote rehearsal, but they generally are. The reason is because one of the "impostor" phenomena actually works *against* the spacing effect on such lists, and cancels it out. Hence, we think that controlling encoding strategies is going to be increasingly important if we want to make theoretical progress on spacing and testing.

The knowledge that one or more "impostor" spacing phenomena are present in a study casts doubt on the validity of the theoretical conclusions drawn from these studies. Our list of impostors can be understood as a potential set of critiques that reviewers can raise when judging whether a new paper should be used for making theoretical arguments, or whether it needs to be repeated with a cleaner design before we can trust the results. We understand that this exercise is fundamentally destructive in that it casts doubt on a large number of well-known empirical results. However, in

order to build a theory that can explain the spacing effect, we first need rules by which studies can be considered trustworthy or suspicious.

Next, we reviewed some existing major theoretical perspectives and tried to evaluate whether they could explain the "true" spacing effect. Various theories have been largely discarded because they fail to capture important aspects of the data (see Table 3 for some of these aspects). Our conclusion was that contextual variability and study-phase retrieval could probably be combined into a model that provides a successful account of most of the important spacing phenomena (cf. Raaijmakers, 2005). However, such a model requires quantitative tests before we can be confident that it works. Nonetheless, the verbal theory on which that model would be built makes a number of correct predictions that seemed to us to be counterintuitive, but that survived empirical tests. For example, it predicts dissociations with delay in free recall and recognition, which we have some evidence exist. It also predicts that directed forgetting will have a larger impact on spaced than on massed items, which was demonstrated by Sahakyan et al. (2008). In our view, a good theory ought to go beyond explaining (some of the) existing data, and make good predictions about future experiments. By that standard, the expanded version of Raaijmakers' SAM/REM account seems to be quite successful.

We next reviewed theoretical accounts of the testing effect in a fashion similar to our review of the spacing effect (see Table 4 for empirical phenomena). The transfer-appropriate processing notion is useful at a practical level, in that when intervening tests are used as an educational tool, it is helpful if characteristics of the intervening test mimic those of the final test. However, as a theoretical explanation of the testing effect, the transfer-appropriate processing account falls short because it has difficulties accommodating important findings from the testing effect literature. That leaves us with the desirable-difficulties framework and the associated concept of effortful retrieval. The idea that retrieval effort plays a pivotal role in the emergence of the testing effect is consistent with important findings from the testing-effect literature. However, why retrieval effort produces the testing effect is not yet clear. On the one hand, there are reasons to suspect that retrieval effort creates a memory advantage over a restudy episode because it adds extra information element to a memory trace (i.e., retrieval effect leads to encoding variability). On the other hand, retrieval effort may bring about its beneficial effect because it activates an elaborative structure of related concepts. It remains to be seen whether these accounts can be distinguished with respect to predictions about the testing effect.

While a hybrid encoding variability and retrieval account is appealing, especially since it postulates similar mechanisms for the spacing and testing effects, it is not entirely clear how such an account deals with the typical interaction between testing and the length of the retention interval. At first sight it seems to follow that testing should produce better final test

performance than restudying at any retention interval. However, the testing effect only emerges after long retention intervals; at short retention intervals restudying often leads to a better final test performance than testing. In contrast, the spacing effect can emerge even on time scales of seconds. Results like these may imply that consolidation-like mechanisms might provide a better account of existing data.

Finally, previous reviews of the spacing and testing literature have emphasized the importance of these phenomena in education (e.g., Dempster, 1988). We agree that implementing spacing and testing in school settings is a promising endeavor both practically and empirically. Borrowing from Daniel and Poole's (2009) educational philosophy, we advocate that spacing and testing research grows out of its current "memory first" approach and embraces a "pedagogical ecology" approach. A pedagogical approach has an interdisciplinary focus and observes students in context with the goal of identifying interactions that lead to various outcomes (Daniel & Poole). The careful control of the laboratory environment is critical for making theoretical progress, but we must be wary of assuming that the results of our laboratory studies can be applied. Learners and teachers both need to be aware of the benefits of spacing and testing, and to be guided to make choices that maximize learning in the long-term instead of minimizing the pain of training. Furthermore, it is worth assessing the value and frequency of existing educational practices by determining whether they encourage spaced practice, are sustainable, and are desired by students.

REFERENCES

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., III, & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*, 861–876.

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior*, *8*, 463–470.

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*, 1063–1087.

Anderson, J. R., & Bower, G. H. (1972). Configural properties in sentence memory. *Journal of Verbal Learning and Verbal Behavior*, *11*, 594–605.

Anderson, J. R., & Lebiere, C. (1998). *Atomic components of thought*. Hillsdale, NJ: Erlbaum.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory*, *2* (pp. 89–195). New York: Academic Press.

Baddeley, A. (1990). *Human memory: Theory and practice*. Needham Heights, MA: Allyn & Bacon.

Baddeley, A. D., & Longman, D. J. (1978). The influence of length and frequency of training session on the rate of learning to type. *Ergonomics*, *21*(8), 627–635.

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, *4*, 316–321.

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language*, *52*, 566–577.

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.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, *128*, 612–637.

Bäuml, K.-H. (1997). The list-strength effect: Strength-dependent competition or suppression? *Psychonomic Bulletin & Review*, *4*, 260–264.

Beins, B. C. (1993). Writing assignments in statistics classes encourage students to learn interpretation. *Teaching of Psychology*, *20*, 161–164.

Benjamin, A. S., & Bird, R. D. (2006). Metacognitive control of the spacing of study repetitions. *Journal of Memory and Language*, *55*, 126–137.

Benjamin, A. S., & Craik, F. I. M. (2001). Parallel effects of aging and time pressure on memory for source: Evidence from the spacing effect. *Memory & Cognition*, *29*, 691–697.

Bentin, S., & Feldman, L. B. (1990). The contribution of morphological and semantic relatedness to repetition priming at short and long lags: Evidence from Hebrew. *Quarterly Journal of Experimental Psychology*, *42A*, 693–711.

Bjork, R. A. (1989). Retrieval inhibition as an adaptive mechanism in human memory. In H. L. Roediger & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 309–330). 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., & Allen, T. W. (1970). The spacing effect: Consolidation or differential encoding? *Journal of Verbal Learning and Verbal Behavior*, *9*, 567–572.

Bobrow, S. A. (1970). Memory for words in sentences. *Journal of Verbal Learning and Behavior*, *9*, 363–372.

Bower, G. H., & Clark, M. C. (1969). Narrative stories as mediators for serial learning. *Psychonomic Science*, *14*, 181–182.

Brainerd, C. J., & Reyna, V. F. (1990). Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review*, *10*, 3–47.

Braun, K., & Rubin, D. C. (1998). The spacing effect depends on an encoding deficit, retrieval, and time in working memory: Evidence from once-presented words. *Memory*, *6*, 37–65.

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

Cahill, A., & Toppino, T. C. (1993). Young children's recognition as a function of the spacing of repetitions and the type of study and test stimuli. *Bulletin of the Psychonomic Society*, *31*, 481–484.

Carew, T. J., Pinsker, H. M., & Kandel, E. R. (1972). Long-term habituation of a defensive withdrawal reflex in aplysia. *Science*, *175*, 451–454.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *35*, 1563–1569.

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

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*, 268–276.

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, *14*, 474–478.
Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, *23*, 760–771.
Carpenter, S. K., Pashler, H., & Jones, J. (2008). The effect of retrieval practice on associative recall of word pairs. In: *Poster presented at the 49th Annual Meeting of the Psychonomic Society*. Chicago, IL.
Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, *13*, 826–830.
Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, *36*, 438–448.
Carrier, M. L., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*, 633–642.
Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*, 236–246.
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*, 354–380.
Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*, 1095–1102.
Challis, B. H. (1993). Spacing effects on cued-memory tests depend on level of processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 389–396.
Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language*, *61*, 153–170.
Chan, J. C. K., McDermott, K. B., & Roediger, H. L., III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*, 553–571.
Chase, W. G., & Ericsson, K. A. (1981). Skilled memory. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 141–189). Hillsdale, NJ: Lawrence Erlbaum Associates.
Ciccone, D. S., & Brelsford, J. W. (1976). Spacing repetitions in paired-associated learning: Experimenter versus subject control. *Journal of Experimental Psychology: Human Learning and Memory*, *2*, 446–455.
Cornell, E. H. (1980). Distributed study facilitates infants' delayed recognition memory. *Memory & Cognition*, *8*, 539–542.
Cornoldi, C., & Longoni, A. (1977). The MP-DP effect and the influence of distinct repetitions on recognition of random shapes. *Italian Journal of Psychology*, *4*, 65–76.
Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.
Cuddy, L. J., & Jacoby, L. L. (1982). When forgetting helps memory: An analysis of repetition effects. *Journal of Verbal Learning and Verbal Behavior*, *21*, 451–467.
Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14*, 215–235.
D'Agostino, P. R., & DeRemer, P. (1973). Item repetition in free and cued recall. *Journal of Verbal Learning and Verbal Behavior*, *11*, 54–58.
Daniel, D. B., & Poole, D. A. (2009). Learning for life: An ecological approach to pedagogical research. *Perspectives on Psychological Science*, *4*, 91–96.
Dannenbring, G. L., & Briand, K. (1982). Semantic priming and the word repetition effect in a lexical decision task. *Canadian Journal of Psychology*, *36*, 435–444.
Delaney, P. F., & Knowles, M. E. (2005). Encoding strategy changes and spacing effects in the free recall of unmixed lists. *Journal of Memory and Language*, *52*, 120–130.

Delaney, P. F., & Verkoeijen, P. P. J. L. (2009). Rehearsal strategies can enlarge or diminish the spacing effect: Pure versus mixed lists and encoding strategy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1148–1161.

Dellarosa, D., & Bourne, L. E. (1985). Surface form and the spacing effect. *Memory & Cognition*, *13*, 529–537.

Dempster, F. N. (1988). Informing classroom practice: What we know about several task characteristics and their effects on learning. *Contemporary Educational Psychology*, *13*, 254–264.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 317–344). San Diego, CA: Academic Press.

DeZazzo, J., & Tully, T. (1995). Dissection of memory formation: From behavioral pharmacology to molecular genetics. *Trends in Neurosciences*, *18*, 212–218.

Diana, R. A., & Reder, L. M. (2005). The list strength effect: A contextual competition account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 1289–1302.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, *84*, 795–805.

Drevenstedt, J., & Bellezza, F. S. (1993). Memory for self-generated narration in the elderly. *Psychology and Aging*, *8*, 187–192.

Durgunoğlu, A. Y., & Roediger, H. L. (1987). Test differences in accessing bilingual memory. *Journal of Memory and Language*, *26*, 377–391.

Eich, J. E. (1980). The cue-dependent nature of state-dependent retrieval. *Memory & Cognition*, *8*, 157–173.

Ericsson, K. A., Delaney, P. F., Weaver, G. A., & Mahadevan, S. (2004). Uncovering the structure of a memorist's superior "basic" memory capacity. *Cognitive Psychology*, *49*, 191–237.

Gartman, L. M., & Johnson, N. F. (1972). Massed versus distributed repetition of homographs: A test of the differential-encoding hypothesis. *Journal of Verbal Learning and Verbal Behavior*, *11*, 801–808.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *40*, 104.

Glanzer, M., & Duarte, A. (1971). Repetition between and within languages in free recall. *Journal of Verbal Learning and Verbal Behavior*, *10*, 625–630.

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.

Glenberg, A. M. (1977). Influences of retrieval processes on the spacing effect in free recall. *Journal of Experimental Psychology: Human Learning and Memory*, *3*, 282–294.

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition*, *7*, 95–112.

Glenberg, A. M., & Smith, S. M. (1981). Spacing repetitions and solving problems are not the same. *Journal of Verbal Learning and Verbal Behavior*, *20*, 110–119.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*, 392–399.

Glover, J. A., & Corkill, A. J. (1987). Influence of paraphrased repetitions on the spacing effect. *Journal of Educational Psychology*, *79*, 198–199.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*, 325–331.

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 371–377.

Greene, R. L. (1990). Spacing effects on implicit memory tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 1004–1011.

Greene, R. L., & Stillwell, A. M. (1995). Effects of encoding variability and spacing on frequency discrimination. *Journal of Memory and Language*, *34*, 468–478.

Hall, J. W. (1992a). Unmixing effects of spacing on free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 608–614.

Hall, J. W. (1992b). Recall of lists of prolonged and repeated (spaced) words. *Bulletin of the Psychonomic Society*, *18*, 183–186.

Hanawalt, N. G., & Tarr, A. G. (1961). The effect of recall on recognition. *Journal of Educational Psychology*, *62*, 361–367.

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 77–99). Hillsdale, NJ: Erlbaum.

Hintzman, D. L., & Block, R. A. (1973). Memory for the spacing of repetitions. *Journal of Experimental Psychology*, *99*, 70–74.

Hintzman, D. L., Summers, J. J., & Block, R. A. (1975). Spacing judgments as an index of study-phase retrieval. *Journal of Experimental Psychology: Human Learning and Memory*, *104*, 31–40.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 302–313.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *10*, 562–567.

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. (1983). Perceptual enhancement: Persistent effects of an experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 21–38.

Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of Consumer Research*, *30*, 138–149.

Jensen, T. D., & Freund, J. S. (1981). Persistence of the spacing effect in incidental free recall: The effect of external list comparisons and intertask correlations. *Bulletin of the Psychonomic Society*, *18*, 183–186.

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, *101*, 621–629.

Johnston, W. A., Coots, J. H., & Flickinger, R. G. (1972). Controlled semantic encoding and the effect of repetition lag on free recall. *Journal of Verbal Learning and Verbal Behavior*, *11*, 784–788.

Johnston, W. A., & Uhl, C. N. (1976). The contributions of encoding effort and variability to the spacing effect on free recall. *Journal of Experimental Psychology: Human Learning and Memory*, *2*, 153–160.

Kahana, M. J., & Howard, M. W. (2005). Spacing and lag effects in free recall of pure lists. *Psychonomic Bulletin & Review*, *12*, 159–164.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 528–558.

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory*, *17*, 471–479.

Karpicke, J. D., & Roediger, H. L., III. (2007). 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.

Kausler, D. H., Wiley, J. G., & Phillips, P. L. (1990). Adult age differences in memory for massed and distributed repeated actions. *Psychology and Aging*, *5*, 530–534.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, *49*, 294–304.

Kirsner, K., Smith, M. C., Lockhart, R. S., King, M. L., & Jain, M. (1984). The bilingual lexicon: Language-specific units in an integrated network. *Journal of Verbal Learning and Verbal Behavior*, *23*, 519–539.

Kolers, P. A. (1966). Interlingual facilitation of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *5*, 314–319.

Konopak, B. C., Martin, S. H., & Martin, M. A. (1990). Using a writing strategy to enhance sixth-grade students' comprehension of content material. *Journal of Reading Behavior*, *22*, 19–37.

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

Kuo, T., & Hirshman, E. (1996). Investigations of the testing effect. *American Journal of Psychology*, *109*, 451–464.

Landauer, T. K. (1969). Reinforcement as consolidation. *Psychological Review*, *76*, 82–96.

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

Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 322–336.

Mammarella, N., Avons, S. E., & Russo, R. (2004). A short-term perceptual priming account of spacing effects in explicit cued-memory tasks for unfamiliar stimuli. *European Journal of Cognitive Psychology*, *16*, 387–402.

Mammarella, N., Russo, R., & Avons, S. E. (2002). Spacing effects in cued-memory tasks for unfamiliar faces and nonwords. *Memory & Cognition*, *30*, 1238–1251.

Mannes, S. M., & Kintsch, W. (1987). Knowledge organization and text organization. *Cognition and Instruction*, *4*, 91–115.

Maskarinec, A. S., & Thompson, C. P. (1976). The within-list distributed practice effect: Tests of the varied context and varied encoding hypothesis. *Memory & Cognition*, *4*, 741–746.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*, 494–513.

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 371–385.

McNamara, T. P. (1992). Theories of priming: I. Associative distance and lag. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1173–1190.

Melton, A. W. (1967). Repetition and retrieval from memory. *Science*, *158*(3800), 532.

Menzel, R., Manz, G., Menzel, R., & Greggers, U. (2001). Massed and Spaced learning in honeybees: The role of CS, US, the intertribal interval, and the test interval. *Learning & Memory*, *8*, 198–208.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General*, *132*, 530–542.

Metcalfe, J., Kornell, N., & Son, L. K. (2007). A cognitive-science based program to enhance study efficacy in a high and low-risk setting. *European Journal of Cognitive Psychology*, *19*, 743–768.

Murdock, B. (2003). The mirror effect and the spacing effect. *Psychonomic Bulletin & Review*, *10*, 570–588.

Murnane, K., & Shiffrin, R. M. (1991). Word repetitions in sentence recognition. *Memory & Cognition*, *19*, 119–130.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*, 226–254.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 264–336). Hillsdale, NJ: Erlbaum.

Nelson, D. L., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review*, *105*, 299–324.

Norman, K. A. (2002). Differential effects of list strength on recollection and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1083–1094.

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology*, *74*, 18–22.

Offner, M. (1911). *Das Gedaächtnis: Die Ergebnisse der experimentellen Psychologie und ihre Anwendung in Unterricht und Erziehung*. Berlin, Germany: Reuther & Reichard.

Paivio, A. (1974). Spacing of repetitions in the incidental and intentional free recall of pictures and words. *Journal of Verbal Learning and Verbal Behavior*, *13*, 497–511.

Paivio, A., & Yuille, J. C. (1969). Changes in associative strategies and paired-associate learning over trials as a function of word imagery and type of learning set. *Journal of Experimental Psychology*, *79*, 458–463.

Parkin, A. J., Gardiner, J. M., & Rosser, R. (1995). Functional aspects of recollective experience in face recognition. *Consciousness & Cognition*, *4*, 387–398.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, *14*, 187–193.

Pastötter, B., & Bäuml, K.-H. (2007). The crucial role of postcue encoding in directed forgetting and context-dependent forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 977–982.

Pastötter, B., & Bäuml, K.-H. (2010). Amount of postcue encoding predicts amount of directed forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 54–65.

Pavlik, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, *29*, 559–586.

Peterson, L. R., Hillner, K., & Saltzman, D. (1962). Time between pairings and short-term retention. *Journal of Experimental Psychology*, *64*, 550–551.

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

Postman, L., & Knecht, K. (1983). Encoding variability and retention. *Journal of Verbal Learning and Verbal Behavior*, *22*, 133–152.

Price, H. L., Connolly, D. A., & Gordon, H. M. (2006). Children's memory for complex autobiographical events: Does spacing of repeated instances matter? *Memory*, *14*, 977–989.

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*, 437–447.

Quitadamo, I. J., & Kurtz, M. J. (2007). Learning to improve: Using writing to increase critical thinking performance in general education biology. *Life Sciences Education*, *6*, 140–154.

Raaijmakers, J. G. W. (2005). Modeling implicit and explicit memory. In C. Izawa & N. Ohta (Eds.), *Human Learning and Memory: Advances in Theory and Application* (pp. 85–105). Mahwah, NJ: Lawrence Erlbaum Associates.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search in associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory*, *14* (pp. 207–262). New York: Academic Press.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134.

Ratcliff, R., Clark, S., & Shiffrin, R. M. (1990). The list-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 163–178.

Rea, C. P., & Modigliani, V. (1987). The spacing effect in 4- to 9-year-old children. *Memory & Cognition*, *15*, 436–443.

Reddy, B. G., & Bellezza, F. S. (1983). Encoding specificity in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 167–174.

Reed, A. V. (1977). Quantitative prediction of spacing effects in learning. *Journal of Verbal Learning and Verbal Behavior*, *16*, 693–698.

Rickard, T. C., Lau, J. S., & Pashler, H. (2008). Spacing and the transition from calculation to retrieval. *Psychonomic Bulletin & Review*, *15*, 656–661.

Roediger, H. L., Agarwal, P. K., Kang, S. H. K., & Marsh, E. J. (2010). Benefits of testing memory: Best practices and boundary conditions. In G. M. Davies & D. B. Wright (Eds.), *New frontiers in applied memory*. (pp. 13–49). Brighton UK: Psychology Press.

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

Rohrer, D. (2009). The effects of spacing and mixing practice problems. *Journal for Research in Mathematics Education*, *40*, 4–17.

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology*, *20*, 1209–1224.

Rose, R. J. (1984). Processing time for repetitions and the spacing effect. *Canadian Journal of Experimental Psychology*, *83*, 537–550.

Rose, R. J., & Rowe, E. J. (1976). Effects of orienting task and spacing of repetitions on frequency judgments. *Journal of Experimental Psychology: Human Learning and Memory*, *2*, 142–152.

Ross, B. H., & Landauer, T. K. (1978). Memory for at least one of two items: Test and failure of several theories of spacing effects. *Journal of Verbal Learning and Verbal Behavior*, *17*, 669–680.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, *89*, 63–77.

Russo, R., & Mammarella, N. (2002). Spacing effects in recognition memory: When meaning matters. *European Journal of Cognitive Psychology*, *14*, 49–59.

Russo, R., Mammarella, N., & Avons, S. E. (2002). Spacing effects in cued memory tasks for unfamiliar faces and nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 819–829.

Russo, R., Parkin, A. J., Taylor, S. R., & Wilks, J. (1998). Revising current two-process accounts of spacing effects in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 161–172.

Sahakyan, L., Waldum, E. R., Benjamin, A. S., & Bickett, S. P. (2009). Where is the forgetting with list-method directed forgetting in recognition? *Memory & Cognition*, *37*, 464–476.

Sahakyan, L., & Delaney, P. F. (2003). Can encoding differences explain the benefits of directed forgetting in the list-method paradigm? *Journal of Memory and Language*, *48*, 195–201.

Sahakyan, L., & Delaney, P. F. (2005). Directed forgetting in incidental learning and recognition testing: Support for a two-factor account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 789–801.

Sahakyan, L., Delaney, P. F., & Kelley, C. M. (2004). Self-evaluation as a moderating factor in strategy change in directed forgetting benefits. *Psychonomic Bulletin & Review*, *11*, 131–136.

Sahakyan, L., Delaney, P. F., & Waldum, E. R. (2008). Intentional forgetting is easier after two "shots" than one. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 408–414.

Sahakyan, L., & Goodmon, L. B. (2007). The influence of directional associations on directed forgetting and interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 1035–1049.

Sahakyan, L., & Kelley, C. M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1064–1072.

Scharf, M. T., Woo, N. H., Lattal, K. M., Young, J. Z., Nguyen, P. V., & Abel, T. (2002). Protein synthesis is required for the enhancement of long-term potentiation and long-term memory by spaced training. *Journal of Neurophysiology*, *87*, 2770–2777.

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.

Seabrook, R., Brown, G. D. A., & Solity, J. E. (2005). Distributed and massed practice: From laboratory to classroom. *Applied Cognitive Psychology*, *19*, 107–122.

Shaughnessy, J. J. (1976). Persistence of the spacing effect in free recall under varying incidental learning conditions. *Memory & Cognition*, *4*, 369–377.

Shaughnessy, J. J., Zimmerman, J., & Underwood, B. J. (1972). Further evidence on the MP-DP effect in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *11*, 1–12.

Shiffrin, R. M. (1970). Forgetting, trace erosion or retrieval failure? *Science*, *168*, 1601–1603.

Simon, D. A., & Bjork, R. A. (2001). Metacognition in motor learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 907–912.

Slamecka, N. J., & Katsaiti, L. T. (1988). Normal forgetting of verbal lists as a function of prior testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 716–727.

Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 460–471.

Smith, S. M. (1984). A comparison of two techniques for reducing context-dependent forgetting. *Memory & Cognition*, *12*, 477–482.

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

Smith, M. C., Theodor, L., & Franklin, P. E. (1983). The relationship between contextual facilitation and depth of processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 697–712.

Son, L. K. (2004). Spacing one's study: Evidence for a metacognitive control strategy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 601–604.

Sperber, R. D. (1974). Developmental changes in effects of spacing of trials in retardate discrimination learning and memory. *Journal of Experimental Psychology*, *103*, 204–210.

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

Stern, L. D., & Hintzman, D. L. (1979). Spacing and retention of synonyms. *Bulletin of the Psychonomic Society*, *13*, 363–366.

Stigler, J. W., Fuson, K. C., Ham, M., & Kim, M. S. (1986). An analysis of addition and subtraction word problems in American and Soviet elementary mathematics textbooks. *Cognition and Instruction*, *3*, 153–171.

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.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L., III. (2007). Expectation of a final cumulative test enhances long-term retention. *Memory & Cognition*, *35*, 1007–1013.

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

Tan, L., & Ward, G. (2000). A recency-based account of the primacy effect in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1589–1625.

Thios, S. J. (1972). Memory for words in repeated sentences. *Journal of Verbal Learning and Verbal Behavior*, *11*, 789–793.

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

Toppino, T. C. (1991). The spacing effect in young children's free recall: Support for automatic-process explanations. *Memory & Cognition*, *19*, 159–167.

Toppino, T. C. (1993). The spacing effect in preschool children's free recall of pictures and words. *Bulletin of the Psychonomic Society*, *31*, 27–30.

Toppino, T. C., & Bloom, L. C. (2002). The spacing effect, free recall, and two-process theory: A closer look. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 437–444.

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

Toppino, T. C., Cohen, M. S., Davis, M., & Moors, A. (2009). Metacognitive control over the spacing of practice: When is spacing preferred? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1352–1358.

Toppino, T. C., & DeMesquita, M. (1984). Effects of spacing repetitions on children's memory. *Journal of Experimental Child Psychology*, *37*, 637–648.

Toppino, T. C., & DiGeorge, W. (1984). The spacing effect in free recall emerges with development. *Memory & Cognition*, *12*, 118–122.

Toppino, T. C., Kasserman, J. E., & Mracek, W. A. (1991). The effect of spacing repetitions on the recognition memory of young children and adults. *Journal of Experimental Child Psychology*, *51*, 123–138.

Toppino, T. C., & Schneider, M. A. (1999). The mix-up regarding mixed and unmixed lists in spacing-effect research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1071–1076.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *6*, 175–184.

Tulving, E., & Hastie, R. (1972). Inhibition effects of intralist repetition in free recall. *Journal of Experimental Psychology*, *92*, 297–304.

Underwood, B. J. (1969). Some correlates of item repetition in free recall learning. *Journal of Verbal Learning and Verbal Behavior*, *9*, 573–580.

Underwood, B. J. (1970). The spacing effect: Additions to the theoretical and empirical puzzles. *Memory & Cognition*, *4*, 391–400.

Vander Linde, E., Morrongiello, B. A., & Rovee-Collier, C. (1985). Determinants of retention in 8-week-old infants. *Developmental Psychology*, *21*, 601–613.

Verkoeijen, P. P. J. L., & Bouwmeester, S. (2008). Modeling bimodality in spacing effect data. *Journal of Memory and Language*, *59*, 545–555.

Verkoeijen, P. P. J. L., & Delaney, P. F. (2008). Rote rehearsal and spacing effects in the free recall of pure and mixed lists. *Journal of Memory and Language*, *58*, 35–47.

Verkoeijen, P. P. J. L., & Delaney, P. F. (2010a). *The testing effect depends on the type of encoding strategy*. Manuscript in preparation.

Verkoeijen, P. P. J. L., & Delaney, P. F. (2010b). *The effect of testing on memory for category lists*. Manuscript in preparation.

Verkoeijen, P. P. J. L., Rikers, R. M. J. P., Pecher, D., Zeelenberg, R., & Schmidt, H. G. (2010). *Evidence against the semantic priming account of spacing effects in recognition memory*. Unpublished manuscript.

Verkoeijen, P. P. J. L., Rikers, R. M. J. P., & Schmidt, H. G. (2004). Detrimental influence of contextual change on spacing effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 796–800.

Verkoeijen, P. P. J. L., Rikers, R. M. J. P., & Schmidt, H. G. (2005). Limitations to the spacing effect: Demonstration of an inverted U-shaped relationship between interrepetition spacing and free recall. *Experimental Psychology*, *52*, 257–263.

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition*, *109*, 163–167.

Wagner, A. D., Maril, A., & Schacter, D. L. (2000). Interactions between forms of memory: When priming hinders new episodic learning. *Journal of Cognitive Neuroscience*, *12*, 52–60.

Waugh, N. C. (1962). The effect of intralist repetition on free recall. *Journal of Verbal Learning and Verbal Behavior*, *1*, 95–99.

Waugh, N. C. (1963). Immediate memory as a function of repetition. *Journal of Verbal Learning and Verbal Behavior*, *2*, 107–112.

Waugh, N. C. (1967). Presentation time and free recall. *Journal of Experimental Psychology*, *73*, 39–44.

Waugh, N. C. (1970). On the effective duration of a repeated word. *Journal of Verbal Learning and Verbal Behavior*, *5*, 587–595.

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*, 571–580.

Wiley, J., & Voss, J. F. (1996). The effects of "playing" historian on learning in history. *Applied Cognitive Psychology*, *10*, 63–72.

Wilson, W. P. (1976). Developmental changes in the lag effect: An encoding hypothesis for repeated word recall. *Journal of Experimental Child Psychology*, *22*, 113–122.

Wolfe, C. R., Reyna, V. F., & Brainerd, C. J. (2005). Fuzzy-trace theory: Implications for transfer in teaching and learning. In J. P. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 53–88). Greenwich, CT: Information Age Publishing.

Wright, J., & Brelsford, J. (1978). Changed in the spacing effect with instructional variables in free recall. *American Journal of Psychology*, *91*, 631–643.

Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1992). Tests of the list-strength effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 345–355.

Young, J. L. (1971). Reinforcement-test intervals in paired-associate learning. *Journal of Mathematical Psychology*, *8*, 58–81.

Zeelenberg, R., & Pecher, D. (2002). False memories and lexical decision: Even twelve primes do not cause long-term semantic priming. *Acta Psychologica*, *109*, 269–284.

Zeelenberg, R., Pecher, D., & Tabbers, H. K. (2008). *The effect of testing on memory: Does enhanced retention transfer to new test situations?* Poster presented at the 49th Annual Meeting of the Psychonomic Society, Chicago, IL.

Zimmerman, J. (1975). Free recall after self-paced study: A test of the attention explanation of the spacing effect. *American Journal of Psychology*, *88*, 277–291.