A Test of the Efficiency of Study and a Study on the Efficacy of Tests

Mario de Jonge

Cover photo by Stefan van der Kamp (Whaleshark Productions)
Cover layout by Nikki Vermeulen
Printed by Ridderprint B.V., Ridderkerk, the Netherlands

ISBN: 978-90-5335-860-3

A Test of the Efficiency of Study and a Study on the Efficacy of Tests

Een test van hoe efficiënt het studeren is
en een studie naar de effectiviteit van het testen

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus
Prof.dr. H.A.P. Pols
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
vrijdag 20 juni 2014 om 11.30 uur

door

Mario Olivier de Jonge
geboren te Vlissingen

ERASMUS UNIVERSITEIT ROTTERDAM

Promotiecommissie

Promotor

Prof.dr. R.M.J.P. Rikers

Overige leden

Prof.dr. T. van Gog

Prof.dr. L. Kester

Dr. P.P.J.L. Verkoeijen

Copromotor

Dr. H.K. Tabbers

Contents

Chapter 1

General Introduction

In an episode of the popular 1980's Flemish children's television show called "*het Liegebeest"* (loosely translated, *[liar, liar, pants on fire]*), two of the main characters, Corneel and Carolus, are engaged in a somewhat silly conversation about learning and forgetting. Corneel argues that learning is basically a very foolish thing to do. Carolus, on the other hand, points out to Corneel that had he not learned anything at all, he would probably have been even dumber than he already is. Corneel is not taken aback by Carolus's rebuttal and he insists that learning is unwise. He argues that the more you learn, the more you forget. Hence, if you do not learn anything at all, there is nothing for you to forget! In other words, Corneel's recommendation for educational practice suggests that, to effectively prevent forgetting from occurring, it might be optimal not to commit anything to memory in the first place. Silly it may be, however, the logic is sound. Research has shown that most of what is learned is forgotten relatively quickly after learning has taken place (e.g., Ebbinghaus, 1885/1964; Wixted & Ebbesen, 1991).[1] Thus, it may be no surprise that students are often reluctant to engage in strenuous learning activities when the benefits of their efforts are so short-lived. In some respects, learning might rightfully be regarded as a complete waste of time. In the present thesis, I will not argue otherwise. However, I will try and provide some helpful recommendations for learners on how to waste time more efficiently and effectively.

The situation of interest in the present thesis is one where students have a limited amount of time to learn by themselves a certain amount of information (ranging from a list of foreign vocabulary words to more complex materials like texts). The question of interest is how students can get the most out of their limited study time. On the one hand, we will consider the efficiency of certain manipulations and focus on what conditions result in fast acquisition during initial learning. Importantly, however, it has been emphasized in the literature that manipulations that speed up initial learning often fail to support post-learning retention (e.g., Bjork, 1994, 1999). In other words, what is efficient might not always prove to be effective in the long run. In order to establish useful recommendations for educational practice it seems vital not only to consider the efficiency, but also the efficacy of learning. Accordingly, in the studies presented herein, we investigated promising ways for improving both the initial learning as well as the long-term retention of information.

[1] The consensus in the literature is that most forgetting occurs relatively quickly after learning has taken place. However, based on findings from earlier research on long-term retention it has also been suggested that, when information is very well-learned, it may enter a state of "permastore" and forgetting might not occur at all (Bahrick, 1984). More recent research on the mathematical form of forgetting has challenged the idea of permanent storage of information, and shown that, even for the Bahrick (1984) retention data, forgetting does not necessarily level off to an asymptote above zero (Wixted, 2004).

There are two important strategies that are singled out in the literature as holding great promise for improving learning and retention. One of these strategies concerns the proper distribution of the available study time, and the other strategy pertains to the use of opportunities for retrieval practice (taking tests) during learning. Since these two strategies have much in common, both in terms of theoretical and practical considerations, it is no surprise that they are often jointly discussed in the literature (e.g., Carpenter & Delosh, 2005; Delaney, Verkoeijen, & Spirgel, 2010; Roediger, 2013; Roediger & Pyc, 2012). In fact, in a recent review of the literature, the distributed practice effect and the retrieval practice effect are referred to as being each other's first cousin (Delaney et al., 2010). Before outlining the studies in the present thesis, I will first acquaint the reader with both cousins, and I will discuss how they can both contribute to efficient and durable learning.

The Distributed Practice Effect

The gist of the distributed practice effect has been eloquently depicted in the following limerick by Ulrich Neisser (as quoted by Bjork, 1988, p. 399): "*You can get a good deal from rehearsal, / If it just has the proper dispersal. / You would just be an ass, / To do it en masse: / Your remembering would turn out much worsal."* Neisser's verse refers to several important findings in the literature on distributed practice. First of all, not surprisingly, relative to studying materials just once, additional rehearsal will generally lead to improved recall performance (e.g., Stubin, Heimer, & Tatz, 1970). However, as is also clear from the verse, rehearsal alone is not enough. It matters a great deal how exactly additional rehearsal is arranged. Specifically, repetitions that are dispersed over time (i.e., *spaced*) are known to be more effective than immediate (i.e., *massed*) repetitions (e.g., Calfee, 1968; Greeno, 1964). Some authors have even presented evidence for the idea that the more spaced apart two presentations of the same item are, the more effective they become (e.g., Melton, 1970). However, more recent evidence indicates that the situation with regard to the optimal spacing of repetitions is more complicated than that.

Studies on the distributed practice effect indicate that there is an inverted U-shape relationship between the degree of spacing during initial learning and subsequent recall performance. That is, increasing the spacing of repetitions, first increases, but then decreases the probability of later recall (e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Glenberg, 1976). To complicate matters further, the optimal degree of spacing does not appear to be stable, and shifts depending on the retention interval of interest. For instance, in Cepeda et al.'s (2008) study, for a retention interval of one week, the optimal spacing interval was in the order of days.

However, for a retention interval of about one year, the optimal spacing interval was in the order of weeks. Thus, to put it in the words of Cepeda and colleagues: *"If you want to know the optimal distribution of your study time, you need to decide how long you wish to remember something."* Unfortunately, for most students this does not pose a big dilemma. Students might not be primarily concerned with long-term retention, but they would rather just be able to recall something on an upcoming exam. This has also been referred to as one of the great tragedies of modern education (Anderson, 1995). Although last-minute cramming can be an efficient strategy for enhancing short-term exam performance, it might not be the most effective strategy for enhancing long-term retention.

Theoretical Accounts of the Distributed Practice Effect

Some of the most popular theoretical accounts for explaining the distributed practice effect include *the deficient processing*, *the encoding variability,* and *the study phase retrieval theory*.

First, the deficient processing theory suggests that massed repetitions are not very potent learning events compared to spaced repetitions, because massed repetitions result in deficient processing. For instance, Greene (1989) argued that, beyond the first presentation of an item, a subsequent massed repetition may receive less rehearsal time and processing resources, because the learner mistakenly thinks that an item is already well-learned. However, spaced repetition can diminish this apparent overconfidence, because items will seem less and less familiar during subsequent repetitions as the spacing interval increases. Thus, for spaced repetitions, learners will be more inclined to devote processing resources and rehearsal time.

Second, the encoding variability theory suggests that the spacing of repetitions can be beneficial, because spaced repetitions will result in more variable encoding than massed repetitions (e.g., Glenberg, 1976, 1979; Madigan, 1969; Melton, 1970). For instance, Glenberg (1976) suggested that differential encoding occurs as a result of changes in context. For massed presentations, the change in context from one presentation to the next will be negligible. However, as the spacing interval between two presentations increases, the change in context from one presentation to the next will increase. The resulting memory representation of an item presented at a spaced interval is assumed to be richer in contextual elements and more elaborated. Furthermore, successful retrieval of a learned target item on a subsequent retention test is assumed to be partly dependent on the overlap between the context during encoding (i.e., the study context) and the context during retrieval (i.e., the test context). The idea that contextual overlap between encoding and retrieval events can facilitate success on the latter event is also known as the *encoding specificity principle* (Tulving &

Thomson, 1973). For items encoded in variable contexts (i.e., spaced items) chances are better that the context of a later test matches the context during encoding, and this in turn increases the chance that a target item will be successfully retrieved on the test.

Third and last, the study phase retrieval theory suggests that the second occurrence of an item can also serve as a reminder (i.e., study phase retrieval) of the first occurrence (Greene, 1989; Raaijmakers, 2003). For one thing, as we will discuss in more detail shortly, retrieval of information during learning can facilitate subsequent retrieval (cf. the retrieval practice effect). To some extent, spaced repetitions might also provide learners with a moderate form of retrieval practice (Greene, 1989). Furthermore, if the originally stored memory trace for a target item is successfully retrieved during the subsequent presentation of the item, information can be added to the originally stored trace (e.g., Raaijmakers, 2003). The adding of information to a trace, contextual or otherwise, will then result in a richer, more elaborate, memory trace. However, if the spacing interval becomes too long, study phase retrieval may not be successful, resulting in the formation of a new memory trace rather than adding information to the original trace. One appealing characteristic of the study phase retrieval account is that it may also explain why increasing the spacing interval between two presentations of an item, first increases, but then decreases subsequent recall performance. That is, if the spacing interval gets too long, the second occurrence of an item may not remind the learner of the first occurrence and study phase retrieval may not be successful.

The Retrieval Practice Effect

Another effective strategy for improving learning and retention is having learners take tests during learning (see Roediger & Karpicke, 2006a for a review). Although investigations of the benefits of testing during learning date back to the beginning of the 20th century (e.g., Gates, 1917; Kühn, 1914), it was not until recently that the retrieval practice effect gained renewed interest among cognitive and educational psychologists. To illustrate, in a recent review of the literature, Rawson and Dunlosky (2012) listed no less than 82 studies on the effects of testing published in the years 2000 to 2010 alone, roughly doubling the amount of work published in the preceding century. Like the distributed practice effect, the effect of retrieval practice is known to be a very robust phenomenon that has been demonstrated using a wide variety of learning materials, ranging from simple verbal materials (e.g., Wheeler, Ewers, & Buonanno, 2003) to complex materials like short texts (e.g., Roediger & Karpicke, 2006b). Given the robustness of the effect, and the rapidly growing body of work in this area, it seems fair to say that the first cousin of the distributed practice effect deserves a verse of its own. I have taken it upon myself to write the following limerick, in much the same spirit as Ulrich Neissers' verse, as a description of the retrieval practice effect: *A test is a great intervention, / When learning for long-term retention. / But according to Roddy,*[2]*/ Repeated study, / Will just result in a lot of forgetchen.*

One thing which should be clear from the verse is that repeated study, although being an effective strategy for getting information into memory, may not be the most effective strategy for keeping information available for future usage. That is, following conditions of repeated study, information is forgotten relatively quickly compared to other more demanding learning activities, like for instance repeated testing (e.g., Roediger & Karpicke, 2006a). One important implication of the results from studies on the retrieval practice effect is that regarding a test as a neutral measurement device (which is probably the predominant view in most educational settings) is an oversimplified view (Roediger & Karpicke, 2006b). Akin to the Heisenberg uncertainty principle, probing memory as to obtain a measure of the current state of affairs can at the same time change the state of affairs (Roediger & Karpicke, 2006a; Spellman & Bjork, 1992). That is, successful retrieval of information from memory increases the chance that retrieval of the same information will again be successful in the future. Thus, a test can be used as a learning intervention, complementing rather than just assessing the learning process.

Theoretical Accounts of the Retrieval Practice Effect

Two of the most popular theoretical accounts for explaining the retrieval practice effect are the *retrieval hypothesis*, and the *transfer appropriate processing account*.

The retrieval hypothesis suggests that some aspect of the retrieval process itself is responsible for the retention benefit often observed for tested items (Dempster, 1996; Roediger & Karpicke, 2006a). There are a variety of ideas and theories about which processes learners engage in during the act of retrieval from memory. For instance, one version of the retrieval account suggests that taking tests during learning invokes more elaborative processing compared to less demanding learning strategies like repeatedly studying (e.g., Carpenter & DeLosh, 2006; Glover, 1989). This *elaborative retrieval account* suggests that retrieval from memory can produce an elaboration of an existing memory trace and increase the variability of encoded information (McDaniel & Masson, 1985). Note that there is some conceptual overlap here between the elaborative

[2] Roddy refers to renowned psychologist Henry "Roddy" Roediger III (Washington University, St. Louis) who was, for a large part, responsible for instigating the revival of research on the retrieval practice effect over the last decade.

retrieval account of the retrieval practice effect and the encoding variability account of the distributed practice effect. That is, both accounts assume that information is added to an existing memory trace resulting in a richer, more elaborate, memory representation. Another version of the retrieval hypothesis is the *effortful retrieval hypothesis* (e.g., Pyc & Rawson, 2009). This hypothesis is partly based on the *depth of processing framework* (Craik & Tulving, 1975). It is assumed that the durability of a memory trace is largely dependent on the depth of processing. Retrieval from memory can produce such deep processing, because retrieval requires a great deal of semantic involvement from the learner.

Importantly, the two versions of the retrieval hypothesis described above are not necessarily mutually exclusive, and to a certain degree they may be describing more or less the same processes (Roediger & Karpicke, 2006a). Also, note that the types of processing implicated by both accounts (i.e., elaboration, and deep effortful processing) are not limited to retrieval from memory per se. Elaboration and deep processing are in fact very general types of processing and the hypotheses described above simply suggest that more of it occurs during retrieval practice (compared to less demanding processing types like passive study).

Another theoretical account of the testing effect, the transfer-appropriate processing account (Morris, Bransford, & Franks, 1977), suggests that the benefits of testing might be the result of the overlap between the type of processing engaged in during initial learning and the type of processing required by a later test. Note that this idea bears a striking similarity to the idea of encoding specificity, discussed earlier with regards to the distributed practice effect. In fact, the two accounts have also been referred to as being conceptual twins (Lockhart, 2002). One subtle difference being that the encoding specificity principle suggests that successful retrieval is in part dependent on the contextual overlap during encoding and retrieval, while the transfer appropriate processing account emphasizes the importance of overlap in types of processing rather than the context in which the processing takes place. Repeated study of information might be ineffective for enhancing later test performance because, during repeated study, the learner practices with types of processing that are not transfer-appropriate. To be successful at retrieving something on a later test, learners should focus on practicing types of processing that most closely resemble the processing required by the later test (i.e., retrieving from memory).

Overview of the Thesis

To sum up, proper study time distribution and providing learners with opportunities for retrieval practice are two very promising strategies for improving learning and retention. The present thesis comprises a number of studies in which we investigated the extent to which these strategies can contribute to efficient and effective learning. As noted, the situation of interest is one where students have only a limited amount of time to study, and we were interested in how students might get the most out of such a short single-session learning period.

In the first part of the thesis, we investigated how study time should be distributed within a single (short) learning session to be optimally effective. Note that, based on the existing literature on the distributed practice effect, it is complicated to make any useful recommendations about optimal study time distribution. That is, for multi-session learning, the research by Cepeda et al. (2008) suggests that what is optimal depends on the retention interval of interest. Hence, for the multi-session learning situation, there does not appear to be a single best way of distributing practice to achieve optimal retention. The same appears to be true, although to a lesser extent, for the retention of information over short intervals (in the order of seconds or minutes) following single-session learning (Glenberg, 1976). Thus, the results from these studies leave open the question of what is most efficient or effective with regards to the distribution of study time.

In our experiments on optimizing study time distribution, we took a somewhat different approach compared to some of the previous studies discussed. We asked ourselves the question, given a limited amount of study time, in how many presentations should the available time be divided to be optimally effective? For instance, imagine someone is given just a couple of minutes to study a list of word pairs in a standard paired associate learning paradigm. Word pairs are (repeatedly) presented to the learner one by one and afterwards a test is given to assess recall performance. On the one hand, one could take the extreme position that it does not matter at all how one divides the total study time (e.g., Bugelski, 1962; Murdock, 1960). However, one could also argue that perhaps having as many presentations as possible might be optimal. For instance, one could argue that having a large number of presentations during learning can result in many successful study-phase retrievals (reminders of previous occurrences if a target item). This could in turn result in stronger memory traces. If this is the case, then study time should perhaps preferably be distributed in such a way that the learner receives many presentations of each item at a very fast rate. At the same time, one could also argue for fewer, but longer, presentations, providing the learner with ample time to process and elaborate on each item during presentations. If one assumes that study-phase retrieval takes time, then perhaps having less presentations of a longer duration is preferred over having many presentations with shorter durations. In short, there might be a trade-off between the number of repetitions and the duration of repetitions (i.e. rate of presentation) and the question is what exactly is optimal for learning and retention. Obviously, what is optimal for learning one particular set of items (e.g., unrelated word pairs) might not necessarily be optimal for learning a different set of items (e.g., foreign vocabulary word pairs). Also, one might expect that the optimal study time distribution might differ depending on the difficulty of the to-be-learned materials. Thus, a second question was how our results concerning the optimal distribution of study time would generalize across different kinds of materials. Lastly, in practice, learners might often find themselves in the situation where they have control over the allocation of study time rather than following some predetermined distribution schedule. Thus, a third question was how effective learners are at distributing study time during single-session learning episodes when they have control over the allocation of study time.

In **Chapter 2**, we first briefly review the relevant literature on optimal study time distribution and we present two novel experiments investigating the effect of presentation rate on learning and retention. Previous research has mostly focused on the effect of presentation rate using stimulus materials like lists of nonsense syllables pairs (e.g., Bugelski, 1962; Stubin, Heimer, & Tatz, 1970), or lists of nonsense syllables paired with digits (e.g., Calfee & Anderson, 1971; Johnson, 1964). In our study, we used more meaningful materials (i.e., unrelated word pairs). Most importantly, prior studies have exclusively looked at the effect of presentation rate on recall performance after a single short-term retention interval. In our study, we assessed recall at multiple intervals allowing us to investigate retention and rate of forgetting.

In **Chapter 3**, we replicated and extended the study in Chapter 2 using materials that are more relevant for educational practice. That is, we investigated the effect of presentation rate on the learning and retention of foreign vocabulary word pairs. Most importantly, we also looked at the possibility that translation direction may moderate the effect of presentation rate. For language learners, translating words from a newly learned language into their native language is generally an easier task than translating the other way around (e.g., Schneider, Healy, & Bourne, 2002). It has been suggested that the optimal presentation rate might shift depending on the difficulty of the to-be-learned materials (de Jonge, Tabbers, Pecher, & Zeelenberg, 2012). Moreover, findings from previous research cast doubt on the extent to which the effects of presentation rate generalize across materials. For instance, in a study by Calfee and Anderson (1971), it was found that presentation rate had a substantial effect on the cued recall of pre-experimentally familiar target items (e.g., digits). However, when the to-be-learned target items were pre-experimentally unfamiliar to the learner (e.g., CVC nonsense syllables) presentation rate had little effect on final cued recall performance. Likewise, presentation rate might not be a factor of importance when language learners have to learn and recall unfamiliar (foreign vocabulary) target translations of familiar (native language) cue words.

In Chapters 2 and 3, we examined the effect of presentation rate in the situation where learners have no control over the pacing of study trials. However, in practice, learners are often given the opportunity to self-pace study trials instead of being presented with materials at a predetermined fixed presentation rate. Surprisingly few studies have investigated whether learners use time efficiently when given control over the pacing of study trials. However, the literature on metacognition suggests that learners might not be very proficient when it comes to allocating study time during self-paced study. It has been argued that effective self-guided learning requires one to go against certain intuitions and this in turn requires a reasonably good understanding of the processes that underlie durable learning (Bjork, 1999; Kornell & Bjork, 2007). Unfortunately, most people have many metacognitive misconceptions about remembering and learning (Kornell & Bjork, 2009). Given that people may not be very good at making the right decisions during learning, one could argue that learners might benefit from a situation where they have no control. In **Chapter 4**, we therefore investigated the efficacy of giving learners control over the rate of presentations during learning.

In the first part of the thesis, we focused on the question of how to optimally distribute study time. However, as already noted, another promising strategy for improving the retention of information is to provide learners with retrieval practice (test) trials in addition to study trials. In the second part of the thesis**,** we investigated the efficacy of providing learners with test trials in addition to study trials during single-session learning. It has been noted that time spent on test trials during learning is generally well spent, even though it takes up time that might otherwise have been utilized for additional study (Nungester & Duchastel, 1989). In our studies on the retrieval practice effect, learning conditions where some portion of the available amount of time was reserved for testing were compared to learning (control) conditions where the full amount of available time was spent studying. Of main interest was the potential benefit of the former strategy over the latter. We investigated how retrieval practice might enhance long-term retention of simple verbal material (e.g. vocabulary word pairs), and also more complex material (e.g., science discourse). In our studies, recall performance was assessed after both short and long retention intervals allowing us to assess the degree of forgetting over time. In addition, to further extend the approach taken in previous research, we also investigated the effect of practicing retrieval during initial learning on the delayed relearning of information.

Few studies have directly compared the respective retention benefits of repeated testing with and without the opportunity to restudy the materials.

However, one might expect that testing is especially beneficial if there are such opportunities for restudy (i.e. when study and test trials are alternated during learning). For instance, research suggests that attempting to retrieve information on test trials may also improve later encoding of that information within the same learning session, even when the retrieval attempt was unsuccessful (e.g., Arnold & McDermott, 2013; Izawa, 1966). However, the long-term benefit of this so-called *potentiating effect* of testing has not yet been thoroughly investigated. Moreover, earlier findings cast doubt on the idea that alternating study and test trials during initial learning can enhance long-term retention relative to repeated testing (e.g., Thompson, Wenger, & Bartling, 1978). In **Chapter 5,** we investigated the effect of testing with and without the opportunity for restudy on the long-term retention of foreign vocabulary word pairs. In two experiments, we assessed the rate of forgetting of word pairs learned under a restudy (control) condition, a repeated tests condition, and an alternated tests condition.

In **Chapter 6**, we extended the study in Chapter 5 by investigating the testing effect for word pairs of differential difficulty. Prior research suggests that the benefits of repeated testing might be beneficial primarily for those items that are relatively easy to learn and not so much for the more difficult items (e.g., Jang, Wixted, Pecher, Zeelenberg, & Huber, 2012). If this is the case, then the implications for educational practice will also be limited. In our study we used mixed lists of easy (related) word pairs and difficult (unrelated) word pairs. Recall performance was assessed for a repeated testing and a repeated study (control) condition after a 5-min and a 1-week interval. Also, in addition to recall performance, we looked at delayed (1-week) relearning. Research on the testing effect has mostly focused on single-session learning with long-term retention test performance as the crucial outcome variable. However, researchers have recently advocated other learning outcomes like relearning (Rawson & Dunlosky, 2011). Relearning as a measure of retention might also hold substantial practical relevance for educational purposes. In many situations, be that the controlled environment of a laboratory or an actual educational setting, expecting learners to achieve high levels of recall after relatively long retention intervals seems rather unrealistic. Accordingly, it has been argued that, in most circumstances, the least educators can hope for is rapid relearning of forgotten information (e.g., Nelson, 1971; Rawson & Dunlosky, 2011).

In **Chapter 7**, we further examined the effect of repeated testing on delayed relearning. Most previous research on the testing effect has focused on the situation where only a subset of information is encoded and repeatedly retrieved during initial learning. In our study we used a learning-to-criterion procedure before introducing the experimental manipulation (repeated study vs. repeated testing). All items were first learned to the criterion of one successful retrieval from long-term memory and all items subsequently received three post-retrieval study or test trials. Recall performance for both conditions was assessed after a 1-week retention interval. Most importantly, for both conditions we also looked at delayed relearning to criterion during the 1-week session relative to a new set of similar (not previously presented) items.

The testing effect is a very robust phenomenon that has been frequently reported in studies using simple verbal materials (e.g., Carpenter, Pashler, & Vul, 2006; Carpenter, Pashler, Wixted, & Vul, 2008; Karpicke & Roediger, 2007, 2008; Kuo & Hirshman, 1996; Pyc & Rawson, 2009, 2011; Toppino & Cohen, 2009; Wheeler et al., 2003). However, as we argue in **Chapter 8,** the positive effect of retrieval practice might be less robust for text materials. That is, studies using educationally relevant test formats (e.g., short answer questions) have come up with somewhat conflicting findings (e.g., Hinze & Wiley, 2011; Kang, McDermott, & Roediger, 2009; LaPorte & Voss, 1975; Nungester & Duchastel, 1982). One limitation of prior testing effect studies using short answer or equivalent test formats is that these studies have almost exclusively looked at recall performance at a single point in time, making it impossible to make any strong claims about the degree of retention of information. In our study, we assessed recall performance at multiple retention intervals, allowing us to investigate the rate of forgetting. Furthermore, in our study, we introduce one new possible explanation for the finding that the effects of testing might be less robust for text materials. One potential moderating factor that has not been considered in previous testing effect research is the connectedness of the to-be-learned materials. One distinctive feature of text and discourse is the highly structured and organized fashion by which information is presented. We tested the hypothesis that the coherence of the materials might moderate the effect by manipulating the coherence of the to-be-learned materials.

Lastly, in **Chapter 9** a summary and general discussion of the studies in the thesis is provided. The results and their implications for theory and practice are discussed and we provide recommendations for improving single-session learning and subsequent retention. Also, suggestions are made regarding directions for future research.

Chapter 2

The Effect of Study Time Distribution on Learning and Retention: A Goldilocks Principle for Presentation Rate*

* Chapter 2 was published as: de Jonge, M., Tabbers, H. K., Pecher, D., & Zeelenberg, R. (2012). The effect of study time distribution on learning and retention: A Goldilocks principle for presentation rate. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 405-412.

Abstract

Two experiments investigated the effect of presentation rate on both immediate (5 min) and delayed (2 days) cued recall of paired associates. Word pairs were presented for a total of 16 s per pair with presentation duration of individual presentations varying from 1 to 16 s. In Experiment 1 participants studied word pairs with presentation rates of 16 x 1 s, 8 x 2 s, 4 x 4 s, 2 x 8 s, or 1 x 16 s. A non-monotonic relationship was found between presentation rate and cued recall performance. Both short (e.g., 1 s) and long (e.g., 16 s) presentation durations resulted in poor immediate and delayed recall compared to intermediate presentation durations. In Experiment 2 we replicated these general findings. Moreover, we showed that the 4 s condition resulted in less proportional forgetting than the 1 s and the 16 s conditions.

One major factor that affects memory performance is the amount of time available for study. It is generally agreed upon that if one studies for a longer period of time then more is learned (Ebbinghaus, 1885/1964). Unfortunately, in reality students do not have an unlimited amount of study time at their disposal and even if they did, they would probably never spend it all studying. Since time is such a precious thing and often in short supply, it only makes sense that researchers everywhere spend heaps of it investigating the conditions under which learning is optimal. In the present study we asked ourselves the following question: If one only has a limited amount of time to study, how should the available time be divided to be optimally effective? Or, more specifically, when learning new information with a fixed amount of time available per item, what would be the most efficient rate of presentation?

Past research on the issue of optimal presentation rates in paired associate learning has led to different opinions on the matter. The most extreme position is probably held by researchers advocating that the amount learned is solely affected by the total study time available (e.g., Bugelski, 1962; Murdock, 1960). This idea, often referred to as the *total time hypothesis*, states that the amount of time necessary to learn a specific amount of information is fixed and does not vary as a function of the individual presentation durations into which the available time is divided (see Cooper & Pantle, 1967, for a review of the literature). There is no doubt that total study time plays an important role in determining the amount that can be learned. However, more recent studies have shown that total time is not the sole determinant of learning. For instance, there is a vast amount of research on the spacing effect, showing that spaced presentations of materials will generally result in superior recall compared to massed presentations (for a review see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). Clearly, these findings pose a serious problem for the total time hypothesis (Dempster, 1988; Melton, 1970). Even though the total time hypothesis might have fallen from grace as a theory for understanding human learning, some studies directly testing this hypothesis have led to interesting findings concerning the effect of presentation rate on the learning of verbal material.

In one such study (Johnson, 1964) participants learned a list of paired associates consisting of consonant-vowel-consonant (CVC) nonsense syllables paired with digits (e.g., *FAW-7*). Both total presentation time as well as presentation rate were manipulated between subjects. Items were studied for a total study time of 10, 20, 40, or 80 s and presented 1, 5, 10, or 20 times. Upon completion of the study phase participants received an immediate recall test. Not surprisingly, the results showed that the total study time had a significant effect on recall. More important, however, a non-monotonic relationship was found between presentation rate and recall performance when total presentation time was held constant. Although the relevant statistics were not always provided, the general pattern of results seems to indicate that both short and long presentation durations resulted in suboptimal learning, with optimal learning occurring at an intermediate presentation rate somewhere between 2 and 4 s per item.

The results from the Johnson (1964) study suggest that not only total study time but also the duration of individual presentations exerts an influence on memory performance. However, the results are not that unequivocal. Johnson (1964) used a fixed 4 s intertrial interval in his study. Because conditions consisting of more exposures automatically received more 4 s intertrial intervals, the differences in presentation rate between conditions also resulted in substantial differences in total time available for study (Cooper & Pantle, 1967). In a follow-up study by Stubin, Heimer, and Tatz (1970) an attempt was made to eliminate this confound of presentation rate and total study time. Paired associates (pairs of CVC nonsense syllables) were presented via a slide projector and it took 0.8 s for the projector to change slides. So, even though measures were taken to eliminate the problem with intertrial interval, there still was an effective 0.8 s lag between trials resulting in differences in total study time between conditions. Still, the results from the Stubin et al. (1970) study were largely in agreement with those obtained by Johnson (1964), even though they used an intertrial interval that was considerably shorter. A non-monotonic relationship was found between presentation rate and subsequent recall: both slow (≥ 10 s) as well as fast (2 s) presentation rates resulted in inferior recall performance compared to an intermediate 5 s presentation rate.

The results from Johnson (1964) and Stubin et al. (1970) suggest that, with total time held constant, the presentation duration of individual exposures to study materials influences the extent to which new information is learned. We believe these studies have important implications for both theoretical as well as educational purposes. Quite undeservedly, however, these studies have been largely neglected in the literature and there has been virtually no follow-up since the total time era ended.

In the two experiments reported here we further investigated the effect of presentation rate on paired associate learning. Our first objective was to replicate the Johnson (1964) and Stubin et al. (1970) studies, controlling for the methodological confound discussed earlier. We incorporated the intertrial interval within the presentation duration to make sure no differences in total time available for study would arise between study conditions. So, for instance, a 2 s presentation consisted of a 1.75 s presentation and a 0.25 s intertrial interval.

The second objective of our study was to extend the general findings from earlier studies. In the present experiments, presentation rate was manipulated within subjects (as opposed to between-subjects manipulations in previous studies). Also, we used more meaningful materials (words pairs instead of CVC nonsense syllables or digits). More important, we also wanted to look at longer retention intervals than those used in the earlier studies. In both the Johnson (1964) and the Stubin et al. (1970) study only a single short retention interval was used. In the Johnson (1964) study a test was given immediately after learning and in the Stubin et al. (1970) study participants received a final test only 20 seconds after the learning phase was completed. One could argue that, to some degree, short-term memory was being compared to long-term memory (Bugelski & McMahon, 1971). That is, the contribution of short-term memory to performance in the final test may have been different depending on the presentation rate. In conditions with a relatively slow presentation rate, items would on average be recalled a couple of minutes later on a final test, while in conditions with relatively fast presentation rates this would only be a matter of seconds. In the present study, participants first worked on a 5-min distractor task before taking a final recall test.

Another limitation of the use of a single short retention interval in earlier studies is that these studies do not inform us about the effect of presentation rate on forgetting. In the present study we therefore also included a retention interval of 2 days. Studies have shown that conditions that result in superior performance on an immediate recall test do not always benefit performance at a longer retention interval (e.g., Rawson & Kintsch, 2005; Roediger & Karpicke, 2006b). On a related note, it has been suggested that learning conditions that slow down initial learning can actually benefit long-term retention because these conditions introduce *desirable difficulties* during learning (Bjork, 1994, 1999). It remains to be seen whether presentation rates that are optimal for performance at short retention intervals are also optimal for performance at longer retention intervals.

Experiment 1

Method

Participants

Forty-two students from the Erasmus University Rotterdam participated in partial fulfillment of course requirements. Data from two participants were excluded from analyses, because these participants failed to show up for the 2-day final test.

Materials and Design

Eighty unrelated Dutch word pairs (e.g., *hamer – lift* [*hammer– elevator*], *spin – balkon* [*spider – balcony*]) were used in the experiment. All words were between four and six letters long and consisted of either one or two syllables. The mean word length was 4.87 (*SD* = 0.79). The mean word frequency per million (Keuleers, Brysbaert, & New, 2010) was 16.54 (*SD* = 44.62). Word pairs were divided over five lists of 16 items each. The computer application E-prime (Psychology Software Tools, Pittsburgh, PA) was used to create and run the experiment.

A 2 x 5 mixed-factorial design was used in the experiment with study condition as within-subjects factor, retention interval as between-subjects factor and cued recall score as dependent variable. Participants were randomly assigned to one of two retention interval conditions. One group of participants received a final cued recall test 5 minutes after the study phase ended and the other group received the cued recall test 2 days later. Both groups were required to return for the 2 day session regardless of the retention interval condition they were assigned to.

Participants studied word pairs under five different study conditions; 16 x 1 s, 8 x 2 s, 4 x 4 s, 2 x 8 s, and 1 x 16 s. In the 16 x 1 s condition a list of word pairs was presented 16 times with a presentation rate of 1 s per pair. In the 8 x 2 s condition a list of word pairs was presented 8 times with a presentation rate of 2 s per pair. The 4 x 4 s condition consisted of 4 list presentations of 4 s per pair, the 2 x 8 s condition of 2 list presentations with 8 s per pair, and in the 1 x 16 s condition participants a list of word pairs was presented once with a presentation rate of 16 s per pair. For each of these conditions, all pairs on the list were presented once in a random order before the pairs were presented again in a different random order (except, of course for the 1 x 16 s condition, in which all pairs were presented only once). By manipulating the presentation rate of word pairs in this manner we kept the total study time for each word pair constant across all conditions. Table 1 shows the average spacing in seconds between repetitions of the same pair as well as the average spacing in seconds between the first and last presentations of the same pair. The manipulation of presentation rate in the present experiment also resulted in differential spacing between conditions. However, as we will explain in the General Discussion, our results are not readily accounted for by these differences in spacing.

A total of 10 counterbalanced versions were created according to a scheme proposed by Lewis (1989) using a pair of Latin squares. Both the assignment of word pairs to conditions and the order in which conditions were administered during the study phase were counterbalanced. Across participants all word pairs appeared equally often in each study condition and all word pairs and study conditions were presented equally often in each of five blocks in the presentation order of conditions. Furthermore, immediate sequential effects were counterbalanced so that each condition was preceded as well as followed by each other condition equally often across participants. In the test phase, cue words were presented in a random order; items from the different study conditions were randomly intermixed.

Procedure

Participants were either tested alone or in small groups during two sessions. In the first session participants received verbal as well as onscreen instructions about the experiment. They were told that they would study word pairs at different presentation rates during five consecutive study blocks and that they would receive a memory test afterwards to assess their performance. Participants were also told that they were not allowed to cover part of the computer screen with their hand in order to test themselves during study. This was explicitly stated because during a pilot study we observed a number of participants using this strategy during study. To control for any undesirable effects that might occur as a result of self-testing we stressed that this was not allowed.

Before each study block participants received onscreen instructions telling them in which way the materials would be presented (how many times and at what rate). During each study block word pairs were presented on a computer screen in a different random order for each cycle. The two words of a pair were presented simultaneously, one above the other on the center of the screen. Upon completion of the study phase participants worked on Sudoku puzzles for a period of 5 min as a distractor task. After the 5-min distractor task half of participants received an immediate self-paced cued recall test. The remaining participants were dismissed after the distractor task and received a self-paced cued recall test 2 days later.

Results and Discussion

Figure 1 shows the mean proportion of correctly recalled words for both the 5-min group and the 2-day group as a function of study condition. At both delays, presentation rate and performance showed an inverted u-shape. Performance was optimal in the 4 x 4 s condition and dropped off with higher and lower presentation rates. On the 5-min test the mean percentages of correctly recalled items were 22%, 41%, 52%, 49%, and 37% for the 16 x 1 s, 8 x 2 s, 4 x 4 s, 2 x 8 s, and the 1 x 16 s condition, respectively. Two days later recall was considerably lower: 4%, 18%, 20%, 17%, and 9%, respectively (for the same five study conditions).

The data were analyzed using a 2 x 5 repeated measures ANOVA with retention interval as between subjects factor, study condition as within subjects factor and recall score as dependent variable. The analysis showed a significant effect of retention interval on final test score, $F(1, 38) = 16.47$, $p < .001$, $\eta_p^2 = .64$.

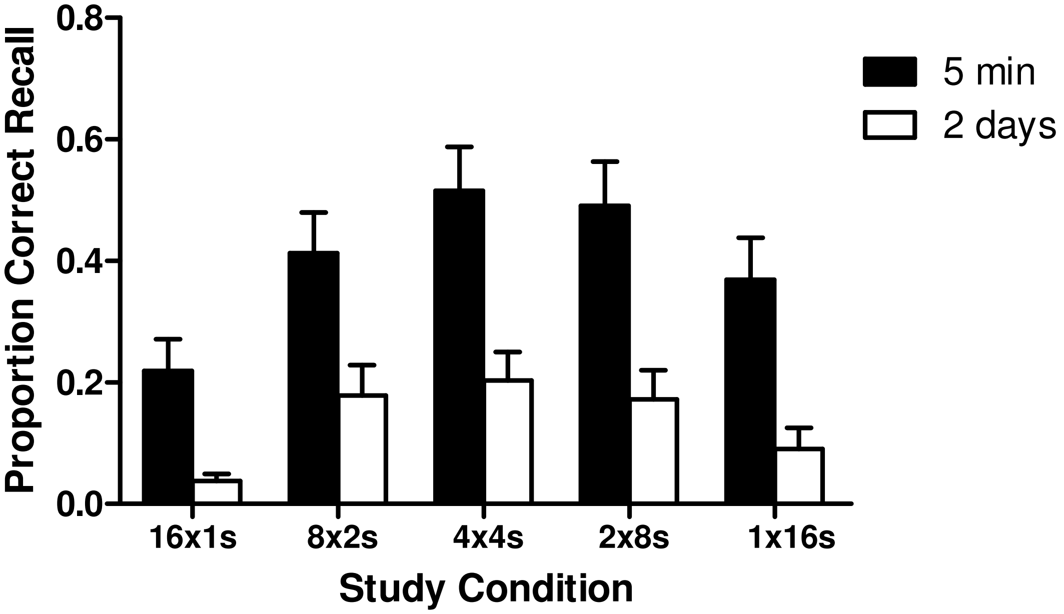

Figure 1. Proportion of words recalled on the 5-min and 2-day cued recall test as a function of study condition in Experiment 1. Error bars represent standard errors of the mean.

Recall scores were considerably lower on the 2-day test (14% correct) compared to recall on the 5-min test (40% correct). More important, study condition also had a significant effect on cued recall performance, $F(4, 152) = 14.22$, $p < .001$, $\eta_p^2 = .27$, indicating that the different rates of presentation during study resulted in differences in final test score. The interaction between retention interval and study condition was not significant, $F(4, 152) = 1.38$, $p > .20$.

We performed a subsequent repeated contrast analysis to determine whether performance for each presentation rate was significantly different from the next slower presentation rate. This analysis showed that studying word pairs 8 times with a presentation duration of 2 s per pair (the 8 x 2 s condition) resulted in superior recall compared to studying word pairs 16 times with 1 s per pair (the 16 x 1 s condition) , $F(1, 57) = 31.26$, $p < .001$, $\eta_p^2 = .38$. Studying word pairs 4 times with 4 s per pair (the 4 x 4 s condition) resulted in superior recall compared to the 8 x 2 s condition, $F(1, 57) = 6.23$, $p < .05$, $\eta_p^2 = .10$. The difference between the 4 x 4 s and the 2 x 8 s condition was not significant, $F < 1$. Finally, studying word pairs once for 16 s (the 1 x 16 s condition) resulted in inferior recall compared to the 2 x 8 s condition, $F(1, 57) = 30.76$, $p < .001$, $\eta_p^2 = .29$. The general pattern of results bears a striking resemblance to the findings of Johnson (1964) and Stubin et al. (1970). Indeed, there appears to be a non-monotonic relationship between presentation rate and recall of paired associates. Participants recalled few words for presentation durations of 1 s, but performance increased as presentation duration increased. However, this trend stalled for presentations of 4 to 8 s and reversed for presentations of 16 s.

Another point of interest in the present experiment was whether or not the general findings would extend over a longer retention interval. Or in other words, does the general pattern of results change over time? As can be seen in Figure 1 the general pattern of results persisted over the 2-day interval. The lack of an interaction between retention interval and study condition reported above supports this observation. If we would interpret the absolute difference in performance between the immediate and 2-day recall test for the different study conditions as an indication of forgetting, then we would have to conclude that the different presentation rates did not result in different rates of forgetting. However, research on forgetting suggests that the course of forgetting is best described by a power function (Wixted & Carpenter, 2007; Wixted & Ebbesen, 1991).[1] A power function of forgetting measures forgetting as a proportional decline of the amount of information that was originally stored in memory (Carpenter, Pashler, Wixted, & Vul, 2008). In this respect a proportional measure of forgetting would be a more appropriate way of looking at the rate of forgetting. Proportional forgetting can sometimes lead to different conclusions about forgetting compared to an absolute measure (e.g., Loftus, 1985). Figure 2 shows the proportional forgetting measures for all five study conditions. As is clear from Figure 2, study conditions that resulted in poor initial recall also resulted in high proportional forgetting, 83% in the 16 x 1 s and 75% in the 1 x 16 s condition. However, study conditions that resulted in superior recall on the 5-min test resulted in less proportional forgetting; 57% in the 8 x 2 s, 60% in the 4 x 4 s, and 65% in the 2 x 8 s condition.

Experiment 2

Experiment 2 was designed to extend the findings from Experiment 1 and to further investigate proportional forgetting. In Experiment 1 we assessed forgetting by comparing average performance on the immediate test with average performance on a delayed test across different groups of subjects. This made it impossible to perform standard statistical analysis on proportional forgetting in Experiment 1. In our second experiment both study condition and

[1] The question of which function provides the best description of the forgetting curve has been debated in the literature. Both power and exponential functions (as well as other functions) often provide excellent fits of forgetting data and it has proved hard to draw firm conclusions about the mathematical form of empirical forgetting functions. Nevertheless, based on different sets of data and different approaches recent studies have argued that power functions provide the best description of forgetting. For elaborate discussions of this issue we refer to Averell and Heathcote (2011), and Wixted (2004).

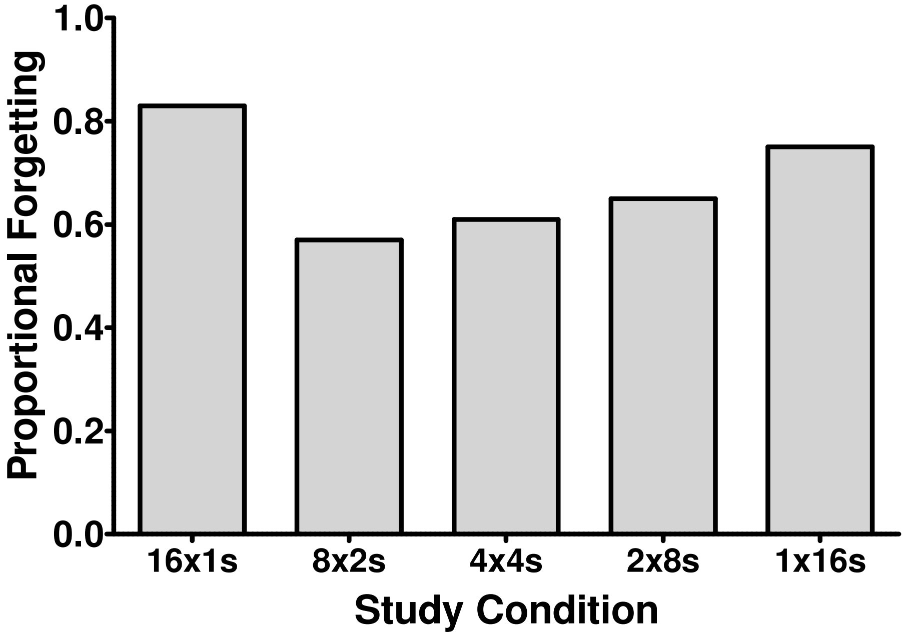

Figure 2. Proportional forgetting during the 2-day interval as a function of study condition in Experiment 1. forgetting in Experiment 1.

retention interval were manipulated within subjects. This enabled us to perform statistical analyses comparing proportional forgetting in the different study conditions. Because including all five study conditions present in Experiment 1 would result in a somewhat tedious experiment from the participants' perspective, we only compared the most extreme study conditions from Experiment 1 (the 16 x 1 s, the 4 x 4 s, and the 1 x 16 s condition).

Method

Participants

Thirty students from the Erasmus University Rotterdam participated in partial fulfillment of course requirements. Six participants were excluded from the analysis because of insufficient performance on the 5-min memory test (recall scores of zero on the 5-min test made assessment of subsequent proportional forgetting impossible). None of the participants had participated in our first experiment.

Materials and Design

A 2 x 3 factorial design was used with both retention interval and study condition as within-subject factors and recall score as dependent variable. Participants studied word pairs under three different study conditions (16 x 1 s, 4 x 4 s, and 1 x 16 s). Participants were tested on half of the word pairs on the 5-min test and the other half was tested after a 2-day interval.

Ninety-six word pairs were used in the experiment. Eighty word pairs were identical to the word pairs used in Experiment 1, and 16 new word pairs were compiled to supplement the original 80 word pair list. The mean word length was 4.86 (*SD* = 0.78). The mean word frequency per million (Keuleers et al., 2010) was 16.77 (*SD* = 43.14). Word pairs were divided over six lists of sixteen items each. Word pairs were assigned to 16-item word pair lists in such a fashion that every list would include an approximately equal number of new items. Six counterbalanced versions were created in the same general manner as in Experiment 1.

Procedure

The procedure was similar to the procedure in Experiment 1. Participants studied word pairs under three different study conditions during three consecutive study blocks. Upon completion of the 25-min study phase participants received a distractor task (Sudoku puzzles), followed by the 5-min cued recall test on half of the word pairs. All participants returned for the cued recall test on the remaining word pairs 2 days later.

Results and Discussion

Figure 3 shows the mean proportion of correctly recalled words on the 5-min and the 2-day recall tests as a function of study condition. On the 5-min test the mean percentages of correctly recalled items were 43%, 73%, and 60% for the 16 x 1 s, 4 x 4 s, and the 1 x 16 s condition, respectively. Two days later recall for the three conditions dropped to 6%, 28%, and 11%, respectively. Thus, after a 2-day delay, cued recall test performance in the 4 x 4 s condition was 348% and 169% higher compared to that in the 16 x 1 s and 1 x 16 s conditions, respectively.

A 2 x 3 repeated measures ANOVA revealed that the effect of retention interval on recall score was significant, $F(1, 23) = 168.37$, $p < .001$, $\eta_p^2 = .88$, indicating that forgetting occurred during the 2-day interval (59% correct on the 5-min test vs. 15% correct on the 2-day test). Also, as in Experiment 1 there was a significant effect of study condition on recall score, $F(2, 46) = 26.08$, $p < .001$, $\eta_p^2 = .53$, indicating that the different presentation rates resulted in different recall scores. The interaction between retention interval and study condition did not reach the conventional level of significance, $F(2, 46) = 2.74$, $p > .05$.

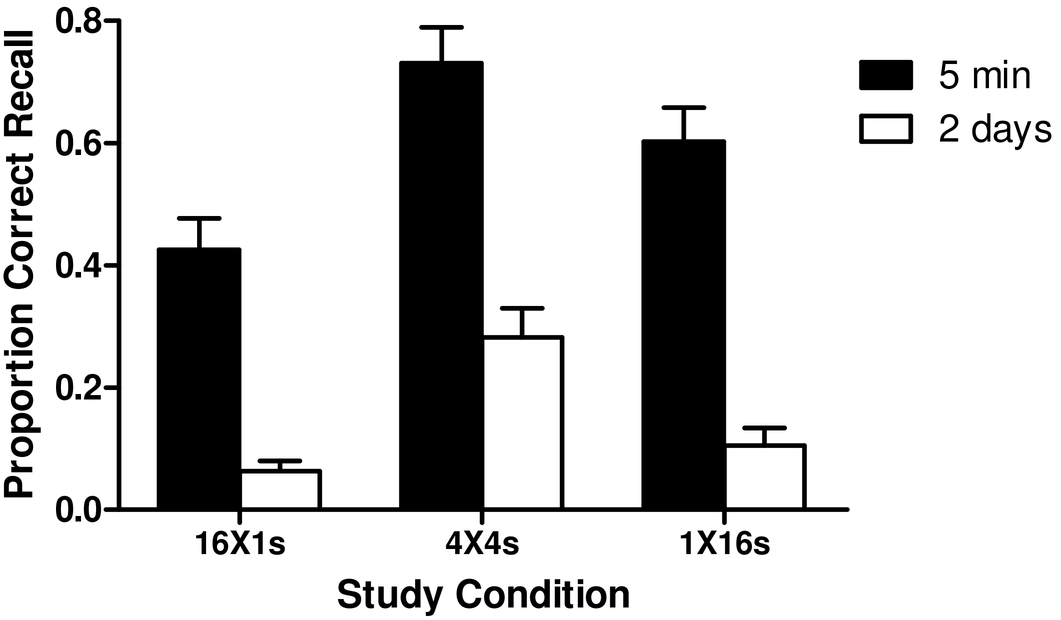

Figure 3. Proportion of words recalled on the 5-min and 2-day cued recall test as a function of study condition in Experiment 2. Error bars represent standard errors of the mean.

Follow-up analyses showed that the 4 x 4 s condition resulted in superior recall compared to both the 16 x 1 s condition, $F(1, 23) = 41.62$, $p < .001$, $\eta_p^2 = .64$, and the 1 x 16 s condition, $F(1, 23) = 29.23$, $p < .001$, $\eta_p^2 = .56$. To summarize, the general findings from Experiment 1 were replicated in Experiment 2, showing that presentation rate exerted a large influence on cued recall performance.

A more important question addressed by the present experiment was whether or not the different study conditions would result in different proportional forgetting. Figure 4 shows proportional forgetting as a function of study condition. As can be seen, the 16 x 1 s and the 1 x 16 s condition resulted in similar proportional forgetting (88% and 86% respectively) while studying word pairs in the 4 x 4 s conditions resulted in less proportional forgetting (64%). A repeated measures ANOVA revealed that the effect of presentation rate on proportional forgetting was significant, $F(2, 46) = 12.46$, $p < .001$, $\eta_p^2 = .35$. Follow-up analyses showed that the 4 x 4 s condition resulted in less proportional forgetting compared to both the 16 x 1 s and the 1 x 16 s condition, $F(1, 23) = 13.79$, $p < .005$, $\eta_p^2 = .38$ and $F(1, 23) = 19.32$, $p < .001$, $\eta_p^2 = .46$, respectively. So, the optimal presentation rate in the present experiment (the 4 x 4 s condition) did not only result in superior recall, but also in less proportional forgetting.

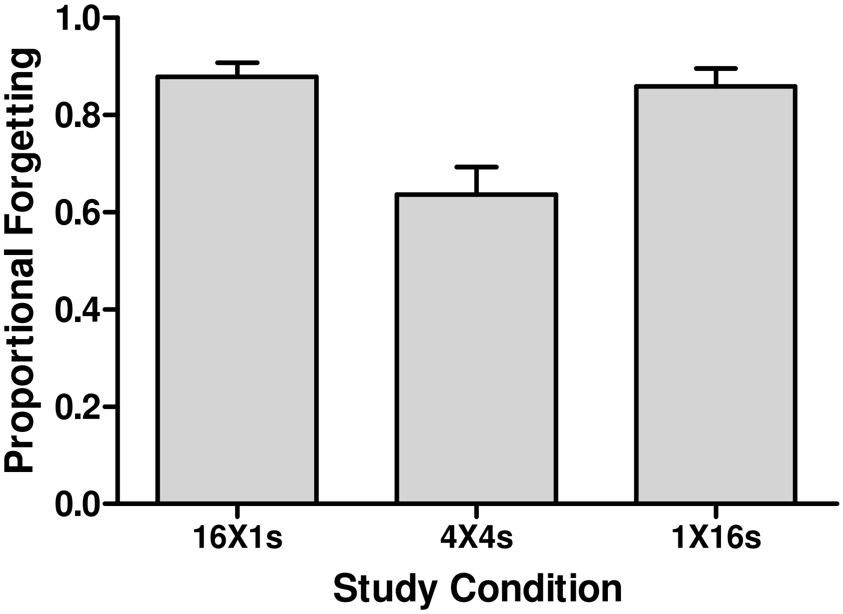

Figure 4. Proportional forgetting during the 2-day interval as a function of study condition in Experiment 2. Error bars represent standard errors of the mean.

General Discussion

In two experiments we investigated the effect of presentation rate on the learning and retention of paired associates. With total study time kept constant we found a non-monotonic relationship between the presentation rate of word pairs and subsequent recall. Performance was poor for short (e.g., 1 s) and long (e.g., 16 s) presentation durations and much better for intermediate (e.g., 4 s) presentation durations. These results indicate that the presentation duration of individual exposures has a large effect on memory performance even when the total study time is kept constant. Our findings extend earlier studies by Johnson (1964) and Stubin et al. (1970) by eliminating the methodological problems present in their studies and by using meaningful stimuli rather than CVC nonsense syllables. Furthermore, we showed that the effect of presentation rate is not only apparent on an immediate test, but also extends to a longer retention interval of 2 days. In Experiment 2 we replicated the general pattern of results and extended the findings by looking at proportional forgetting. We showed that presentation rates that resulted in poor immediate recall also resulted in more proportional forgetting.

In the present study, using unrelated word pairs, we found that a presentation rate of around 4 s resulted in optimal performance. Johnson (1964) and Stubin et al. (1970) obtained similar optimal presentation rates with different types of stimuli (CVC-digit pairs and CVC-CVC pairs, respectively). Nevertheless one should exercise caution in generalizing these optimal presentation rates to other materials. Although we would expect that the same general pattern will emerge across different kinds of materials, the optimal presentation rate might shift depending on the kind of materials used. For example, with more difficult materials a longer presentation rate might turn out to be optimal. Individual differences among learners may also affect the optimal rate of presentation. Also, as noted earlier, total study time is an important factor determining learning outcomes: when more time is available for learning, more can be learned. Both Johnson (1964) and Stubin et al. (1970) found that doubling the amount of time available for study resulted in substantial increases in cued recall performance. Thus we do not deny that total study time is an important determinant of learning. However, how the available time is divided up into study episodes is at least as important a factor.

Our results provide an intriguing puzzle for theoretical accounts of optimal study routines. There is a large body of literature on theoretical frameworks explaining a variety of distribution of practice phenomena like the spacing effect (see Cepeda et al., 2006; Delaney, Verkoeijen, & Spirgel, 2010, for recent reviews of the literature). Unfortunately, the relationship between presentation rate and the amount of spacing is not as straightforward as one might presume. Table 1 shows the average spacing in seconds between repetitions of the same pair as well as the average spacing in seconds between the first and last presentations of the same pair. As can be seen, the average interval between two presentations of a word pair increases as the presentation rate decreases. Following this measure of spacing, one would have to conclude that slower presentation rates resulted in more spacing between presentations. On the other hand, one could also consider the total time between the first and last presentation of a pair as an indication of spacing. Following this measure of spacing one would conclude that faster presentation rates resulted in more spacing of word pairs. Although both measures of spacing seem reasonable, the problem of course is that they lead to different conclusions about which study conditions were more spaced. Of course, this line of reasoning assumes that the evolvement of time is the critical dimension underlying spacing. If one assumes that the number of intervening presentations between repetitions as the critical dimension the picture is somewhat clearer. In this case the average spacing between repetitions is identical for all conditions (except for the 1 x 16 s condition) but the average total spacing (from the first to the last presentation) increases linearly with presentation rate.

Table 1
Average Spacing in Seconds for the Five Study Conditions in Experiment 1

Condition	Interrepetition Spacing	Total Spacing
16 × 1 s	15	239
8 × 2 s	30	222
4 × 4 s	60	188
2 × 8 s	120	120
1 × 16 s	-	-

Note. Interrepetition spacing refers to the average number of seconds between repetitions of the same pair. Total spacing refers to the average number of seconds between the first and last presentation of the same pair.

Although the manipulation of presentation rate in the present experiment inevitably resulted in differential spacing between conditions, we believe the present results are not that easily explained from a spacing point of view. That is, other factors, besides spacing *per se*, seem to play a role. In both our experiments we found similar patterns of results after a short and a long retention interval. So, presentation rates resulting in relatively good performance did so regardless of the delay between study and test. This is unlike research on the spacing effect which actually shows that different distributions of practice are optimal for different retention intervals (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Glenberg, 1976). Furthermore, at the 2-day retention interval the differences in spacing between the different presentation rate conditions of the present study were rather small relative to the length of the retention interval (approximately 170,000 s) and are therefore not expected to have a substantial impact on performance. Yet, large differences in performance were still found. The observed findings are also not simply an effect of massed versus spaced presentations. Numerous studies have shown that spaced presentations result in better performance than massed presentations (see Cepeda et al., 2006, for a review).[2] However, in both Experiment 1 and

[2] Some studies have reported better performance for massed practice than spaced practice on recall tests given almost immediately after learning (e.g., Balota, Ducheck, & Paullin, 1989; Peterson, Wampler, Kirkpatrick, & Saltzman, 1963). Allegedly, Endel Tulving has dubbed this rather paradoxical finding the “Peterson paradox” (Roediger, Balota, & Watson, 2001). In their meta-analysis, Cepeda et al. (2006) found that spaced presentations improved final-test performance by 9% when averaging over 96 studies that used a retention interval of less than 1 min. Improvements were also found for longer retention intervals. It appears that the beneficial effect of massed practice is limited to studies that used retention intervals of 4-8 s; retention intervals that are much shorter than those in the present study.

Experiment 2 of the present study performance was higher in the massed condition (1 x 16 s) than in the 16 x 1 s condition.[3] So, even though spacing of items and pacing of items can both be considered as accounts of distributed practice, we believe that there are some fundamental differences between the two.

We are inclined to propose an alternative explanation for the effect of presentation rate on subsequent recall, namely the *effective study time hypothesis*. It has been argued that some minimal amount of time might be necessary in order to optimally form an association (Stubin et al., 1970). This could explain why a fast presentation rate results in poor recall on a subsequent test. On the other hand, it has also been argued that presentation durations beyond some optimal value might cause inattention, decreased concentration, and boredom (Bugelski & McMahon, 1971). In this way a 16 s presentation might be inefficient, because less time is needed to form an associative link between two unrelated words. As a result, the remaining time beyond some optimum will be utilized in a less efficient way; that is less additional information will be stored in memory per unit time. This idea is reminiscent of the famous story of Goldilocks and the three bears. In the story Goldilocks successively tries three different bowls of porridge. She finds that one bowl is too cold, the other one is too hot, but the one in the middle is just right. The same principle appears to be true for presentation rates during the learning of paired associates. Presentation durations should be not too long, not too short, but just right.

In the present study we did not look at the kind of processing that took place during the different presentation rate conditions. So, we can only speculate about the strategies participants used during learning. However, it has been argued that elaborative study strategies take a certain amount of time to be effective (Bugelski, 1970). For instance, the results of a study by Bugelski, Kidd, and Segmen (1968) suggested that participants who studied paired associates under imagery instructions needed 4-8 s to form a useful image. At presentation rates of 4 and 8 s participants in the imagery group outperformed those in the control group. However, at a presentation rate of 2 s participants in the imagery group failed to outperform those in the control group, suggesting that they were unable to form an effective mental image. Perhaps a relatively fast presentation rate provides too little time for elaborative processing and learners will be forced to rely on less effective learning strategies (e.g., rote rehearsal).

[3] In both Experiment 1 and Experiment 2, more words were recalled in the 1 x 16 s condition than in the 16 x 1 s condition, $F(1, 38) = 9.85$, $p < .01$, $\eta_p^2 = .21$, and $F(1, 23) = 7.79$, $p < .01$, $\eta_p^2 = .25$, respectively.

Other related factors may be at play as well and provide a possible account of why intermediate presentation rates enhance initial learning and reduce forgetting. Note that although our effects are not simply the result of spacing, some mechanisms proposed in the spacing literature may provide a (partial) account of our results. One such mechanism is encoding variability. Encoding variability assumes that context fluctuates over time (Glenberg, 1976; Melton, 1967). Furthermore, encoding materials in different contexts enhances memory performance. More diverse contextual elements would be encoded for items presented four times (as in the 4 x 4 s condition) than for items presented only once (as in the 1 x 16 s condition). Without additional assumptions this account would predict optimal performance for the condition with the largest number of presentations, the 16 x 1 s condition. This prediction is clearly violated by our results. One could make the additional assumption that context storage takes time and little context information is stored during brief presentations of word pairs. Such a hypothesis, however, seems to conflict with findings that suggest context information is stored early on in processing (Malmberg & Shiffrin, 2005).

Another possible explanation is provided by the study phase retrieval account (see Raaijmakers, 2003, for a theory that combines context fluctuation and study phase retrieval to account for spacing effects). This account assumes that when an item is repeated it is retrieved from long-term memory and additional information is stored in the original trace (provided that retrieval of the item is successful). Spacing is beneficial because it results in more contextual information (as well as item and associative information) being stored in the memory trace. Like the encoding variability account, the study phase retrieval account could explain why performance in the 4 x 4 s condition is better than in the 1 x 16 s condition. More repetitions result in more successful study phase retrievals. However, without additional assumptions this account too would predict optimal performance for the condition with the largest number of presentations, the 16 x 1 s condition. It is plausible though, that successful study phase retrieval depends on the amount of time an item is presented; for brief presentation times of 1 or 2 s study phase retrieval may not be successful. However, to arrive at testable predictions, such an account would have to make specific assumptions about the time course of study phase retrieval. To summarize, spacing theories do not readily account for all aspects of our results. Factors such as encoding variability and study phase retrieval may play a role in our findings, but in order to account for the entire pattern of results spacing theories would need to make additional assumptions.

Recent years have seen a renewed interest in the factors that enhance learning and retention. Of particular importance, these studies have looked at the effects of study manipulations on performance not only on immediate recall but also after retention intervals ranging from several days to several months and even up to a year. Recent studies have shown that testing enhances long-term retention for a variety of materials. For example, Roediger and Karpicke (2006) found that recall of prose passages after a 1-week retention interval was substantially better for subjects who had been tested on those passages after initial study compared to subjects who received additional study opportunities. Similar benefits of testing over study have been found for the recall of Swahili-English word pairs (e.g., Karpicke & Roediger, 2008; Pyc & Rawson, 2010). In some cases, the advantages of testing over study amounted to improvements in performance of more than 150% (e.g., Karpicke & Roediger, 2008). Spacing also has substantial effects on memory performance. For example, in a very ambitious study Cepeda et al. (2008) investigated the effect of spacing (gap varied from 0 to 105 days) and retention interval (from 7 to 350 days) on cued recall and recognition of trivia facts. They found improvements in cued recall performance of up to 111% for the optimal gap between study trials as compared to a zero-day gap. Together, these studies and the present one indicate that testing, spacing, and appropriate presentation rates can have a large impact on memory. Not only immediate memory, but also delayed memory can benefit enormously from the right set of study conditions.

In sum, the present study indicates that there is a Goldilocks principle at work with regard to the presentation rate during the learning of paired associates within a fixed amount of time. We showed that presentation rates that are optimal for a short 5-min retention interval also benefit retention after a longer 2-day delay. We believe these results are not just interesting from a theoretical point of view, but they might also be of particular relevance for educational purposes. For instance, they could be used for optimizing foreign vocabulary learning. Most computer programs for learning foreign vocabulary provide their users with the opportunity for self-paced learning. In the present study we compared learning conditions with different presentation rates that remained constant during learning. However, when learning foreign vocabulary under self-paced instructions, learners tend to speed up the presentation rate as learning progresses. Even though they employ a reasonable presentation rate the first time through a list, they ultimately devolve to a presentation rate of less than 1 s per item (Kornell & Bjork, 2007). In the present study we showed that fast presentation rates of 1 s per pair resulted in suboptimal learning. Thus, it is doubtful whether or not the opportunity for self-paced study will result in efficient use of study time. Research on metacognition and learning generally shows that students are not very proficient when it comes to allocating self-paced study time (e.g., Nelson & Leonesio, 1988). Since self pacing often results in suboptimal study time allocation, it could be interesting to look at the usefulness of externally paced study schedules for improving learning and long-term retention.

Chapter 3

The Effect of Presentation Rate on Foreign Language Vocabulary Learning[*]

[*] Chapter 3 was submitted for publication as: Zeelenberg, R., de Jonge, M., Tabbers, H. K., & Pecher, D. (2014). The effect of presentation rate on foreign language vocabulary learning.

Abstract

The present study examined the effect of presentation rate on foreign language vocabulary learning. Experiment 1 varied presentation rates from 1 s to 16 s per pair while keeping the total study time per pair constant. Speakers of English studied Dutch-English translation pairs (e.g., *kikker* - *frog*) for 16 × 1 s, 8 × 2 s, 4 × 4 s, 2 × 8 s, or 1 × 16 s. The results showed a non-monotonic relationship between presentation rate and recall performance for both translation directions (Dutch → English and English → Dutch). Performance was best for intermediate presentation rates and dropped off for short (1 s) or long (16 s) presentation rates. Experiment 2 showed that the non-monotonic relationship between presentation rate and recall performance was still present after a 1-day retention interval for both translation directions. In Experiment 3, we replicated the findings of Experiment 1 using digit-trigram pairs as to-be-learned materials. Our results suggest that a presentation rate in the order of 4 s results in optimal learning of simple verbal materials like foreign language vocabulary.

Learning a new language can be a daunting task. One critical aspect of learning a new language consists of the acquisition of its vocabulary (e.g., Groot, 2000). Because time is valuable, learners would like to have a maximum return on their investment. The question of interest in this study is how to best use the limited time available for the acquisition of foreign language vocabulary. In particular we are interested in the question whether and how presentation rate affects foreign vocabulary learning. Given a limited total study time, is it best to study translation pairs (e.g., *oog* [Dutch]-*eye* [English]) just once for an extended period of time, or is it better to study translation pairs more often for a relatively short duration?

Several studies have found that presentation rate can have a substantial effect on paired-associate learning (e.g., Cull, d'Anna, Hill, Zechmeister, & Hall, 1991; Johnson, 1964; Stubin, Heimer, & Tatz, 1970). None of these studies, however, investigated the effect of presentation rate with foreign vocabulary learning. Moreover, these studies suffer from a number of methodological problems such as an improper control of the total time available for study, or the use of a limited range of presentation rates. In addition, some studies (e.g., Johnson, 1964; Stubin et al., 1970) have used very short study lists and retention intervals of 20 s or less so that performance was probably, at least in part, based on retrieval from short-term memory.

In a recent study, de Jonge, Tabbers, Pecher, and Zeelenberg (2012) re-examined the effect of presentation rate on paired-associate learning for unrelated word pairs while eliminating the methodological problems present in earlier studies. De Jonge et al. kept the total time available for study constant across a range of different presentation rate conditions (i.e., 16 × 1 s, 8 × 2 s, 4 × 4 s, 2 × 8 s or 1 × 16 s). Their results showed a non-monotonic relationship between presentation rate and cued recall performance. Both fast (i.e., 16 × 1 s) and slow (i.e., 1 × 16 s) presentation rates resulted in poor cued recall performance, compared to conditions with intermediate presentation rates. Moreover, extending the results of prior studies that used only short retention intervals, the non-monotonic relationship between presentation rate and cued recall was still present after a 2-day retention interval. The optimal learning with intermediate presentation rates was dubbed the *Goldilocks principle of presentation rate* (de Jonge et al., 2012).

De Jonge et al. (2012) argued that their findings might be particularly relevant for educational purposes and could be used to optimize foreign language vocabulary learning. They suggested that a non-monotonic relation between presentation rate and recall performance would obtain for different types of stimulus materials, but speculated that the optimal presentation rate might shift depending on the difficulty of to-be-learned materials. More specifically, with more difficult materials, such as those presented in the acquisition of foreign language vocabulary, longer presentation rates might turn out to be optimal. However, it is still an open question whether and how presentation rate affects foreign vocabulary learning.

Foreign vocabulary learning can be viewed as a paired-associate task involving three processes: learning the cue word (stimulus), learning the target word (response) and learning the association between cue and target (e.g., McGuire, 1961; Schneider, Healy, & Bourne, 2002). When learning new vocabulary in a foreign language, translation performance is usually better when participants are given the foreign language word as a cue (e.g. *oog*), and have to provide the corresponding familiar language equivalent (e.g. *eye*) than vice versa. Memory traces for newly learned words may be relatively weak, incomplete or error prone. As a consequence, producing the foreign language translation equivalent to the familiar language cue word may often be unsuccessful, resulting in relatively poor recall. Recall of the familiar language word may be relatively successful because even a weak memory trace for the foreign word may be sufficient to differentiate the foreign word from other foreign words and thus allow for retrieval of the familiar language translation equivalent.

Some findings suggest that the effect of presentation rate depends on the direction of recall. In Calfee and Anderson's (1971) study, participants studied trigram-digit (e.g., *LUB - 91*) or digit-trigram pairs (e.g., *91 - LUB*) under different presentation rate conditions (1 s, 2 s, 3 s, 4 s, 10 s, or 20 s), while keeping total study time constant across conditions. Thus, analogous to the situation of foreign vocabulary learning, participants had to learn the association between familiar items (i.e., digits) and unfamiliar items (i.e., nonsense syllables). Calfee and Anderson found a non-monotonic relationship between presentation rate and cued recall performance, when participants were given the trigram as a cue during test (e.g., *LUB - ?*) and had to recall the digit (e.g., *91*). Performance was best when presentation rates were in the 2-4 s range, but dropped off with shorter and longer presentation rates. Importantly, however, presentation rate had little systematic effect when participants were given the digit as a cue (e.g., *91 - ?*) and had to recall the trigram (e.g., *LUB*). These results suggest that presentation rate can have little effect on the cued recall of pre-experimentally unfamiliar target items. If the results of Calfee and Anderson generalize to foreign vocabulary learning, one would expect presentation rate to have a differential effect on the recall of foreign language translation equivalents depending on translation direction.

In the present study, speakers of English learned new words in a foreign language: Dutch. Translation equivalents (e.g., *oog - eye*) were studied for 16 s per pair. The presentation rate of word pairs was manipulated in a similar fashion as in the de Jonge et al. (2012) study. That is, with total study time equated across conditions, pairs were presented for 16 × 1 s, 8 × 2 s, 4 × 4 s, 2 × 8 s or 1 × 16 s. In addition, we manipulated translation direction. Participants were presented with either the Dutch word (e.g., *oog*) during test and had to recall the English translation equivalent (e.g., *eye*) or were presented with the English word (e.g., *eye*) during test and had to recall the Dutch translation equivalent (e.g., *oog*). In line with previous findings (e.g., Schneider et al., 2002), we expected better performance in the Dutch→English than in the English→Dutch condition. In addition, given the previous findings of de Jonge et al. (2012) and of Calfee and Anderson (1971), we expected to find a substantial effect of presentation rate on recall performance for the Dutch→English condition. One question of interest was whether the optimal presentation rate would be in the order of 4 s or whether a longer presentation rate would be optimal. An additional question was whether similar findings would be obtained for both translation directions or whether presentation rate would have little effect on recall performance when participants have to recall unfamiliar foreign language translation equivalents (English→Dutch), as is suggested by the results of Calfee and Anderson.

Experiment 1

Method

Participants

One hundred students from the University of California, San Diego participated in partial fulfillment of course requirements. Fifty participants participated in the Dutch→English condition; the remaining 50 participated in the English→Dutch condition.

Materials and Design

Sixty Dutch-English translation pairs (e.g., *oog - eye*, *–fles - bottle*, *ridder - knight*) were used in the experiment. Translation pairs were non-cognates: each Dutch word and its English translation equivalent were orthographically and phonologically dissimilar. All words (both Dutch and English) were between 3 and 7 letters long and consisted of either one or two syllables. The mean word length of the Dutch words was 4.80 (*SD* = 1.24); the mean word length of the English words was 4.80 (*SD* = 1.06). The mean word frequency per million of the English words (Brysbaert & New, 2009) was 60 (*SD* = 104). Translation pairs were divided over five lists of 12 items each. The computer application E-prime (Psychology Software Tools, Pittsburgh, PA) was used to create and run the experiment.

A 5 × 2 mixed-factorial design was used with presentation rate as within-subjects factor, and translation direction as a between-subjects factor.

Participants studied translation pairs under five different presentation rate conditions: 16 × 1 s, 8 × 2 s, 4 × 4 s, 2 × 8 s, and 1 × 16 s. In the 16 × 1 s condition a list of translation pairs was presented 16 times with a presentation rate of 1 s per pair. The 8 × 2 s condition consisted of 8 list presentations of 2 s per pair, the 4 × 4 s condition consisted of 4 list presentations of 4 s per pair, the 2 × 8 s condition of 2 list presentations of 8 s per pair, and the 1 × 16 s condition of one list presentation of 16 s per pair. Thus, the presentation rate was varied, but the total time for each translation pair was kept constant.

Ten counterbalanced versions were created according to a scheme proposed by Lewis (1989) using a pair of Latin squares. Both the assignment of translation pair lists to the five presentation rate conditions and the order in which conditions were administered during the study phase were counterbalanced. Across participants, all lists appeared equally often in each condition, and all lists and conditions appeared equally often in each of five study blocks. Furthermore, immediate sequential effects were counterbalanced so that, across participants, each condition was preceded as well as followed equally often by each other condition. In the test phase, the items from the different presentation rate conditions were intermixed and presented in a random order.

Translation direction (Dutch→English vs. English→Dutch) was manipulated between subjects. Within each translation direction group, the translation direction remained the same throughout the experiment. Thus, in each group, participants were tested in the direction congruent to the direction used during study. Half of the participants had to translate Dutch words into English (e.g., *oog* - *?*); the other half of the participants had to translate English words into Dutch (e.g., *eye* - *?*). Participants were randomly assigned to either the Dutch→English or the English→Dutch translation direction.

Procedure

Participants received on-screen instructions about the experiment. They were told that they would study translation pairs at different presentation rates during five consecutive study blocks, and that, afterwards, they would receive a memory test to assess their performance. To control for any unwanted effects that might occur as a result of self-testing (e.g., by covering part of the computer screen with their hand during study), it was explicitly stated that this was not allowed.

Before each study block, participants received on-screen instructions telling them in which way the materials would be presented (how many times and at what rate). Each time a list was repeated, the translation pairs were presented in a different random order (except, of course for the 1 × 16 s condition, in which all pairs were presented only once). Different random orders were generated for each participant. The two words of a pair were presented simultaneously, one above the other on the center of the screen (e.g., the English word *eye* above its Dutch translation equivalent *oog*, for the English→Dutch condition). As in the de Jonge et al. (2012) study, we incorporated the inter-trial interval of 0.25 s within the presentation duration. So, for instance, a 2 s presentation consisted of a 1.75 s presentation and a 0.25 s inter-trial interval. By manipulating the presentation rate of translation pairs in this manner, we kept the total time available for study constant across all conditions.

Upon completion of the study phase, participants worked on multiplication problems (e.g., *7 × 43 = ?*, *57 × 8 = ?)* for a period of 5 minutes. Upon completion of the distractor task, participants received a self-paced cued recall test. The cue words of the translation pairs were presented on the screen one at the time, and participants were required to type the target translations on the keyboard. The letters typed on the keyboard were displayed on the screen directly below the to-be-translated cue word. Participants could use the <Backspace> key to correct errors. Participants were instructed to type carefully in order to minimize the number of typos. When participants felt unable to provide the translation they were instructed to type "I don't know". The next cue was presented after participants pressed the <Enter> key.

Results and Discussion

Figure 1 shows the mean proportion of correctly recalled target words as a function of presentation rate and translation direction. As can be seen, performance was optimal for intermediate presentation rates and dropped off with higher and lower presentation rates. Moreover, performance was better when participants had to provide the translation in their native language, English, than when they had to provide the translation in the newly learned language, Dutch. The data were analyzed using a 5 × 2 repeated measures ANOVA with presentation rate as within-subjects factor, translation direction as between-subject factor and proportion correct recall as the dependent variable. Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of presentation rate. Degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity. There was a significant main effect of presentation rate on cued recall performance, $F(5.51, 344{,}01) = 13.11$, $p < .001$, $\eta_p^2 = .12$. The main effect of translation direction was also significant, $F(1, 98) = 13.70$, $p < .001$, $\eta_p^2 = .12$. The interaction between presentation rate and translation direction was not significant, $F < 1$. Follow-up contrasts showed that studying translation pairs 16 times for 1 s per pair (the 16 × 1 s condition) resulted in inferior cued recall performance compared to studying translation pairs 8 times for 2 s per pair (the 8 × 2 s condition), $F(1, 98) = 34.67$, $p < .001$, $\eta_p^2 = .26$. Cued recall performance in the 8 × 2 s condition did not differ significantly

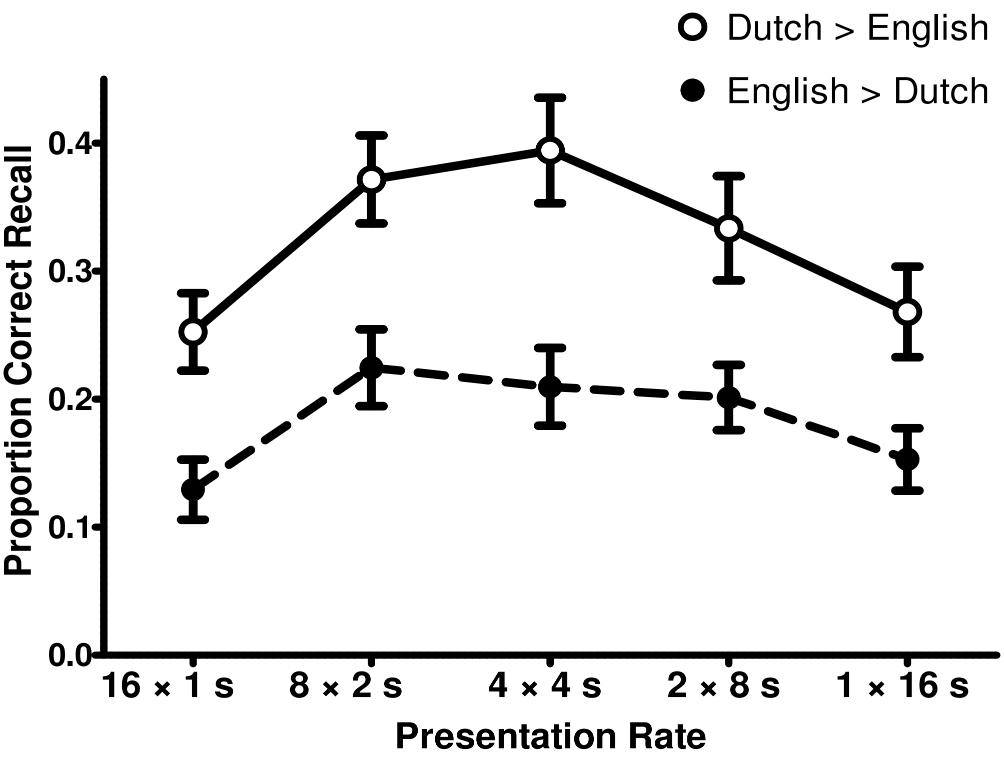

Figure 1. Proportions of correctly recalled translation equivalents as a function of presentation rate and translation direction in Experiment 1. Error bars represent standard errors of the mean.

from that in the 4 × 4 s condition, $F < 1$. Cued recall performance in the 4 × 4 s condition did not differ significantly from that in the 2 × 8 s condition, $F(1, 98) = 2.67$, $p = .11$. Finally, cued recall performance in the 2 × 8 s condition was superior to performance in the 1 × 16 s condition, $F(1, 98) = 8.47$, $p = .005$, $\eta_p^2 = .08$.

In sum, the results of Experiment 1 extend those of previous work. As with the unrelated word pairs used in the de Jonge et al. (2012) study, we found a non-monotonic relationship between presentation rate and cued recall performance in a foreign vocabulary learning task. Interestingly, however, our results indicate that the pattern of results does not seem to depend on translation direction, a result that was unexpected (cf. Calfee & Anderson, 1971).

Experiment 2

The first objective of Experiment 2 was to replicate the findings from Experiment 1. Furthermore, we also wanted to investigate whether or not the findings would extend to a longer retention interval. Thus, in addition to a 5-min retention interval we included a retention interval of 1 day. Previous studies indicate that conditions resulting in superior performance on an immediate recall test can sometimes result in inferior performance at longer delays (e.g., Rawson & Kintsch, 2005; Roediger & Karpicke, 2006; Schmidt & Bjork, 1992). However, for unrelated word pairs, de Jonge et al. (2012) found that the effect of presentation rate extended over a longer retention interval. The question is whether the same is true for both translation directions in a foreign vocabulary learning task. Because including all five presentation rates present in Experiment 1 would result in a somewhat tedious experiment from the participants' perspective, we only compared the most extreme and the intermediate presentation rates from Experiment 1 (the 16 × 1 s, the 4 × 4 s, and the 1 × 16 s conditions).

Method

Participants

Seventy-two students from the University of California, San Diego participated in partial fulfillment of course requirements. None of the participants had participated in Experiment 1.

Materials and Design

A 3 × 2 × 2 factorial design was used with presentation rate and retention interval as within-subject factors, and translation direction as a between-subjects factor. Participants studied translation pairs under three different presentation rates (16 × 1 s, 4 × 4 s, and 1 × 16 s), and were tested on half of the pairs on the 5-min test and on the other half after a 1-day retention interval. As in Experiment 1, half of the participants were assigned to the Dutch→English translation direction and the other half was assigned to the English→Dutch translation direction.

Twelve additional noncognate Dutch - English translation pairs were added to the set of 60 pairs from Experiment 1. All words (both Dutch and English) were between 3 and 7 letters long and consisted of either one or two syllables. The mean word length of the Dutch words was 4.76 (*SD* = 1.16); the mean word length of the English words was 4.89 (*SD* = 1.10). The mean word frequency per million of the English words (Brysbaert & New, 2009) was 54 (*SD* = 96). The 72 translation pairs were divided over 6 lists of 12 items each. Six counterbalanced versions of the study procedure were created in the same general manner as in Experiment 1. In addition, we counterbalanced the assignment of stimulus sets to retention interval condition, resulting in twelve counterbalanced versions.

Procedure

The procedure was similar to the procedure of Experiment 1. Participants studied translation pairs under three different presentation rates during three consecutive study blocks. Upon completion of the study phase, participants received a 5-min distractor task (solving multiplication problems) followed by a cued recall test on half of the translation pairs from each presentation rate condition. All participants returned for the translation test on the remaining translation pairs 1 day later.

Results and Discussion

Figure 2 shows the mean proportion of correctly recalled words on the 5-min and the 1-day tests as a function of presentation rate and translation direction. As can be seen, participants translated more words correctly in the 4 × 4 s condition than in the 16 × 1 s and 1 × 16 s conditions, both when tested after a 5-min and a 1-day retention interval. Also, participants translated more words correctly when they had to provide a translation in their native language, English, than when they had to provide a translation in the newly learned language, Dutch. Lastly, and not surprisingly, performance dropped when participants were tested after a 1 day retention interval compared to a 5 min retention interval.

The data were analyzed using a 3 × 2 × 2 mixed ANOVA with presentation rate and retention interval as within-subjects factors, translation direction as a between-subjects factor, and proportion correct recall as dependent variable. Mauchly's test indicated that the assumption of sphericity had been violated for some of the data. In these cases, degrees of freedom were corrected using Greenhouse-Geiser estimates of sphericity. There was a significant main effect of presentation rate, $F(1.63, 113.93) = 27.57$, $p < .001$, $\eta_p^2 = .28$, translation direction, $F(1, 70) = 19.69$, $p < .001$, $\eta_p^2 = .22$, and retention interval, $F(2, 140) = 114.73$, $p < .001$, $\eta_p^2 = .62$. In addition, the interaction between retention interval and translation direction and the interaction between retention interval and presentation rate were both significant, $F(1, 70) = 6.86$, $p < .05$, $\eta_p^2 = .09$, and $F(2, 140) = 6.49$, $p < .01$, $\eta_p^2 = .09$, respectively. Inspection of Figure 2 suggests that these interaction effects were probably the result of recall performance being close to floor after the 1-day interval in the 16 × 1 s and 1 × 16 s condition for the English→Dutch direction group, but not for the Dutch→English group. The interaction between presentation rate and translation direction as well as the three-way interaction failed to reach significance (both *Fs* < 1). Separate follow-up analyses were performed for the two translation direction groups. For the Dutch→English group, a repeated measures ANOVA revealed that there was a significant main effect of presentation rate, $F(1.68, 58.93) = 12.44$, $p < .001$, $\eta_p^2 = .26$, and a significant main effect of retention interval, $F(2, 70) = 65.15$, $p < .001$, $\eta_p^2 = .65$. The interaction between presentation rate and retention interval was not significant, $F < 1$. Follow-up contrasts showed that performance in the 4 × 4 s condition was significantly better than performance in the 16 × 1 s condition,

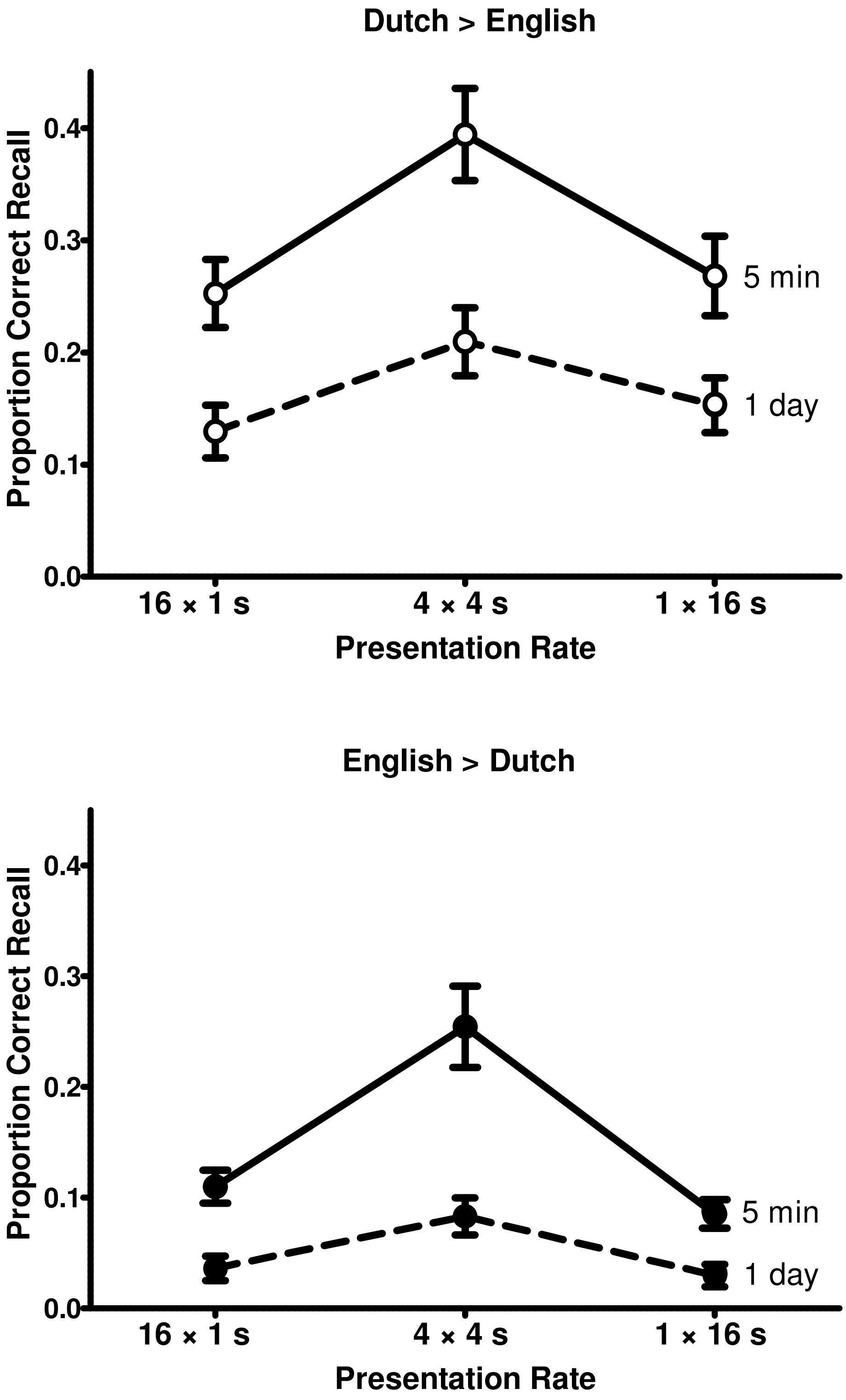

Figure 2. Proportions of correctly recalled translation equivalents on the 5-min and 1-day translation test as a function of presentation rate and translation direction in Experiment 2. Error bars represent standard errors of the mean.

$F(1, 35) = 11.72$, $p < .01$, $\eta_p^2 = .25$, and that in the 1 × 16 s condition, $F(1, 35) = 18.92$, $p < .001$, $\eta_p^2 = .35$.

For the English→Dutch group, there was a significant main effect of presentation rate, $F(1.44, 50.48) = 19.65$, $p < .001$, $\eta_p^2 = .36$, and a significant main effect of retention interval, $F(2, 70) = 51.46$, $p < .001$, $\eta_p^2 = .60$. Also, the interaction between presentation rate and retention interval was significant, $F(1.59, 55.73) = 8.68$, $p < .001$, $\eta_p^2 = .20$. As already noted, this interaction was likely due to performance being close to floor in the 16 × 1 s and 1 × 16 s conditions (even after a retention interval of only 5 min, performance in these conditions was already low at approximately 10%, hence the absolute amount of forgetting could never have attained the amount of the 17% observed in the 4 × 4 s condition). Thus, this interaction is not theoretically meaningful. Importantly, however, at both the 5-min and the 1-day retention intervals there was a significant effect of presentation rate, $F(1.40, 48.88) = 18.60$, $p < .001$, $\eta_p^2 = .35$, and $F(1.65, 57.86) = 5.90$, $p < .001$, $\eta_p^2 = .19$, respectively. Follow-up contrasts showed that, at the 5 min retention interval, the 4 × 4 s condition outperformed both the 16 × 1 s condition, $F(1, 35) = 14.99$, $p < .001$, $\eta_p^2 = .30$, and the 1 × 16 s condition, $F(1, 35) = 30.76$, $p < .001$, $\eta_p^2 = .77$. At the 1-day retention interval too, the 4 × 4 s condition outperformed both the 16 × 1 s condition, $F(1, 35) = 7.19$, $p < .05$, $\eta_p^2 = .17$, and the 1 × 16 s condition, $F(1, 35) = 7.28$, $p < .05$, $\eta_p^2 = .17$.

To summarize, these results replicate and extend those obtained in Experiment1. We replicated the non-monotonic relationship between presentation rate and translation recall performance for both language directions and showed that this relation is still present after a 1-day retention interval.

As noted, results from previous research on paired associate learning indicated that presentation rate can sometimes have little systematic effect on subsequent cued recall of pre-experimentally unfamiliar target items. That is, in the Calfee and Anderson (1971) study, when participants studied digit-trigram pairs (*91-LUB*), subsequent cued recall performance of target trigrams did not appear to be systematically affected by the rate of presentation. Thus, one might wonder why Calfee and Anderson did not find a non-monotonic relationship between presentation rate and cued recall performance for digit-trigram recall. Experiments 1 and 2 of the present study differed from the Calfee and Anderson study on several factors. For example, Calfee and Anderson used a between-participants manipulation of presentation rate, participants in their study learned a small number of pairs (16 pairs vs. 60 and 72 pairs in Experiments 1 and 2 of the present study), alternating study and test cycles were used during the study phase of their experiment (no test trials were used in the present study until the final recall test) and they used a longer total study time (60 s vs. 16 s per pair in the present study). None of these factors, however, seem plausible candidates for explaining why no systematic effect of presentation rate on performance was observed for digit-trigram recall (*91 - ?*). Note that Calfee and Anderson did find a non-monotonic relation between presentation rate and performance for trigram-digit recall (*LUB - ?*). Moreover, other studies using relatively short study lists, a longer total study time than the present study and a between-participants manipulation of presentation rate have also found a non-monotonic relationship between presentation rate and subsequent recall performance (e.g., Johnson, 1964; Stubin, et al.,1970). One possibility is that there is something special about digit- trigram recall (e.g., *91 - ?*) causing presentation rate to have little effect on performance. Before speculating on a possible explanation we wanted to make sure the results of Calfee and Anderson would replicate. The study of de Jonge et al. (2012) and the present experiments suggest that the typical finding is that of a non-monotonic relation between presentation rate and cued recall performance. The results of Calfee and Anderson for digit-trigram recall have, to our knowledge, not been replicated. In Experiment 3 we therefore re-examined the effect of presentation rate on digit-trigram recall.

Experiment 3

Method

Participants

Thirty students from the Erasmus University Rotterdam participated in partial fulfillment of course requirements.

Materials, Design and Procedure

The experiment closely resembled Experiment 1. The major difference was in the materials being presented for study. Sixty digit-trigram pairs (e.g., *37- KOG, 83 - FEH*) similar to the ones used in the Calfee and Anderson (1970) study were created for the experiment. Digits were selected from the range 10-99, with the digits 1 to 9 appearing about equally often in final and initial positions. The nonsense trigrams were selected from Archer's (1960) norms for CVC syllables. All trigrams were pronounceable CVC syllables and none of the trigrams spelled a Dutch word. In each block, participants studied 12 digit-trigram pairs in one of the five presentation rate conditions (i.e., 16 × 1 s, 8 × 2 s, 4 × 4 s, 2 × 8 s, and 1 ×16 s). The order of conditions and assignment of stimuli to conditions was counterbalanced in the same manner as Experiment 1. We anticipated that learning digit-trigram pairs would be harder than learning English-Dutch translation pairs. To prevent floor effects, memory was tested after each study block of 12 pairs, rather than after study of all the pairs. Upon completion of each study block, participants received a 2-min distractor task (color decision):

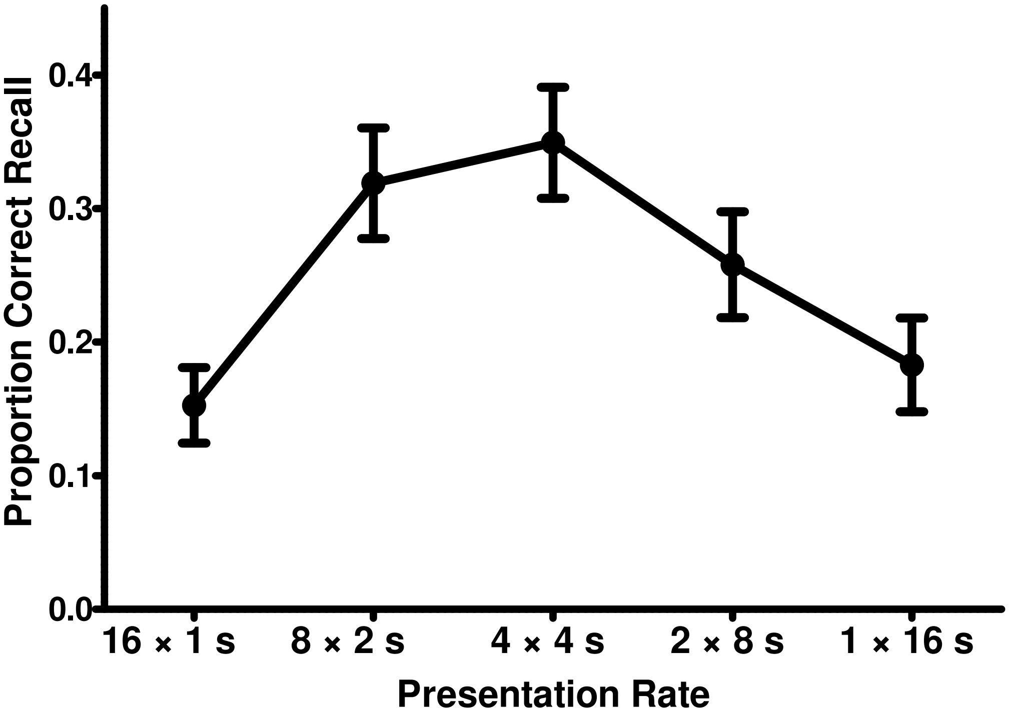

Figure 3. Proportions of correctly recalled digit-cued trigrams as a function of presentation rate in Experiment 3. Error bars represent standard errors of the mean.

two color patches were presented simultaneously and subjects decided whether or not the color of the two patches was identical. The distractor task was followed by the memory test: digit-cued trigram recall (*37 - ?*). Note that the prevalent finding reported in the literature is one of a non-monotonic relation between presentation rate and cued recall performance, a finding that Calfee and Anderson obtained for trigram-digit recall, but not for digit-trigram recall. Because the question of interest was whether we would obtain a non-monotonic relation between presentation time and digit-trigram recall we did not assess trigram-digit recall.

Results and Discussion

Figure 3 shows the mean proportion of correctly recalled target trigrams as a function of presentation rate. As can be seen, the pattern of results was similar to the patterns observed in Experiments 1 and 2. Performance was optimal for intermediate presentation rates and dropped off with higher and lower presentation rates. A repeated measures ANOVA showed that the effect of presentation rate was significant, $F(4, 116) = 7.80$, $p < .001$, $\eta_p^2 = .21$. Follow-up contrasts revealed that the 16 × 1 s condition resulted in inferior cued recall performance compared to the 8 × 2 s condition, $F(1, 29) = 20.90$, $p < .001$, η_p^2

= .42. The difference between the 8 × 2 s, and the 4 × 4 s condition was not significant, $F < 1$. Performance in the 4 × 4 s condition was superior to performance in the 2 × 8 s condition, $F(1, 29) = 4.91$, $p < .05$, $\eta_p^2 = .14$. Lastly, performance in the 2 × 8 s condition was superior to performance in the 1 × 16 s condition, $F(1, 29) = 5.93$, $p = .05$, $\eta_p^2 = .17$.

Thus, contrary to Calfee and Anderson (1971), we obtained a clear non-monotonic relation between presentation rate and digit-trigram recall performance. Of course, we can only speculate why Calfee and Anderson did not find such an effect for digit-trigram recall in their experiment. Perhaps the absence of a presentation rate effect for digit-trigram recall in the Calfee and Anderson study was simply the result of happenstance (i.e., a type II error). Importantly, however, our results clearly show that presentation rate can have a substantial effect on subsequent recall performance, even when learners are required to recall a pre-experimentally unfamiliar target item.

General Discussion

In the present study we investigated the effect of presentation rate on the learning and retention of foreign language vocabulary. We manipulated the presentation rate while keeping total study time for each translation pair constant. In all experiments, students (speakers of English) studied Dutch - English vocabulary pairs. In Experiment 1 we found a non-monotonic relation between presentation rate during study (16 × 1 s, 8 × 2 s, 4 × 4 s, 2 × 8 s or 1 × 16 s) and recall of translation equivalents thereby replicating and extending the findings of de Jonge et al. (2012). Overall performance was best for presentation rates of around 4 s with presentation rates of 1 s and 16 s resulting in significantly lower recall performance. Experiment 2 extended these findings to a retention interval of 1 day indicating that the manipulation of presentation rate affected long-term recall. Of primary interest, a non-monotonic relation between presentation rate and recall was present both when subjects had to recall the English translation equivalent of a Dutch word (Dutch → English) and when they had to recall the Dutch translation equivalent of an English word (English → Dutch).

The main question of interest in the present study was whether presentation rate would have an effect on the learning and retention of the English → Dutch translation pairs. As noted, results from previous research on paired associate learning of digit-trigram pairs (e.g., *91-LUB,* Calfee & Anderson, 1971) indicated that presentation rate can sometimes have little systematic effect on subsequent cued recall of pre-experimentally unfamiliar target items. Our finding of a sizeable effect of presentation rate on translation recall for both translation directions (Dutch → English and English → Dutch) caused us to question the reliability of the findings of Calfee and Anderson. In Experiment 3 we therefore re-examined the effect of presentation rate on digit-trigram recall. Again, we found a non-monotonic effect of presentation rate on recall performance. Based on these findings and previous ones (e.g., de Jonge et al., 2012; Johnson, 1964; Stubin et al., 1970) we conclude that presentation rate has a substantial influence on paired-associate learning across a range of materials.

Two factors may help to explain why presentation rate affects paired-associate learning. A first factor is related to what de Jong et al. (2012) referred to as the *effective study time hypothesis*. According to this hypothesis some minimal amount of time is necessary to optimally form an association between two stimuli (Stubin et al., 1970). Elaborative encoding processes such as forming a mental image, generating a sentence or other processes connecting the elements of two stimuli take a certain amount of time (e.g., Bugelski, Kidd, & Segmen, 1968). With short presentation times such processes cannot be effective employed, resulting in poor recall for fast presentation rates. On the other hand, long presentation times may result in decreased concentration and inattention (Bugelski & McMahon, 1971); beyond some optimum the remaining study time may be used in a less efficient way, causing relatively little additional information to be stored in memory (cf. Nelson & Leonesio, 1988).

A second factor concerns what happens when an item is repeated. According to several accounts, memory benefits from repetitions if the initial encoding of an item in memory is retrieved on a subsequent presentation of that item (e.g., Benjamin & Tullis, 2010; Raaijmakers, 2003). If retrieval of earlier encodings of an item benefits memory, more presentations should result in better memory. However, with total study time held constant, as was the case in the present study, it is reasonable to assume there is a trade-off. More presentations imply shorter presentation durations and short presentation durations may not allow for successful retrieval of earlier encodings of an item. A presentation duration of 1 s provides many potential opportunities for retrieval of an earlier encoding of the same item, but very few of these retrieval opportunities may be successful. A presentation time of 4 s, in contrast, may optimize the beneficial effect of successful retrievals by providing multiple opportunities for retrieval, several of which may be successful. With even longer presentation times, retrieval may often be successful but final recall performance is limited by the relatively small number of trials.

To summarize, the present study, as well as earlier ones (e.g., de Jonge et al., 2012; Stubin et al., 1970), suggest that a presentation rate in the order of 4 s results in optimal paired associate learning. These results were obtained in studies using different types of study materials. Moreover, the observation of a non-monotonic relation between presentation rate and cued recall performance has been obtained in studies that differed in several procedural aspects such as total study time, the length of the study list (i.e., the number of items studied) and the length of the retention interval, and is present both when presentation rate is manipulated between subjects as well as within subjects. Altogether, this suggests that the effect of presentation rate on cued recall is rather robust. It would, however, be premature to conclude that a presentation rate of around 4 s would be optimal for all types of study materials. Optimal presentation rates might be different for paired associates that, in addition to verbal stimuli, involve nonverbal stimuli (e.g., learning the names of anatomical structures or learning the names of the countries on the map of Africa) or even for vocabulary learning of languages that do not use alphabetic writing systems such as Chinese. Our results do indicate, however, that presentation rate during study has a substantial effect on subsequent memory. Moreover, it seems that the optimal presentation rate during learning does not necessarily shift with difficulty of recall. Even though translation recall of foreign language words (English→Dutch) was much lower than that of familiar language words (Dutch→English), presentation rate had similar effects on both language directions. For alphabetic language materials, at least, a presentation rate of 4 s seems a good rule of the thumb to achieve optimal paired-associate learning.

Chapter 4

The Efficacy of Self-Paced Study in Multitrial Learning[*]

[*] Chapter 4 was submitted for publication as: de Jonge, M., Tabbers, H. K., Pecher, D., Jang, Y., & Zeelenberg, R. (2014). The efficacy of self-paced study in multitrial learning.

Abstract

In two experiments we investigated the efficacy of self-paced study in multitrial learning. In Experiment 1, native speakers of English studied lists of Dutch-English word pairs under one of four imposed fixed presentation rate conditions (24 × 1 s, 12 × 2 s, 6 × 4 s, or 3 × 8 s) and a self-paced study condition. Total study time per list was equated for all conditions. We found that self-paced study resulted in better recall performance than most of the fixed presentation rates, with the exception of the 12 × 2 s condition which did not differ from the self-paced condition. Additional correlational analyses suggested that the allocation of more study time to difficult pairs than to easy pairs might be a beneficial strategy for self-paced learning. Experiment 2 was designed to test this hypothesis. In one condition, participants studied word pairs in a self-paced fashion without any restrictions. In the other condition, participants studied word pairs in a self-paced fashion, but total study time per item was equated. The results showed that allowing self-paced learners to freely allocate study time over items resulted in better recall performance.

Intuitively, giving learners control over the pacing of their own study seems the right thing to do. But is it really wise to give learners control? In general, literature on metacognition paints a pretty bleak picture concerning the decisions learners make during study. It has been argued that, in order to become an effective self-guided learner, one needs to go against certain intuitions and have a reasonably good understanding of the processes that underlie durable learning (Bjork, 1999; Kornell & Bjork, 2007). Unfortunately, people often do not understand all of the complexities of their own memory, and they have many metacognitive misconceptions about remembering and learning (Kornell & Bjork, 2009). Although research suggests that, in some situations, people do have accurate metacognitions, it is unclear if they are able to put this knowledge to use (Son & Metcalfe, 2000). Given that people may not be very good at making the right decisions during learning, a pessimist could argue that it might be best to take away control from learners as much as possible. On the other hand, it might be a bit rash to give up on the self-paced learner altogether. Although learners might not make optimal decisions during self-paced study, it is still not clear whether what they do is really that ineffective (Metcalfe & Kornell, 2003, 2005; Tullis & Benjamin, 2011). In the present study, therefore, we investigated to what extent learners are able to effectively allocate study time during multitrial learning. To this end, we compared a situation where learners have control over the allocation of study time to conditions where learners have no control.

Only a few studies have directly compared a self-paced condition to an experimenter-imposed fixed-pace condition. Moreover, these comparative studies on the effectiveness of self-paced study have come up with somewhat equivocal results (Tullis & Benjamin, 2011). For instance, in a study by Mazzoni and Cornoldi (1993), participants who self-paced their study rate showed better recall performance compared to those who studied words presented with a fixed pace (the average rate of presentation in the self-paced condition). However, Koriat, Ma'ayan, and Nussinson (2006) did not replicate this result. Furthermore, several of these studies (e.g., Koriat et al., 2006; Mazzoni & Cornoldi, 1993) incorporated test trials or asked participants for metacognitive judgments during study. Research has shown that test trials given during study are not merely neutral assessment trials, but can have a profound effect on later recall (see Roediger & Karpicke, 2006a for a review). The same argument has been made about judgments of learning and remember/know judgments. It has been suggested that, analogous to the Heisenberg uncertainty principle, measuring the state of memory during study may change the state of memory itself (Jönsson, Hedner, & Olsson, 2012; Kimball & Metcalfe, 2003; Naveh-Benjamin & Kilb, 2012; Spellman & Bjork, 1992).

Recently, Tullis and Benjamin (2011) investigated the effectiveness of self-paced study in isolation (without test trials or metacognitive judgments given during study) on later recognition test performance. In Experiment 1 of their study, one group of participants studied a list of words in a self-paced fashion. They could study each word for as long as they wanted before proceeding to the next item on the list. In the other condition, participants were yoked to one of the self-paced participants. The yoked control group did not have any control over study time; the presentation time of the words was determined by calculating the average presentation time per word of the previous participant in the self-paced condition. This way total study time was equated between the two study conditions. The results showed that self-paced learning resulted in better performance on a subsequent recognition test compared to the yoked control condition. In Experiment 2 of their study, this result was replicated and extended by showing that self-paced study was even more effective than a condition in which study time was allocated to individual items based on normative item difficulty (based on performance of the yoked control condition in Experiment 1). In addition to test performance, Tullis and Benjamin also looked at the study strategies used by the self-paced group. They noted that the advantage of self-pacing was apparent only in those participants who allocated more study time to the more difficult items. This strategy is often referred to in the literature as *discrepancy reduction* (Dunlosky & Herzog, 1998), and suggests that students try to cope with the experienced difficulty of items in a list by differentially allocating study time. Tullis and Benjamin's (2011) results thus seem to suggest that, during single trial learning, learners can be quite proficient when it comes to allocating study time.

Research on self-pacing and study time allocation has mostly focused on single-trial learning instead of multitrial learning. In practice, however, when students acquire new knowledge (e.g., foreign vocabulary or anatomy), they probably do not study each item just once. Rather, one would expect students to go over the materials multiple times before terminating study. Also, memory researchers have considered self-pacing mainly as an incidental procedural aspect of their experimental design rather than the object of actual investigation. Therefore, little is known about what learners actually do during multitrial self-paced study. Hence, for practical considerations as well as to extend existing theoretical frameworks, it is important to find out how effectively students allocate study time during multitrial learning.

In a review of the literature on self-regulated learning, Kornell and Bjork (2007) also reported their own data from a pilot experiment on multitrial learning, in which participants were instructed to study a list of word pairs multiple times during a 10-minute self-paced learning phase. The results showed that participants started out with a reasonably long (7.4 s) presentation rate per item during the first study cycle, but that they eventually ended up with a very fast (< 1 s) presentation rate by the last study cycle. Although the authors did not report any statistical analysis concerning these self-paced study data, as these were not their primary interest, the pattern of results suggests that learners increased the rate of self-paced presentations as learning progressed. On the one hand, one could argue that increasing the rate of presentation could be an effective strategy, because participants experienced a larger number of study trials than they would have if they had stuck to their initial presentation rate. On the other hand, research has also shown that, with total study time equated, a large number of very fast (e.g., 1 s) presentation rates results in suboptimal learning compared to a smaller number of intermediate (e.g., 4 s) presentation rates (de Jonge, Tabbers, Pecher, & Zeelenberg, 2012). The pilot experiment of Kornell and Bjork contained no fixed-paced control condition to which performance in the self-paced condition could be compared. Thus, it is still unclear whether learners' distribution of study time during multitrial self-paced learning is effective or not.

In the present study we investigated the effectiveness of self-paced study in a foreign vocabulary learning task. In Experiment 1, we investigated the efficacy of self-paced multitrial learning relative to fixed-pace multitrial learning (i.e., when presentation duration is determined by the experimenter and not under the control of the learner). Because presentation rate has a large influence on learning, even when total study time is held constant (de Jonge et al., 2012), we compared a variety of fixed-presentation rates to a condition where participants were allowed to self-pace. For the self-paced condition, we expected the study time per item to decrease across cycles. Also, we expected that more study time would be allocated to items of high normative item difficulty (discrepancy reduction). Most important, if it is beneficial to control study time allocation during multitrial self-paced learning, then self-pacing should result in better recall performance relative to the fixed presentation rates. In Experiment 2, we investigated whether or not differential allocation of study time over items is a crucial factor in self-paced study during multitrial learning with regards to later recall performance. To this end, we compared two self-paced study conditions: one in which participants were allowed to freely allocate study time over items (unrestricted) and one in which total study time per item was equated (restricted).

Experiment 1

Method

Participants

One hundred and twenty-eight undergraduate students at the University of California, San Diego, participated for course credit. The data from one participant were discarded because of a computer malfunction. This participant was replaced so that the design of the experiment remained completely counterbalanced across participants.

Materials

A total of forty-eight Dutch-English word pairs (e.g., *kikker - frog*) were used in the experiment. Translation pairs were noncognates, that is, the Dutch word and its English translation equivalent were orthographically and phonologically dissimilar. All words (both Dutch and English) were between 3 and 7 letters long, and consisted of one or two syllables. The mean word length of the Dutch words was 4.75 (*SD* = 1.23); the mean word length of the English words was 4.90 (*SD* = 1.01). The mean word frequency per million of the English words (Brysbaert & New, 2009) was 63.66 (*SD* = 115.75). The 48 word pairs were divided over four 12-item lists. E-prime (Psychology Software Tools, Pittsburgh, PA) was used to create and run the experiment.

Design and Procedure

We used a 2 × 4 mixed design with pacing (self-paced vs. fixed pace) as a within-subjects factor and fixed presentation rate (24 × 1 s, 12 × 2 s, 6 × 4 s, and 3 × 8 s) as a between-subjects factor. Each participant received the self-paced condition in combination with one of the four fixed presentation rate conditions. Half of the participants started with self-paced study followed by fixed-pace study; the other half received the opposite order. Participants were randomly assigned to one of the four fixed presentation rate conditions and to one of the two orders.

In the self-paced condition, participants studied a total of 24 word pairs divided over two lists during two consecutive self-paced study blocks. In each block, participants were given 288 s of total study time to learn a list of 12 items (i.e., an average of 24 s per word pair). Participants were told that they could determine the rate of individual study presentations. The instructions emphasized that each study block would take approximately 5 minutes to complete regardless of pacing. Word pairs were presented one at a time on the computer screen in a random order, and participants could progress to the next item by pressing the ENTER-key. If participants did not press the ENTER-key in the first 16 s of the block, a reminder appeared on the screen informing them that, if they wanted, they could use the ENTER-key to move on to the next pair. This was done because in a pilot study some participants studied the first presented word pair in the self-paced condition for a disproportionately large amount of time (perhaps due to a failure to carefully read or remember the instructions). Importantly, due to the reminder used in the present study, this problem did not reoccur. As discussed in the results section, most of the participants cycled through the study materials several times. All pairs on the list were presented once in a random order before the pairs were presented again in a different random order.

Upon completion of the two self-paced study blocks, participants first solved multiplication problems for 1 min as a distractor task and then were given a cued recall test. On the test, the 24 Dutch words were presented on the computer screen in a random order, one at a time, and participants were asked to type the correct English translations. The cued recall test was self-paced and participants could simply progress to the next item by pressing the ENTER-key.

In the fixed-pace condition, participants studied two lists of 12 word pairs during two consecutive study blocks. As in the self-paced condition, participants were given 288 s of total study time for each list. However, unlike the self-paced condition, participants had no control over the presentation rate. In the 24 × 1 s condition, each list of word pairs was presented 24 times with a presentation rate of 1 s per pair. In the 12 × 2 s condition, each list was presented 12 times with a presentation rate of 2 s per pair. In the 6 × 4 s condition, each list was presented six times with a presentation rate of 4 s per pair. Finally, in the 3 × 8 s condition, each list was presented three times with a presentation rate of 8 s per pair. All pairs on the list were presented once in a random order before the pairs were presented again in a different random order. Participants were informed in advance how many times each word pair would be presented and at what rate. They were also informed that each study block would take approximately 5 minutes to complete. Upon completion of the two fixed-pace study blocks, participants received a distractor task followed by a cued recall test. The procedure for the distractor task and cued recall task were identical to those in the self-paced condition.

A total of eight counterbalanced versions were used. Across participants, each word pair was presented equally often in each condition (i.e., self-paced vs. fixed pace), each of the fixed pace presentation rates, and each of the four study blocks.

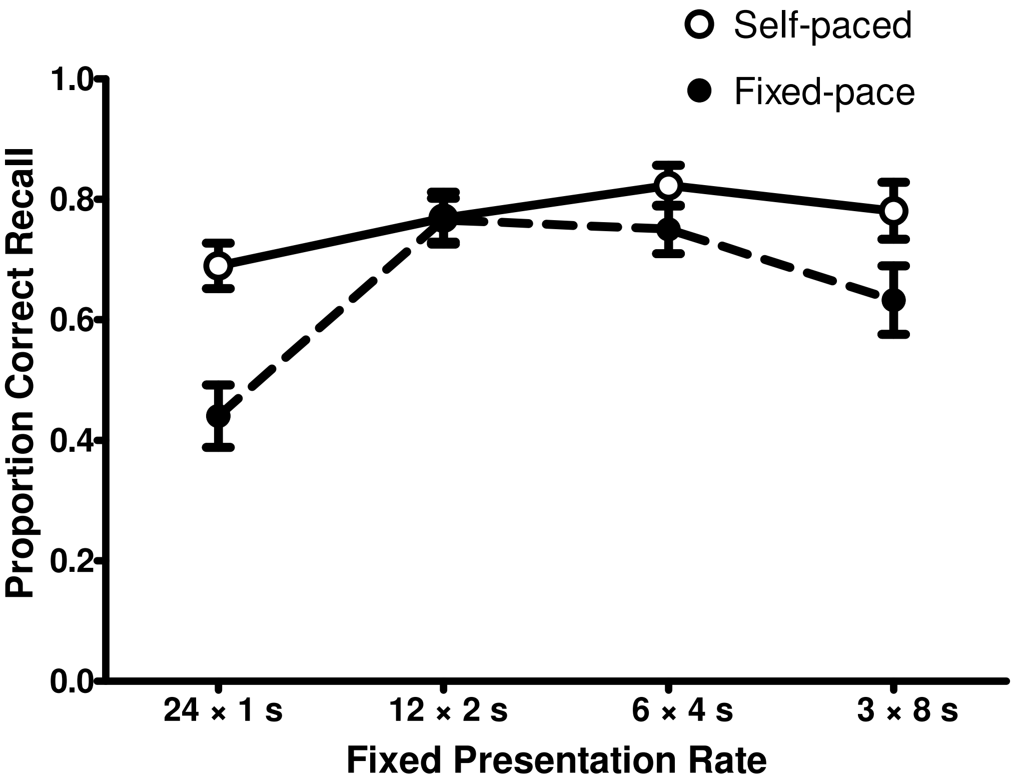

Figure 1. Proportion correct cued recall in Experiment 1 as a function of study condition (self-paced vs. fixed-pace) and fixed presentation rate group. Error bars represent standard errors of the means.

Results and Discussion

Recall Performance

Figure 1 shows the mean proportion of correct cued recall in Experiment 1. The results show that, overall, self-paced study resulted in higher performance than fixed-pace study. In all but one of the fixed-pace conditions participants recalled more words when they could determine the presentation durations themselves than when presentation rate was imposed by the experimenter. These observations were supported by a 2 × 4 mixed ANOVA with study condition (self-paced vs. fixed-pace) as a within-subjects factor and presentation rate (24 × 1 s, 12 × 2 s, 6 × 4 s or 3 × 8 s) as a between-subjects factor.[1] The ANOVA showed a main effect of study condition, $F(1, 124) = 45.06$, $p < .001$, $\eta_p^2 = .27$, indicating that overall, more words were recalled in the self-paced study condition than in the fixed-pace study condition. There also was a significant main effect of

[1] An initial ANOVA also included condition order (self-paced study first vs. fixed-paced study first). The main effect of condition order and all interactions involving condition order were nonsignificant (all *p*s > .30). Condition order was therefore not included in the analyses reported here.

presentation rate, $F(3,\ 124) = 6.14$, $p < .001$, $\eta_p^2 = .13$. Importantly, however, these main effects were qualified by a significant study condition × presentation rate interaction, $F(3, 124) = 8.92$, $p < .001$, $\eta_p^2 = .18$, indicating that the difference between self-paced and fixed-pace study was not the same for each presentation rate. Follow-up analysis revealed that recall performance was unaffected by presentation rate for the self-paced condition, $F(3, 124) = 1.85$, $p = .14$. However, in the fixed-pace condition, there was a significant effect of presentation rate, $F(3, 124) = 10.15$, $p < .001$, $\eta_p^2 = .20$. This was to be expected because presentation rate was manipulated for the fixed-pace condition, but not for the self-paced condition. Note that the inverted U-shape relation between presentation rate and recall performance in the fixed-pace condition observed in Figure 1 is in line with earlier research on the effect of presentation rate on recall (e.g., de Jonge et al., 2012).

In subsequent analyses, we compared test performance in the self-paced condition to that in the fixed-pace condition for each of the presentation rates separately. For participants in the 24 × 1 s condition, performance in the self-paced condition was better than that in the fixed-pace condition, $t(31) = 6.12$, $p < .001$, $d = 1.14$. For participants in the 12 × 2 s condition, performance in the self-paced condition did not differ from that in the fixed-pace condition, $t(31) < 1$. For participants in the 6 × 4 s condition, performance in the self-paced condition was better than that in the fixed-pace condition, $t(31) = 2.21$, $p < .05$, $d = 0.40$. Finally, for participants in the 3 × 8 s condition, performance in the self-paced condition was better than that in the fixed-pace condition, $t(31) = 4.28$, $p <. 001$, $d = 0.78$. Thus for all but the 12 × 2 s condition, participants performed better in the self-paced condition than in fixed-pace condition.

Self-Paced Study

In order to gain insight in how people had distributed study time during self-paced study and how this may have affected their learning outcomes, we took a closer look at study behavior during the self-paced study blocks. Figure 2 shows the average self-paced study time per item as a function of study cycle for the first 10 cycles. As is clear from the figure, the average study time per item decreased across cycles. Study time per item decreased rapidly at first and subsequently decreased more slowly. For practical considerations (i.e., because different participants completed a different number of study cycles), in our statistical analysis we compared only the first, second, and last full cycle of each participant. Data were collapsed across study blocks and for participants with missing data, cases were excluded listwise. The data were analyzed using a repeated measures ANOVA. Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of study cycle. Degrees of freedom were corrected using Greenhouse-Geiser estimates of sphericity. There

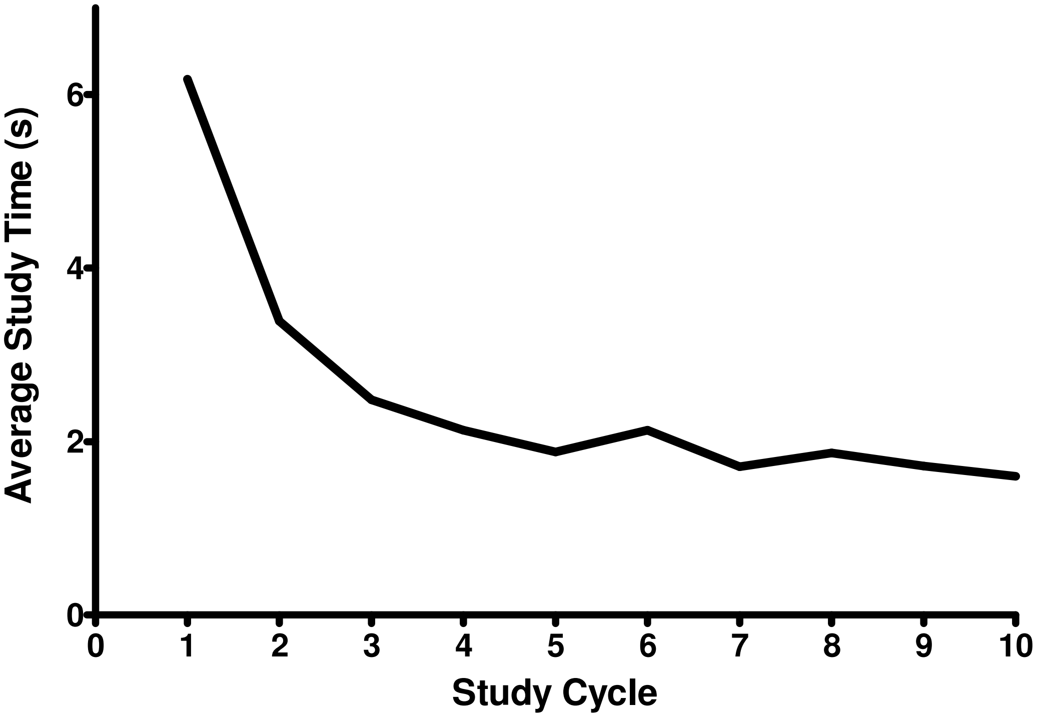

Figure 2. Self-paced study time per item in Experiment 1 as a function of study cycle averaged over participants.

was a significant effect of study cycle, $F(1.55, 170.72) = 98.92$, $p < .001$, $\eta_p^2 = .47$. Follow-up analysis showed that study time decreased from the first to the second cycle, $F(1, 110) = 55.17$, $p < .001$, $\eta_p^2 = .33$, as well as from the second to the last full cycle, $F(1, 110) = 81.55$, $p < .001$, $\eta_p^2 = .43$.

Allocation of Self-Paced Study Time

Figure 3 shows the average total study time for each item in the self-paced condition, plotted against normative item difficulty (defined as 1 minus the average proportion correct recall for the same item in the fixed-pace conditions, for a similar procedure, see Tullis and Benjamin, 2011). As can be seen in the figure, there was a strong positive correlation between self-paced study time allocated to the word pairs and normative item difficulty, $r(46) = .68$, $p < .001$. This finding is in line with the general finding that participants tend to allocate more self-paced study time to the more difficult items (Son & Metcalfe, 2000).

To sum up, in Experiment 1, we found that self-paced study resulted in relatively good performance compared to a variety of fixed-pace study conditions. Except for the 12 × 2 s condition, where recall performance was more or less equivalent, having control over pacing and study time allocation resulted in a significant recall advantage on a later test. One possible explanation for the results of the present experiment could be related to the allocation strategy

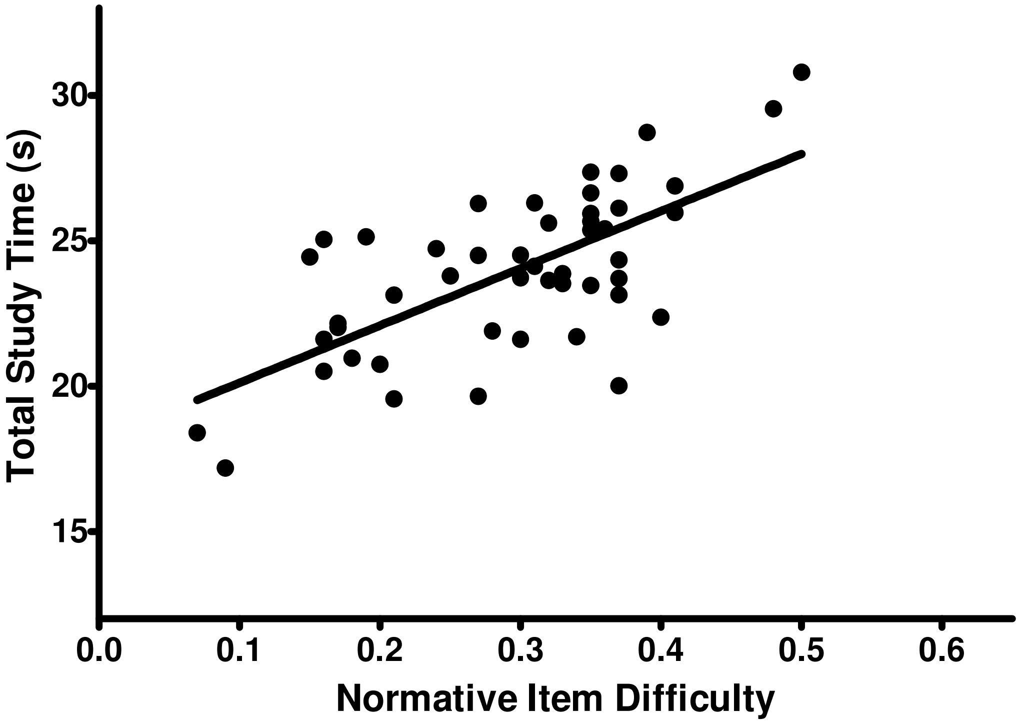

Figure 3. Average total study time for each item in the self-paced condition in Experiment 1 plotted against normative item difficulty (1 minus the average proportion correct recall in the fixed-pace conditions).

employed by learners in the self-paced condition. In the present study, we replicated the general finding that learners tend to allocate more self-paced study time to the more difficult items (e.g., Dunlosky & Herzog, 1998; Nelson & Leonesio, 1988). As already noted, in the Tullis and Benjamin (2011) study, the benefit of self-pacing was apparent only for those participants that were classified as discrepancy reducers. Likewise, in the present experiment, we explored the relationship between the degree of discrepancy reduction and subsequent recall performance in the self-paced condition. For each participant, we calculated the correlation across items between normative item difficulty and total study time allocated to each item. A more positive correlation indicated a higher degree of discrepancy reduction because more time was spent on items of higher normative difficulty. The data showed that 110 out of 128 participants (86%) in the present experiment could be classified as discrepancy reducers, in the sense that these participants spent more study time on the more difficult items. Analysis revealed that there was a significant correlation between the degree of discrepancy reduction and subsequent recall performance, $r(126) = .38$, $p < .001$. In other words, participants who displayed a strong tendency to allocate more study time to items of high normative difficulty recalled more items than participants who displayed only a weak (or no) tendency.

Experiment 2

In Experiment 1 we found that recall performance following self-paced study was at least as good and in most conditions even better than fixed-pace study. Moreover, the data suggested that one possible advantage of self-pacing study could be related to differential allocation of study time (discrepancy reduction). Experiment 2 was designed to further test this hypothesis. To this end we compared an unrestricted self-paced condition, virtually identical to the one used in Experiment 1 (in which the total amount of available study time could be freely distributed over items), to a restricted self-paced condition where the total study time per item was equated. If discrepancy reduction (differential study time allocation) is a beneficial strategy then one would expect that self-pacing without the opportunity to differentially allocate study time over items would result in lower recall performance compared to self-pacing without any restrictions.

Method

Participants

Forty-four undergraduate Psychology students at the University of California, San Diego participated for course credit. None of the participants had participated in Experiment 1.

Materials, Design, and Procedure

The materials were identical to those used in Experiment 1. We used a within-subjects design with study condition (unrestricted vs. restricted) as independent variable. For both unrestricted and restricted study conditions, the procedure was identical to that of the self-paced study condition of Experiment 1 except as noted.

In the restricted self-paced condition the total study time per item was fixed. Participants were told that each item was allocated 24 s of total study time, and that as soon as the study time for an item had run out, the program would automatically terminate the presentation and continue to the next item. We anticipated that this procedure could result in a rather unpredictable study experience from the participants' perspective. Hence, to indicate that the time for an item had almost expired, the word pair changed color (from blue to red) during the final 1000 ms of total study time. Items for which the total amount of available study time had expired, did not reappear for further study.

In the unrestricted self-paced condition, participants were free to differentially allocate study time to the different items in the list. Participants simply studied the entire list of items continuously until the total study time

(288 s) for the list had run out. To equate as much as possible with the restricted condition, items were presented in red during the final 12 seconds of the total study time to indicate that time had almost expired.

Immediately following each of the self-paced conditions (every two blocks), participants first received a 5-min distractor task solving multiplication problems followed by a cued recall test. Four counterbalanced versions were created in the same general manner as in Experiment 1.

Results and Discussion

Recall Performance

Proportion correct recall was .71 (*SD* = .26) in the unrestricted self-paced condition versus .61 (*SD* = .24) in the restricted self-paced condition. A t-test for paired samples showed that the difference between the two conditions was significant, $t(43) = 2.76$, $p < .01$, $d = .42$. Thus, withholding the possibility to differentially allocate total study time to the items in the lists during self-paced study resulted in lower recall performance.

Self-Paced Study

As in Experiment 1, we also looked at self-paced study behavior. Figure 4 shows the average study time per item as a function of study cycle and study condition for the first 10 cycles averaged over participants. The pattern of study times across cycles for the self-paced conditions was similar to the pattern observed in Experiment 1. In both conditions, study time per item rapidly decreased at first and then leveled off. Secondly, as is also clear from the figure, the average study times in the first study cycle were somewhat larger in the restricted condition compared to the unrestricted condition. Note that the dropping of items from the lists in the restricted condition resulted in increasingly shorter lists of items in this condition, while in the unrestricted condition lists remained intact throughout the study phase. Thus, a direct comparison of the study times in the two self-paced study conditions is problematic. For practical considerations, we only compared the study times in the first study cycle. A paired-samples t-test confirmed that there was a significant difference between the two conditions in the first cycle, $t(43) = 3.31$, $p < .01$, $d = 3.31$.[3]

[3] Closer inspection of our data revealed that, in the restricted self-paced condition, eight of the participants used the total amount of available study time (24 s) for at least half of the items in the very first study cycle. Since this self-imposed strategy might have disadvantaged recall performance for these participants in the restricted self-paced condition, we conducted an additional exploratory analysis that excluded these participants. In this analysis, we still found a recall benefit for the unrestricted over the restricted self-paced condition, $t(35) = 2.30$, $p < .05$, $d = 0.38$.

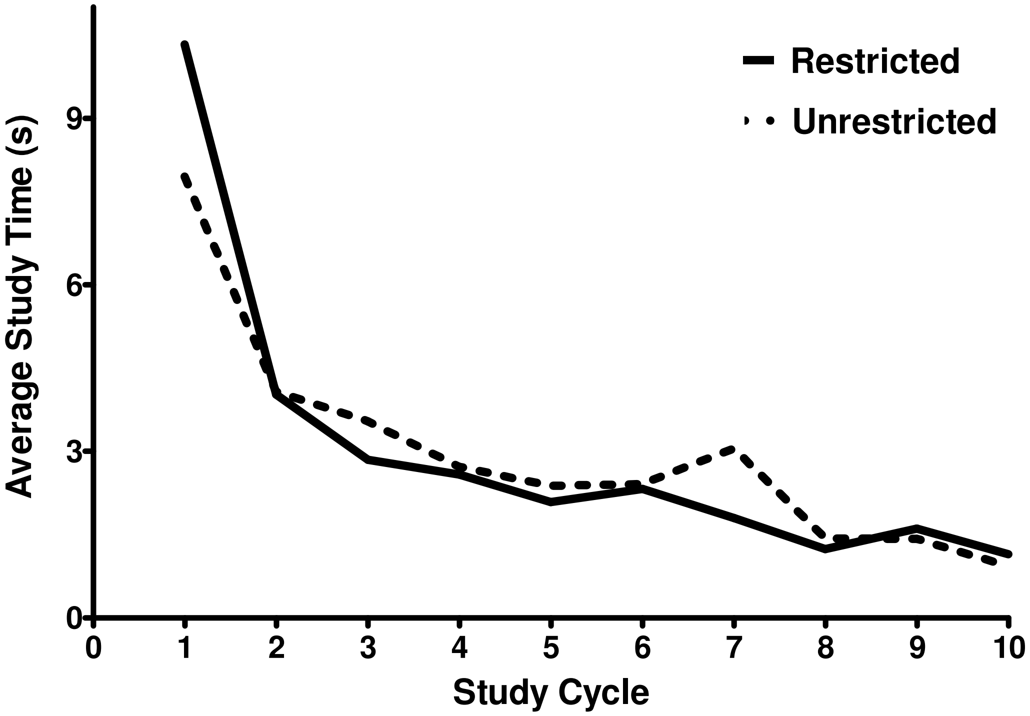

Figure 4. Self-paced study time per item in Experiment 2 as a function of study cycle and study condition averaged over participants.

As in Experiment 1, we analyzed the decrease in study times in the unrestricted condition during the first, second, and last full cycle of each participant. Data were collapsed across study blocks and for participants with missing data, cases were excluded listwise. The data were analyzed using a repeated measures ANOVA. Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of study cycle. Degrees of freedom were corrected using Greenhouse-Geiser estimates of sphericity. There was a significant effect of study cycle, $F(1.44, 46.12) = 20.88$, $p < .001$, $\eta_p^2 = .40$. Follow-up analysis showed that study time decreased from the first to the second cycle, $F(1, 32) = 16.07$, $p < .001$, $\eta_p^2 = .33$, as well as from the second to the last full cycle, $F(1, 32) = 12.42$, $p < .005$, $\eta_p^2 = .28$. For the restricted condition, analysis of the study times across cycles was limited to the first and second cycle. We did not look at the last cycle, because the durations of the last presentations for items in the restricted self-paced condition were not under the participants' control. The data were analyzed using a paired samples t-test. As in the unrestricted condition, there was a significant decrease in study times from the first to the second study cycle, $t(39) = 6.32$, $p < .001$, $d = 1.0$.

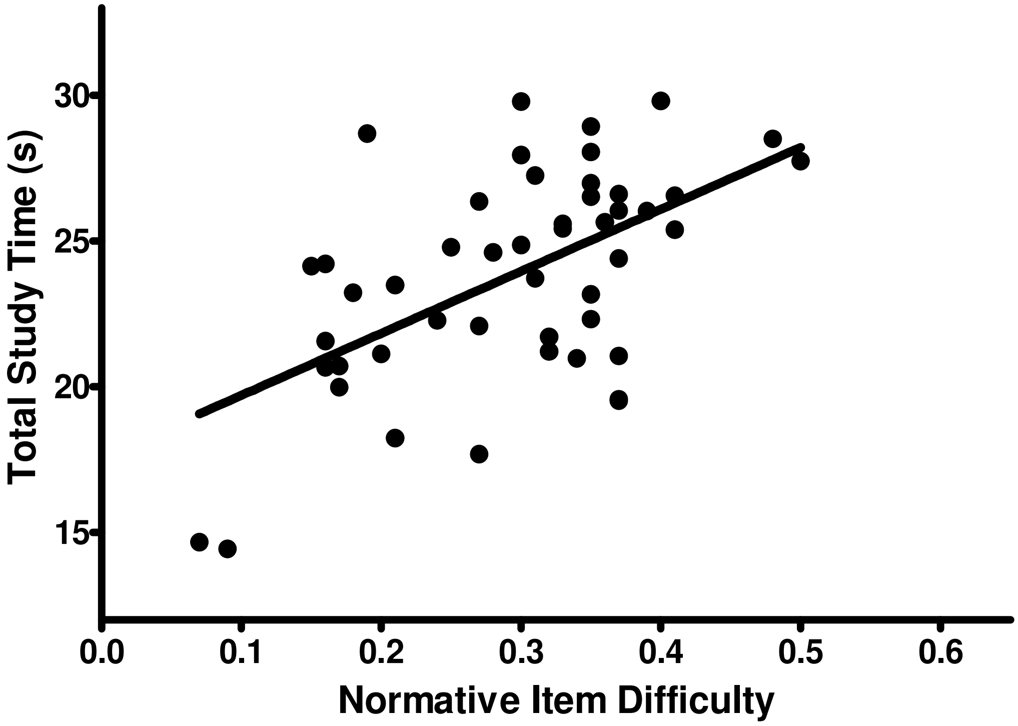

Figure 5. Average total study time for each item in the unrestricted self-paced condition of Experiment 2 plotted against normative item difficulty (1 minus the average proportion correct recall in the fixed-pace conditions in Experiment 1).

Allocation of Self-Paced Study Time

Figure 5 shows the average total study time for each item in the unrestricted condition, plotted against normative item difficulty (i.e., 1 minus the average proportion correct recall for that item in the fixed-pace conditions of Experiment 1). Again, we found a positive correlation between self-paced study time allocated to word pairs and normative item difficulty, $r(46) = .56, p < .001$, indicating that participants differentially allocated total study time to items as a function of normative item difficulty in the unrestricted self-paced condition.

Like in Experiment 1, we also evaluated the relationship between the degree of discrepancy reduction and subsequent recall performance. We found that 36 out of 44 participants (i.e., 82%) spent more study time on the more difficult items. Also, there was a correlation between the degree of discrepancy reduction and subsequent recall performance, $r(42) = .27, p < .05$ (one-tailed). Thus, a larger degree of discrepancy reduction tended to be associated with better recall performance in the unrestricted self-paced condition.

General Discussion

In the present study we investigated the effectiveness of self-paced study during multitrial learning. In Experiment 1 we found that self-paced study resulted in higher performance than fixed-pace study. In all but one of the fixed-pace conditions having control over pacing and study time allocation resulted in a significant recall advantage. Experiment 1 also showed that participants allocated more self-paced study time to normatively more difficult items. In Experiment 2, we found evidence suggesting that the opportunity to allocate more study time to the more difficult items on a list can be one important factor determining later test performance. That is, test performance deteriorated when total study time per item was equated during self-paced study. Taken together, our results suggest that learners can be proficient when it comes to allocating self-paced study time during multitrial learning.

One particularly consistent result across the two experiments in the present study was the positive correlation between normative item difficulty and the amount of study time allocated to items. In both experiments, we found that the majority of the participants (82%, and 70%, respectively) allocated more self-paced study time to the more difficult items. The finding that learners tend to devote more study time to the more difficult items is in line with earlier research (see Son & Metcalfe, 2000 for a review). It has been suggested that, by differentially allocating study time, learners try to compensate for the experienced difficulty of items in a list (Dunlosky & Herzog, 1998). Although, at first glance, this might seem like a logical strategy to improve learning, some studies on study time allocation suggest that it could in fact be suboptimal. For instance, it has been argued that learners are often unable to successfully compensate for the difficulty of items in a list and that allocating more study time to difficult items often yields little or no gain in later recall performance (Mazzoni & Cornoldi, 1993; Mazzoni, Cornoldi, & Marchitelli, 1990; Nelson & Leonesio, 1988). This finding has lead researchers to suggest that the strategy of allocating more study time to items in a list during self-paced study might be *labor-in-vain* (Nelson & Leonesio, 1988). Taking this point even further, some researchers have even suggested that metacognitive self-monitoring itself might be labor-in-vain (Begg, Martin, & Needham, 1992). Clearly, these claims seem hard to reconcile with the results from the present study and those of other recent studies (e.g., Tullis & Benjamin, 2011), where learners saw a return on their investment rather than having labored in vain.

One possible explanation for these seemingly conflicting results could be related to the research designs employed in some of the earlier experiments on study time allocation during self-paced learning. First of all, as we already noted, most of the earlier studies have investigated the effects of study time allocation during single-trial learning instead of looking at multitrial learning. Interestingly, some researchers have already suggested that the labor-in-vain effect might disappear during multitrial learning (Nelson & Leonesio, 1988). Second, and more important, earlier research on the effect of study time allocation has mostly focused on correlational evidence for the relationship between normative item difficulty, study time allocation, and subsequent recall performance (i.e., the finding that participants allocate more study time to normatively difficult items, yet recall these items less often than normatively easy items to which less study time is allocated). Although these correlational data have provided important insights about the kind of strategies learners employ during self-paced study, they do not enable us to answer the question whether what people do is effective or not. For instance, self-paced learners might be allocating study time effectively, and, at the same time, show a negative correlation between allocated study time and subsequent recall test performance. The extra time invested may not fully compensate for differences in item difficulty, but still improve overall memory performance. Moreover, if learners use a discrepancy reduction strategy and the resulting correlation between item difficulty and subsequent item recall is equal to or greater than zero, this still does not imply greater efficacy. The observed correlation could suggest that learners were able to effectively compensate for the difficulty of the materials (increased recall of difficult items). However, at the same time, it could reflect deteriorated recall of the easier items. Thus, an experimental manipulation is essential to ascertain a causal relationship between study time allocation and subsequent recall performance. Our study provides experimental evidence suggesting that, to a certain extent, learners are able to allocate study time effectively during multitrial self-paced learning. Although the benefits of differential study time allocation might not become apparent in a correlational design focusing on normative item difficulty, the results from our study show that, on an idiosyncratic level, self-paced learners can *effectively* compensate for some of the experienced difficulty of items in a list. When learners are forced to indiscriminately use an equal amount of study time for all items, their recall performance will deteriorate.

In addition to recall performance and the allocation of study time we also looked at the presentation rate during self-paced learning over the course of the consecutive study cycles. As expected, we found that learners increased the rate of presentation over study cycles during self-paced learning confirming earlier observations of Kornell and Bjork (2007) and we provided statistical evidence supporting this conclusion. In both Experiment 1 and 2, participants tended to speed up presentation rate as learning progressed.

Research focussing on the effect of presentation rate has mostly focused on the situation where learners study materials with a constant fixed rate (e.g., de Jonge et al., 2012; Stubin, Heimer, & Tatz, 1970). These studies have shown that presentation rate can have profound effects on later recall performance. For instance, de Jonge et al. (2012) found that, with total study time equated, both slow and fast presentation rates can result in poor recall performance compared to intermediate presentation rates. This finding was coined *the Goldilocks principle of presentation rate*. De Jonge et al. (2012) have proposed *the effective study time hypothesis* to account for this finding. That is, they suggest that some minimal amount of time is necessary to optimally form an association (see also Stubin et al., 1970). However, presentation rates beyond some optimal value might cause inattention, decreased concentration, and boredom (see also Bugelski & McMahon, 1971). In that case, the excess study time of a presentation beyond some optimal value is not effectively utilized and might be better spent when set aside for later presentations. However, since studies investigating the effects of presentation rate have focused solely on situations where learners study with a fixed constant presentation rate, it would be interesting to investigate the situation where a pattern of study time durations is used similar to the one observed for the self-pacers in the present study (e.g. increased pacing across cycles). A relatively long (8 s) presentation rate might be optimal to form an association in the first study cycle of a learning sequence, however, it could be suboptimal (too slow) for presentations during subsequent study cycles.

To conclude, the results from the present study seem to rehabilitate the self-paced learner concerning the allocation of study time policy employed during multitrial learning. In both Experiments 1 and 2 of the present study, we found evidence suggesting that self-pacers allocated more total study time to the more difficult items. This is in line with the idea that learners try to compensate for the difficulty of the materials by differentially allocating study time (Dunlosky & Herzog, 1998). Although it has been suggested that differential study time allocation can be considered labor-in-vain (e.g., Nelson & Leonesio, 1988), we found that overall recall performance was actually relatively good when participants were allowed to differentially allocate study time. Moreover, both experiments indicated that participants displaying a strong tendency to allocate more study time to items of high normative difficulty recalled more items than participants who did not display such a tendency. Naturally, we would not want to suggest that learners are able to fully compensate for the difficulty of to-be-learned materials. However, to a certain extent, learners seem well able to discriminate between items of differential difficulty and allocate study time accordingly in a way to be effective.

Chapter 5

Using Test Trials to Improve Learning and Retention of Foreign Vocabulary[*]

[*] The data presented in Chapter 5 were collected by the first author during an internship under the supervision of René Zeelenberg in partial fulfillment of the requirements for the degree of Master of Science at Erasmus University Rotterdam. This chapter is in preparation for publication as de Jonge, M., Tabbers, H. K., Pecher, D., & Zeelenberg, R. (in preparation). Using test trials to improve learning and retention of foreign vocabulary.

Abstract

In the present study, we investigated the effect of testing on long-term retention of foreign vocabulary word pairs. Word pairs were learned under repeated testing, alternated testing, or a restudy (control) condition. In Experiment 1, we found that taking tests during learning slowed down the rate of forgetting over a 1-week interval compared to a restudy (control) condition. In Experiment 2, using an extended retention interval of four weeks, we replicated the finding that testing can slow down the rate of forgetting and we showed that, after the extended 4-week interval, the respective forgetting functions crossed over. On the 4-week final retention test, both the repeated tests and alternated tests condition outperformed the restudy (control) condition. Taken together, the results of our study provide a clear demonstration of the powerful effect retrieval practice can have on long-term retention. Furthermore, our results indicate that the benefit of retrieval practice can get more pronounced as the retention interval gets longer.

An important goal of education is not only to enhance the initial learning of new materials, but also to enhance its long-term retention. Although repeated study of information can result in fairly good recall performance in the short term, the memorial shelf life of repeatedly studied information is often short-lived. Indeed, in some of our own experiments on optimal study time distribution and subsequent retention, we observed substantial degrees of forgetting over retention intervals of just a few days (e.g., de Jonge, Tabbers, Pecher, & Zeelenberg, 2012; Zeelenberg, de Jonge, Tabbers, & Pecher, 2013). For instance, the results of de Jonge et al. (2012) indicated that, even in the best performing study conditions, more than half of what had initially been learned was forgotten within just two days. Thus, although repeated study can be an efficient way of encoding information (i.e. learning), it might not the most effective strategy for keeping information accessible over the long-term (i.e., retention). In the present study, we investigated the effect of one strategy, *retrieval practice*, which holds great potential for enhancing long-term retention and retarding forgetting.

The retrieval practice effect (also known as *the testing effect*) is a well-established phenomenon in the literature on learning and retention (for a review see Roediger & Karpicke, 2006a). In a typical study by Wheeler, Ewers, and Buonanno (2003), one group of participants repeatedly studied a list of words during four consecutive study cycles (the study condition), while another group of participants studied a list of words once, followed by three consecutive recall tests without feedback (the test condition). Not surprisingly, since the participants in the study condition were given more time to study during initial learning, they outperformed the test condition on a final recall test given after a short (5-min) retention interval. Importantly, however, on a final test given one week later, the tables had turned. That is, the test condition outperformed the study condition, indicating that practicing retrieval during learning can effectively slow down the rate of forgetting for successfully retrieved information.

In most previous testing effects studies, the effect of retrieval practice on the rate of forgetting has been investigated over intervals ranging from a couple of days up to 1 week (e.g., Wheeler et al., 2003; Roediger & Karpicke, 2006b; Toppino & Cohen, 2009). Surprisingly few studies have investigated the rate of forgetting over intervals beyond a 1-week interval. In one classic study by Spitzer (1939), the effect of testing on retention was investigated across an interval up to 62 days. Although the results from Spitzer's study suggested that testing improved long-term retention, he did not provide statistical analysis to confirm this observation. Furthermore, the Spitzer study did not include a restudy (control) group and the observed benefit of testing was relative to a situation where learners were not re-exposed to the study materials at all. Thus, the benefit of testing observed in his study might have been in part due to additional exposure to the materials in the testing conditions (see also Roediger

& Karpicke, 2006b). Other, more recent studies on the effects of testing in actual educational settings have also looked at recall performance after relatively long intervals of up to a couple of months (e.g., Carpenter, Pashler, & Cepeda, 2009; Roediger, Agarwal, McDaniel, & McDermott, 2011). The results of these studies also indicate that testing can benefit delayed recall test performance after relatively long intervals. However, because in these studies recall performance was assessed at a single point in time, these studies do not inform us about the rate of forgetting.

In one study, Carpenter, Pashler, Wixted, and Vul (2008) investigated whether test trials (with feedback) can reduce forgetting relative to restudy trials over a period of 6-weeks. In Experiment 1 and 2, participants learned obscure facts under study and testing conditions, and in Experiment 3, Swahili-English word pairs were used as stimuli. Recall of facts/word pairs was tested at six different points in time ranging from 5 min up to 42 days after initial learning. The difference in rate of forgetting between the study and the testing condition was explored using both an ANOVA-based approach (which is the more traditional method used in most previous studies), and a curve fitting method (which is an alternative approach using a mathematical characterization of forgetting). In two out of three experiments, the curve fitting method suggested that test trials reduced forgetting more than restudy trials. Interestingly, however, the results from the more traditional ANOVA-based approach sometimes led to different conclusions. That is, based on the ANOVA approach, in just one out of three experiments, test trials slowed down the rate of forgetting. Carpenter et al. (2008) note that, in their study, the tendency for tests to reduce forgetting was less pronounced compared to the effects reported in prior studies (e.g., Roediger & Karpicke, 2006b; Wheeler et al., 2003), and they suggest that the use of feedback, which was one factor in which their study differed from previous studies, might be the factor responsible for this apparent discrepancy. Thus, one limitation of the Carpenter et al. (2008) study is that they did not include a testing without feedback condition.

Few studies have directly compared the long-term benefits of repeated testing with feedback or restudy opportunity relative to retrieval practice without feedback or restudy opportunity. However, there are reasons why one might expect that testing can be especially beneficial when learners are given the opportunity to restudy following (i.e. when study and test trials are alternated during learning). For instance, research suggests that attempting to retrieve information on test trials may also facilitate later encoding of that information within the same learning session, even when the retrieval attempt was unsuccessful (e.g., Arnold & McDermott, 2013; Izawa, 1966). However, aside from this immediate short-term benefit, very little is known about the long-term benefits of alternated study and test trials. In one study by Thompson, Wenger, and Bartling (1978) the effect of retrieval practice on the rate of forgetting across a 2-day retention interval was investigated. Participants studied lists of words under one of three conditions. In all three conditions, the words were first presented once, before the procedures for the respective conditions diverged. In one condition, participants subsequently received three presentation trials (the multiple-presentation condition). In the other condition, participants received three subsequent recall trials without feedback (the multiple-recall condition). Finally, in the last condition, participants received three subsequent recall plus re-presentation trials. That is, in this condition, every recall trial was followed by re-presentation of the unrecalled items. Surprisingly, Thompson et al. found no reliable difference in rate of forgetting between the multiple-presentation and the recall plus re-presentation condition. However, for the multiple-recall condition, the rate of forgetting was slowed down relative to both the multiple-presentation and the recall plus re-presentation condition. These results indicate that, in terms of retention, repeated testing might be a more effective strategy than repeated testing with re-presentation. This might also explain why the retention benefit for the testing with feedback condition in the Carpenter et al. (2008) study was less pronounced compared to other studies looking at testing without feedback or re-presentation trials.

In short, as has been emphasized in the literature, it is important both for theoretical as well as practical purposes to establish to which extent testing can slow down the rate of forgetting especially over longer periods of time (e.g., Carpenter et al., 2008). In the present study we investigated the effect of testing on long-term retention of paired associates. In Experiment 1, we looked at the rate of forgetting over a 1-week interval. In Experiment 2, we looked at the rate of forgetting over an extended period of 4 weeks. Also, as noted, few studies have directly compared the respective retention benefits of repeated testing with and without the opportunity to restudy the materials. Therefore, in the present study, we compared the respective effects on the rate of forgetting of repeated testing without restudy, alternated study and test trials, and a restudy (control) condition.

Experiment 1

Method

Participants

Sixty students from the Erasmus University Rotterdam participated for course credit or a small monetary reward (€ 7.00). Participants were randomly assigned to one of three conditions: the restudy (control) condition, the alternated tests condition or the repeated tests condition.

Materials

Forty-eight Swahili-Dutch translation pairs (e.g., *gari-auto [car], joka-slang [snake]*) were used in the present experiment. The mean word frequency per million for the Dutch words (Keuleers, Brysbaert, & New, 2010) was 898.46 (*SD* = 89.93). The mean word length of the Dutch words was 4.8 (range 3-7 letters). The mean word length of the Swahili words was 5.1 (range 4-7 letters).

Design and Procedure

We used a 3 × 2 mixed design with learning condition (restudy, alternated tests, and repeated tests) as a between-participants factor and retention test (1-min or 1-week) as a within-participants factor. Participants in the restudy (control) condition were given six study blocks. During study blocks all 48 Swahili-Dutch translation pairs were presented in the center of the screen one at a time for 8 s each. The next pair was presented after an interval of 500 ms. Participants in the alternated tests condition were given alternating study and test blocks. The study blocks were identical to those of the restudy (control) condition. The test blocks consisted of a cued recall test in which all 48 Swahili words were presented and participants attempted to recall the Dutch translation equivalents. Participants entered their response on the keyboard. After 8 seconds the cue disappeared (regardless of whether participants had entered a response) and 500 ms later the next cue was presented. Finally, participants in the repeated tests condition started with three study blocks followed by three test blocks. The study and test blocks were identical to those of the other two conditions. Note that, total time on task was equated for the three learning conditions. In all study and test blocks, items were presented in a random order. New random orders were generated for each block and each subject.

Following the initial learning phase of the experiment, participants received the first of two retention tests. On the retention test, participants were shown half of the cue words one at a time in a random order. They were asked to type in the correct target words and they were told that they could progress to the next item by pressing the ENTER-key. One week later participants received the second retention test on the remaining half of the word pairs. Each subject received one of two counterbalanced lists. The lists of word pairs were assigned to retention interval conditions in such a fashion that both lists appeared equally often in both conditions.

Results and Discussion

Initial Learning

Table 1 shows how test scores evolved across cycles for both the repeated tests and the alternated tests condition. As can be seen in the table, there was a small,

Table 1

Mean Proportion of Correctly Recalled Target Words on the Three Initial Practice Tests as a Function of Testing condition in Experiments 1 and 2

Experiment and test cycle	Repeated tests	Alternated tests
Experiment 1		
T1	.59 (.27)	.20 (.13)
T2	.61 (.27)	.53 (.16)
T3	.62 (.28)	.82 (.13)
Experiment 2		
T1	.56 (.22)	.17 (.13)
T2	.56 (.25)	.46 (.22)
T3	.57 (.25)	.66 (.23)

Note. Standard deviations are given in parentheses. T1, T2, and T3 refer to the first, second, and third test cycle, respectively.

but gradual, increase in recall performance across cycles in the repeated tests condition. A repeated measures ANOVA revealed that the increase in recall performance was significant, $F(2, 38) = 6.23$, $p < .01$, $\eta_p^2 = .25$. Follow-up analysis showed a significant increase between the first and the second test cycle, $F(1, 19) = 7.28$, $p < .05$, $\eta_p^2 = .28$. However the difference between the second and the third test cycle did not reach the level of significance, $F < 1$. Note that the finding that recall performance can increase over the course of successive tests even without the use of corrective feedback is a well-documented phenomenon known as *the hypermnesic effect* (for a review see Payne, 1987). In the alternated tests condition, not surprisingly, there also was an increase in recall test performance across cycles. A repeated measures ANOVA revealed a significant effect of test cycle on recall performance, $F(2, 38) = 459.14$, $p < .001$, $\eta_p^2 = .96$. Follow-up analysis showed that there was a significant increase between the first and the second test cycle, $F(1, 19) = 397.69$, $p < .001$, $\eta_p^2 = .95$, as well as from the second to the third cycle, $F(1, 19) = 213.04$, $p < .001$, $\eta_p^2 = .92$.

Retention

Figure 1 shows the proportion of correctly recalled target words on the 1-min and the 1-week retention tests as a function of learning condition. As can be seen, the restudy (control) and alternated tests conditions resulted in better performance on

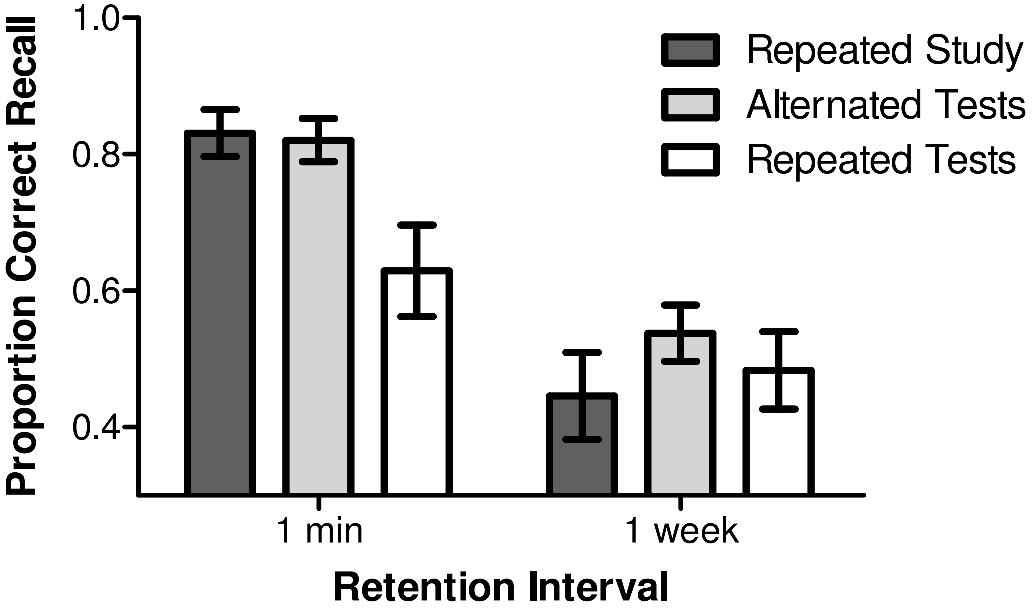

Figure 1. Proportion correctly recalled target words in Experiment 1 as function of learning condition and retention interval. Error bars represent standard errors of the means.

on the 1-min test than the repeated tests condition. However, after a 1-week retention interval, performance across the three conditions converged. A 3 × 2 mixed ANOVA revealed that there was a significant main effect of retention interval, $F(1,\ 57) = 128.99$, $p < .001$, $\eta_p^2 = .69$. The main effect of learning condition was not significant, $F(2,\ 57) = 1.17$, $p > .10$, $\eta_p^2 = .059$. More important, the interaction between learning condition and retention interval was significant, $F(1,\ 57) = 8.43$, $p < .001$, $\eta_p^2 = .23$, indicating that the absolute amount of information forgotten over the 1-week interval was different for the three learning conditions. Follow-up contrasts showed that the test conditions showed less forgetting over the 1-week interval compared to the restudy (control) condition, $t(57) = 3.86$, $p < .05$. Also the repeated tests condition showed less forgetting compared to the alternated tests condition, $t(57) = 2.10$, $p < .05$.

Our conclusion that forgetting differed for the different learning conditions was confirmed by additional analyses in which we looked at proportional forgetting rather than absolute forgetting. In these analyses, we calculated, for each subject, the number of correctly recalled target words after the 1-week retention interval as proportion of the number of correctly recalled target words after the 1-min retention interval. The use of proportional forgetting measures is consistent with power functions of forgetting such as Wickelgren's power law (Wickelgren, 1974; Wixted & Carpenter, 2007). Recently, this forgetting function has been used to fit memory performance and forgetting in studies on the testing

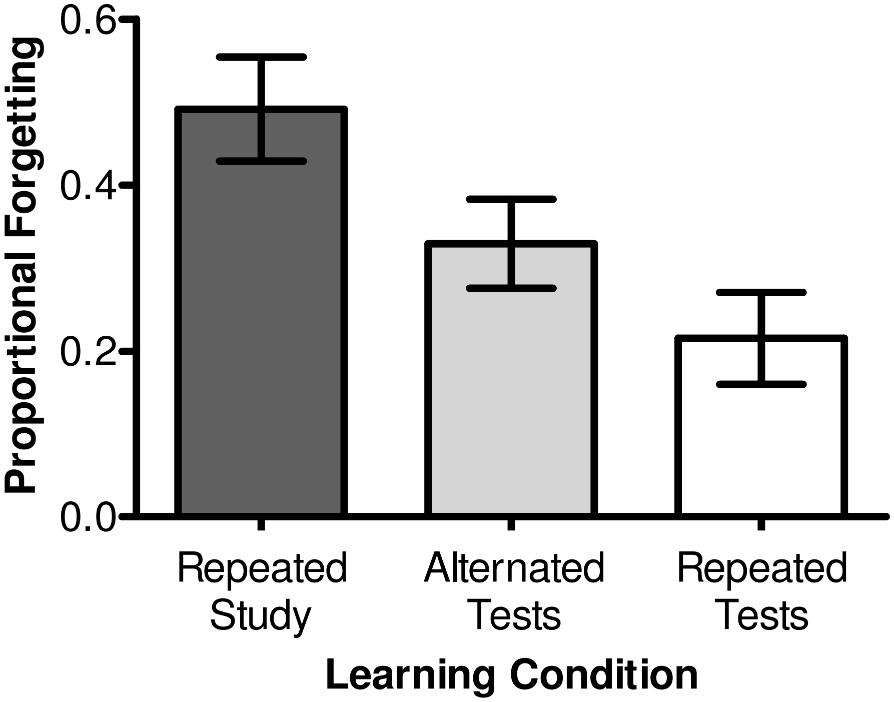

Figure 2. Proportional forgetting in Experiment 1 as function of learning condition. Error bars represent standard errors of the means.

effect (Carpenter et al., 2008). Figure 2 shows proportional forgetting in Experiment 1 as a function of learning condition. A one-way ANOVA showed a significant effect of learning condition, $F(2, 57) = 5.85$, $p < .01$, $\eta_p^2 = .17$. Follow-up contrasts showed that the test conditions showed less proportional forgetting compared to the restudy (control) condition, $t(57) = 3.11$, $p < .01$. However, the difference between the repeated tests and the alternated tests condition did not reach the level of statistical significance, $t(57) = 1.41$, $p = .16$.

Separate follow-up analyses were performed to identify which learning condition resulted in the best memory performance after the 1-min and 1-week retention interval. A one-way ANOVA for the 1-min retention interval showed a significant effect of learning condition, $F(1, 57) = 5.80$, $p < .01$, $\eta_p^2 = .17$. Post-hoc t-tests showed that both the restudy (control) and alternated tests conditions resulted in better performance than the repeated tests condition, $t(38) = 2.68$, $p < .05$ and $t(38) = 2.58$, $p < .05$, respectively. The restudy (control) and alternated tests conditions did not differ significantly from each other ($t < 1$). Lastly, a one-way ANOVA for the 1-week retention interval showed no significant effect of learning condition, $F < 1$, indicating that the initial advantage for the restudy (control) and alternated tests conditions had disappeared after a week.

Experiment 2

The absolute and the proportional forgetting measures used in Experiment 1 both suggest that the rate of forgetting was slower for the testing conditions relative to the restudy (control) condition. However, importantly, contrary to previous findings (e.g., Karpicke & Roediger, 2006; Wheeler et al., 2003) we did not observe the typical crossover interaction between learning condition and retention interval. That is, after the 1-week retention interval, there was no apparent advantage of prior testing over restudy and the three learning conditions resulted in comparable recall performance. One possible explanation for the absence of a testing benefit after the 1-week interval in Experiment 1 could be related to the retention interval. Previous research suggests that the benefit of testing can get more pronounced as the retention interval gets longer. For instance in Experiment 1 of the Wheeler et al. (2003) study, after a 2-day interval, no benefit of testing was observed relative to a restudy (control) condition. However, in Experiment 2, when the retention interval was extended and the final retention test was given one week later, a crossover interaction was observed and the testing condition outperformed the restudy (control) condition on the 1-week test. Likewise, in Experiment 1 of the present study, the benefit of testing might have become apparent if only a longer retention interval had been used. In Experiment 2, we investigated this possibility by including a retention interval of four weeks in addition to a 1-min and 1-week retention interval.

Method

Participants

Fifty-four students from the Erasmus University Rotterdam participated for course credit or a monetary reward (€ 20.00). All participants were native speakers of Dutch.

Materials, Design, and Procedure

The materials used in Experiment 2 were identical to those used in Experiment 1. However, the design was slightly different. In Experiment 2, we used a 3 × 3 mixed design with learning condition as a within-subjects factor and retention interval as a between-subjects factor. Thus, participants were tested either after a retention interval of one minute, one week, or four weeks.

As in Experiment 1, the study phase of the experiment consisted of six blocks. In each block each of the 48 word pairs was presented either for study or test. Within each of the six blocks, the three conditions were presented in a blocked order. Thus, for example, for a given subject, in block 1 first all 16 word pairs from the restudy (control) condition could be presented, followed by all 16 word pairs from the repeated tests condition and finally all 16 word pairs from the alternated tests condition. The order of these restudy (control), alternated tests and repeated tests sub-blocks within a block was determined randomly. Likewise, the word pairs within sub-blocks were presented in a random order. New random orders were generated for each sub-block, block and participant.

Participants were randomly assigned to one of the three retention interval conditions. Thus, 18 participants were tested after a 1-min interval, 18 were tested one week later, and 18 were tested after a 4-week interval. Note that all participants were asked to return and attend the 1-week and 4-week sessions regardless of the retention interval condition they had been assigned to. Those participants whose memory was not tested in a particular session participated in unrelated experiments. All other aspects of the method were identical to those of Experiment 1.

Results and Discussion

Initial Learning

Table 1 shows how test scores evolved across test cycles for both the repeated tests and the alternated tests condition. Unlike in Experiment 1, the increase in recall performance for the repeated tests condition did not reach the level of statistical significance, $F(2, 106) = 1.76$, $p = .18$, $\eta_p^2 = .03$. However, not surprisingly, in the alternated tests condition there was a significant increase across test cycles, $F(2, 106) = 232.71$, $p < .001$, $\eta_p^2 = .81$. Follow-up analysis showed that there was a significant increase between the first and the second test cycle, $F(1, 53) = 162.89$, $p < .001$, $\eta_p^2 = .75$, as well as from the second to the third cycle, $F(1, 53) = 136.71$, $p < .001$, $\eta_p^2 = .72$.

Retention

Figure 3 shows the proportion of correctly recalled target words as a function of learning condition and retention interval. A 3 × 3 mixed ANOVA showed a significant main effect of retention interval, $F(2, 51) = 37.03$, $p < .001$, $\eta_p^2 = .59$. The main effect of learning condition was not significant, $F(2, 102) = 1.08$, $p > .25$, $\eta_p^2 = .021$. More important, the interaction between learning condition and retention interval was significant, $F(4, 102) = 11.24$, $p < .001$, $\eta_p^2 = .31$, indicating that the absolute amount of forgetting was different for the three learning conditions. Follow-up contrasts showed that the test conditions showed less absolute forgetting over the 4-week interval compared to the restudy (control) condition, $F(2, 51) = 17.07$, $p < .001$, $\eta_p^2 = .40$. Also, the repeated tests condition showed less absolute forgetting than the alternated tests condition, $F(2, 51) = 6.01$, $p < .01$, $\eta_p^2 = .19$.

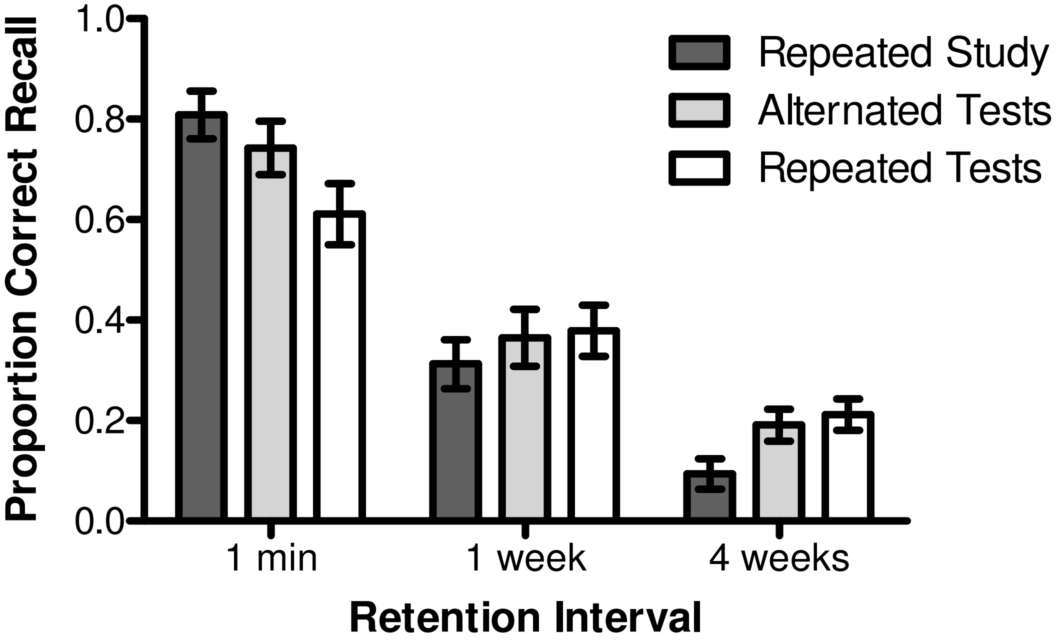

Figure 3. Proportion correctly recalled target words in Experiment 2 as function of learning condition and retention interval. Error bars represent standard errors of the means.

In Experiment 1, we also investigated proportional forgetting in addition to absolute forgetting. As noted, the course of forgetting is best described by a power function originally proposed by Wickelgren (1974): $y = a(bt + 1)^{-c}$. This power function measures forgetting as a proportional loss of the amount of originally learned information (Carpenter et al., 2008). In this function *a* represents the degree of original learning, *b* is a scaling constant, and *c* represents the rate of forgetting. Although the design of Experiment 2 did not allow us to analyze proportional forgetting, the use of more than two retention intervals in Experiment 2 enabled us to fit forgetting functions to the averaged data for the three learning conditions. First, the scaling constant was estimated by fitting the function to the data averaged over conditions (see also Wixted & Carpenter, 2007). Subsequently, we fit the data for the three learning conditions separately. Figure 4 shows the forgetting curves for each learning condition with the corresponding power functions. Of main interest are the parameter estimates for the rate of forgetting. As can be seen the difference between conditions in the respective rate of forgetting parameter estimates corresponds with the conclusions from the ANOVA-based method. The value of the parameter estimate was lowest for the repeated test condition (c = -.43), followed by the alternated condition (c = -.60), and the repeated study condition (c = -.82), respectively.

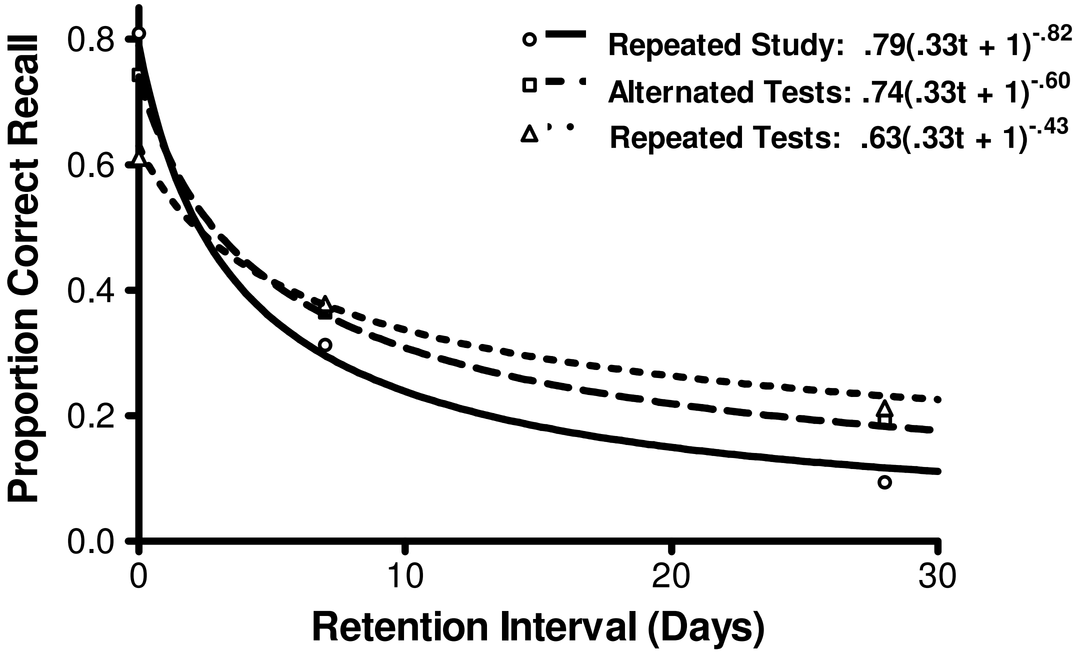

Figure 4. Proportion correctly recalled target words as a function of retention interval and learning condition in Experiment 2 with corresponding forgetting curves.

To identify for each retention interval separately which learning condition resulted in the best memory performance, we performed three additional repeated measures ANOVAs. On the 1-min test, there was a significant effect of learning condition, $F(2, 34) = 16.49$, $p < .001$, $\eta_p^2 = .49$. Paired samples t-tests showed that both the restudy (control) and alternated tests conditions resulted in better performance than the repeated tests condition, $t(17) = 5.37$, $p < .001$ and $t(17) = 4.30$, $p < .001$, respectively. The difference between the restudy (control) and alternated tests condition failed to reach statistical significance, $t(17) = 2.22$, $p = .12$. On the 1-week test, there was no significant effect of learning condition, $F(2, 34) = 1.27$, $p > .25$, $\eta_p^2 = .07$. However, on the 4-week test, the effect of learning condition was significant, $F(2, 34) = 13.02$, $p < .001$, $\eta_p^2 = .43$. Paired samples t-tests showed that both the repeated tests and alternated tests conditions resulted in better performance than the restudy (control) condition, $t(17) = 4.68$, $p < .001$, and $t(17) = 4.51$, $p < .01$, respectively. The difference between the repeated tests and alternated tests condition failed to reach statistical significance ($t < 1$).

General discussion

In two experiments, we investigated the effect of testing on long-term retention of foreign vocabulary word pairs. In Experiment 1, using both absolute and proportional forgetting measures, we found that retrieval practice during learning slowed down the rate of forgetting over a 1-week interval compared to a restudy (control) condition. However, we did not observe the often found crossover interaction (e.g., Roediger & Karpicke, 2006b; Wheeler et al., 2003). That is, after a 1-week interval, recall performance was more or less equivalent for all learning conditions. In Experiment 2, using an extended retention interval of four weeks, we replicated the finding that testing can slow down the rate of forgetting. Furthermore, we showed that, after the extended 4-week interval, the respective forgetting functions crossed over, ultimately resulting in a long-term recall benefit of retrieval practice. Taken together, the results of our study provide a clear demonstration of the powerful effect retrieval practice can have on long-term retention. Furthermore, our results indicate that the benefit of retrieval practice can get more pronounced as the retention interval gets longer.

To our knowledge, the present study is among the first to show a long-term retention benefit of testing across an interval as long as four weeks. Other studies investigating the rate of forgetting over retention intervals of a length comparable to the one used in the present study have come up with somewhat equivocal results. For instance, Carpenter et al. (2008) explored the rate of forgetting following testing with feedback across a 6-week interval. Their conclusions concerning the rate of forgetting were dependent on the approach taken to assess forgetting. That is, the more stringent ANOVA-based approach sometimes led to different conclusions than the curve fitting method. Based on the ANOVA approach, in just one out of three experiments, it was found that test trials slowed down the rate of forgetting. In the present study, we extended their findings by providing strong confirmatory support for the idea that testing can slow down the rate of forgetting even over an interval as long as four weeks.

A second merit of the present study is that we directly compared forgetting of a testing without re-presentation to a testing with re-presentation condition. As noted, one of the limitations in the testing effect literature is that few studies have made such a comparison. However, those studies that have, have come up with some interesting and surprising results. For instance, Thompson et al. (1978) found that, across a 2-day interval, forgetting was slowed down more for testing without re-presentation of unrecalled items relative to testing with re-presentation of unrecalled items and a repeated study (control) condition. In the present study, we replicated and extended this finding across a 4-week retention interval. Repeated testing resulted in a slower rate of forgetting compared to both the alternated tests condition and the restudy (control) condition. Note that this was the case even though in the alternated tests condition more items were successfully retrieved during the initial learning phase compared to the repeated test condition.

One possible explanation for the retention benefit observed for the repeated tests condition, could be due to the number of successful retrieval attempts during the initial learning phase. That is, even though fewer items were successfully retrieved in the repeated tests condition, those items that were successfully retrieved, were so multiple times during the successive test cycles. Thus, in the repeated tests condition, a small subset of items received many successful retrieval practice trials. In the alternated tests condition, however, far fewer items were successfully retrieved on multiple occasions. For instance, a large portion of retrieved items received just one successful retrieval practice trial, because many items were not successfully retrieved until the very last test cycle in the alternated sequence. Thus, in the alternated condition, a larger subset of items received considerably less successful retrieval practice trials compared to the repeated tests condition. Closer inspection of our data revealed that this was the case both in Experiments 1 and 2.[1] Prior research has established the powerful effect of repeated retrieval on long-term retention (e.g., Karpicke & Roediger, 2007; Karpicke and Roediger, 2008). For instance, in a study by Karpicke and Roediger (2007) participants learned a list of words and were subsequently tested on the materials. In one condition correctly recalled items were dropped from further study while in another condition correctly recalled items were dropped from further testing during the initial learning phase. When participants were tested one week later, the data indicated that long-term retention did not benefit from additional study of previously recalled items compared to dropping these items from further study. However, additional testing of previously recalled items had a profound positive effect on long-term retention compared to dropping these items from further testing. Furthermore, in an additional conditional analysis, Karpicke and Roediger showed that the probability of final recall for words was a function of the number of times these words were retrieved from memory during the initial learning phase. Thus final recall test performance was generally better for words that were successfully retrieved more often during the initial learning phase. However, as they also note, one should keep in mind that such correlational evidence might also reflect

[1] In Experiment 1, items in the repeated tests condition were successfully retrieved 61% (*SD* = 27%) of the time versus 52% (*SD* = 13%) retrieval success in the alternated tests condition. Additional exploratory analysis showed that the difference did not reach the level of statistical significance, $t(26.89) = 1.32$, $p = .20$, $d = 0.42$. In Experiment 2, items in the repeated tests condition were successfully retrieved 56% (*SD* = 24%) of the time versus 43% (*SD* = 17%) retrieval success in the alternated tests condition. Additional exploratory analysis revealed that this difference was significant, $t(53) = 5.20$, $p < .001$, $d = 0.74$.

an item-selection artifact. That is, easy items are more likely to be retrieved during both the initial learning phase as well as on the final test. In a similar vein, Thompson et al. (1978) have also suggested that the benefit observed for repeated testing (without feedback) might be the result of the overlearning of a small selection of easy items. Future research should be directed at investigating the role of item selection in the testing effect.

In the present study, we investigated the potential of testing as a strategy to retard the rate of forgetting. In our study, the results indicated that repeated testing without re-presentation trials was most effective for achieving this goal. However, it should be noted that, even though repeated testing resulted in a slower rate of forgetting compared to the alternated tests condition, the alternated tests condition did not ultimately result in inferior final test performance compared to the repeated tests condition. That is, in neither of our experiments did the respective forgetting functions of the test conditions cross-over. Thus, for the retention intervals used in the present study, the results suggest that the alternated tests condition might be the preferred strategy. That is, alternated testing resulted in relatively good recall performance both on the short-term and the long-term.

To conclude, in the present study, we demonstrated the powerful effect testing can have on long-term retention. Testing during initial learning more than doubled 4-week final test performance compared to the restudy (control) condition. However, it should also be noted that, even in the test conditions, the larger part of what had been initially learned was still forgotten across the 4-week interval. Thus, the observed benefit of testing was obtained for a relatively small subset of items that were still recallable at the time of the final test. For future directions, it would therefore be interesting to look at more comprehensive retention measures, like for instance rate of relearning. For one thing, it has been suggested that the relearning method is one of the most sensitive tools available for measuring retention (Macleod, 1988; Nelson, 1971, 1985). More importantly, the relearning method might also hold more practical relevance compared to other, more popular, measures of retention (e.g. recall or recognition test performance) that are often used in research on learning and retention. For educational practice, it seems vital to establish whether testing during initial learning can also facilitate the reinstatement of knowledge once acquired, but subsequently forgotten.

Chapter 6

Repeated Testing, Item Selection, and Relearning: The Benefits of Testing Outweigh the Costs[*]

[*] Chapter 6 was published as: de Jonge, M., & Tabbers, H. K. (2013). Repeated testing, item selection, and relearning: The benefits of testing outweigh the costs. *Experimental Psychology, 60,* 206-212.

Abstract

In the present study we investigated the effect of repeated testing on item selection, retention, and delayed relearning of paired associates. Participants learned both related (easy) and unrelated (difficult) word pairs under conditions of repeated study and repeated testing. A retention test was given after both a 5-minute and a 1-week interval. Following the 1-week retention test, participants received a relearning task. During the initial learning phase of the experiment, more related word pairs were successfully recalled on the practice tests compared to unrelated word pairs. Also, long-term retention benefits were found for items that were repeatedly tested compared to items that were repeatedly studied, regardless of item difficulty. The results suggest that the testing benefit following conditions of repeated testing cannot be attributed to mere item selection. Secondly, we found that delayed relearning was faster for previously restudied items compared to previously tested items. However, at the end of the relearning phase, repeated study and repeated testing one week prior to relearning resulted in comparable levels of recall performance. The results suggest that repeated testing can enhance delayed recall performance with little additional cost in terms of delayed relearning.

Rigorous restudy can be a very effective learning strategy for students when short-term retention is concerned. Unfortunately, the ravages of time creep unrelentingly. What has been learned is soon forgotten and after a relatively long retention interval very little of what was initially stored in memory can be successfully retrieved. However, research suggests that there is a solution to this problem. Numerous studies have shown that retrieving information (i.e., taking a test) during learning can greatly enhance the retention of retrieved information and slow down the rate of forgetting (for a review see Roediger & Karpicke, 2006). This general finding has received a considerable amount of attention in recent years and it has been emphasized in the literature that this so-called *testing effect* has important implications for educational purposes (McDaniel, Roediger, & McDermott, 2007).

In a typical testing effect study by Wheeler, Ewers, and Buonanno (2003), participants learned a list of words under the condition of repeated study or repeated testing. The repeated study group studied the list of words during four consecutive study cycles, while the test group studied the list just once followed by three consecutive recall tests without feedback. On a 5-min test the repeated study group outperformed the repeated test group. However, on a final test given 1 week later, the results were reversed and the repeated testing group outperformed the restudy group. The results indicate that taking tests during learning can slow down the rate of forgetting for successfully retrieved information.

Although the implications of the results from testing effect studies like the one by Wheeler et al. (2003) seem straightforward, there are some issues that deserve consideration. For instance, it has been noted that, under conditions of repeated testing without feedback, only a subset of items is effectively strengthened (e.g., Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011). The benefits of testing are largely limited to the subset of items that have been successfully retrieved on a practice test (e.g., Jang, Wixted, Pecher, Zeelenberg, & Huber, 2012). In all probability, this subset predominantly consists of items that are relatively easy to learn. In other words, it is likely that item selection occurs under conditions of repeated testing. However, the role of item selection has received little consideration in the testing effect literature.

Since it is likely that item selection occurs under conditions of repeated testing, it is important to establish whether or not repeated testing will indiscriminately improve retention for both easy and difficult items. Research on the testing effect has shown that testing can sometimes fail to improve long-term retention when recall performance on a practice test is relatively low (e.g., Kang, McDermott, & Roediger, 2007). Since recall performance for difficult items is expected to be relatively low on a practice test, it would be interesting to see whether or not repeated testing will still produce a testing effect for these items.

In terms of benefits, repeated testing might improve the retention of easy items and not so much for difficult items.

In a recent contribution to the field it has been argued that taking repeated tests without feedback will bifurcate the distribution of item strengths on a target list, whereas repeated study will not result in a bifurcated distribution (Halamish & Bjork, 2011; Kornell et al., 2011). More specifically, Kornell et al. argued that repeated testing divides item distributions into weak and strong items. On a first practice test, only a subset of items is successfully retrieved. These items are strengthened during subsequent test trials, whereas previously unrecalled items do not receive further practice and will weaken as a consequence. However, under conditions of repeated study, all items in a set are practiced continuously. As a consequence, all restudied items will get strengthened, yet to a lesser extent than the successfully retrieved items. Importantly, the bifurcation model does not make specific predictions about the role of a priori item difficulty. Rather, it is assumed that testing will enhance the memory strength of successfully recalled items, regardless of their difficulty.

Another popular account for explaining the testing effect is the *retrieval effort hypothesis* (Pyc & Rawson, 2009). This hypothesis suggests that the memorial benefits of testing are largely determined by the amount of effort invested in a retrieval attempt. Assuming that it takes more effort to retrieve difficult items from long-term memory than easy items, successful retrieval should result in a larger benefit for difficult items . In short, the retrieval effort hypothesis would predict a larger testing benefit for difficult items compared to easy items.

To sum up, it is unclear how item difficulty will affect the relative benefits of testing as different accounts of the testing effect lead to different predictions. In the present study, we investigated the role of item difficulty under conditions of repeated testing without feedback. To this end we manipulated the relative difficulty of the materials used. Research suggests that both ease of learning and long-term retention can be affected by the degree of association of word pairs (Heim, Watts, Bower, & Hawton, 1966). In the present study we used mixed word pair lists containing related and unrelated word pairs. We expected item selection to occur during learning. More specifically, we expected that more easy (related) items would be successfully retrieved on the practice tests than difficult (unrelated) items. Secondly, if item selection does play a role in the testing effect then we would expect that the relative benefit of testing on long-term retention would vary as a function of item difficulty.

A second question we addressed in the present study was whether or not repeated testing would also result in a relearning benefit following a delay. Research on the testing effect has mostly focused on single-session learning with long-term retention test performance as the crucial outcome variable. However, researchers have recently advocated other learning outcomes like relearning (Rawson & Dunlosky, 2011). It has been argued that relearning might be a more sensitive tool for measuring what has been saved in memory (Nelson, 1971, 1978, 1985). Especially when long-term retention is the subject of investigation, relearning can tell us a great deal about what resides in memory even when this information cannot be consciously retrieved (MacLeod, 1988).

Another reason why it could be interesting to look at relearning is that it is not clear whether or not repeated testing will result in faster relearning. In fact, there are reasons to expect just the opposite: a relearning advantage following repeated study. As noted earlier, in the study by Wheeler et al. (2003), the repeated study group learned more items than the testing group during initial practice. It was only after a 1-week interval that the retention benefit of testing became apparent. Given that forgetting is a decremental process rather than occurring in all-or-none fashion (Nelson, 1971), one would expect that some residual information is still left in memory for items that could not be recalled after a delay. Therefore, one could argue that the restudy group in the Wheeler et al. (2003) study might have had the advantage if an opportunity to relearn the materials had been given following the 1-week interval. This point has been raised by other researchers as well (e.g., Kornell et al., 2011). For instance, the bifurcation model would predict that previously restudied items that are forgotten over time are expected to be closer to the threshold for successful retrieval compared to previously tested items that were never successfully retrieved to begin with. Consequently, one would expect that rate of relearning after a delay should be faster for items that were learned under conditions of repeated study compared to items that were learned under conditions of repeated testing.

Thus, in the present study, we also aimed to investigate the relative benefits of repeated study and repeated testing on rate of relearning following a 1-week delay. We expected that rate of relearning would be faster for items that were learned under conditions of repeated study compared to items that were learned under conditions of repeated testing.

Method

Participants

Twenty-six undergraduate Psychology students at the Erasmus University in Rotterdam, ages 17 - 25, participated in partial fulfilment of course requirements. Eight participants were male and 18 female. Data from two participants were excluded, because these participants failed to show up for the 1-week session of the experiment.

Materials

A total of 96 word pairs were used in the experiment. The mean word frequency per million (Keuleers, Brysbaert, & New, 2010) was 31.88 (*SD* = 91.16). Forty-eight weakly related word pairs (e.g., *fakkel - grot* [*torch - cave*]) were compiled using free association norms for Dutch words (De Deyne & Storms, 2008). The related word pairs had a mean word length of 4.9 letters (*SD* = 0.9) and a mean forward strength of 0.045 (*SD* = 0.02). The other 48 word pairs used in the experiment were unrelated (e.g., *gebit – balkon* [*jaw - balcony*]). The mean word length for these word pairs was 4.9 letters (SD = 0.8). Word pairs were divided over two different mixed-item lists, each list containing 24 related and 24 unrelated word pairs. The computer application E-prime (Psychology Software Tools, Pittsburgh, PA) was used to create and run the experiment.

Design and Procedure

In the present study we used a within-subjects design. The experiment consisted of two sessions separated by a 1-week interval. The first session of the experiment consisted of an initial learning phase followed by a retention test. During the initial learning phase, participants learned two lists of word pairs under two different learning conditions (repeated study vs. repeated testing). In the repeated study condition a list of word pairs was learned during four consecutive study cycles. During study cycles word pairs were presented one at a time with a presentation rate of 5 s per pair. In the repeated testing condition participants studied a list of word pairs just once and then received three cued recall tests. During test cycles, the cue-words were presented one at a time and participants were given 5 s to type in the correct target word. No feedback was provided after giving a response. Items from each list were presented in a random order during both study and test cycles.

Following the initial learning phase of the experiment, participants first worked on multiplication problems for 5 minutes before taking a self-paced retention test on half of the word pairs from each learning condition. Items from the repeated study condition were intermixed with items from the repeated testing condition on the retention test. Participants were shown the cue words one at a time. They were asked to type in the correct target words and they were told that they could progress to the next item by pressing the ENTER-key. Upon completion of the retention test participants were dismissed. One week later participants received a final retention test on the remaining half of the word pairs. Immediately following the 1-week final retention test, participants received instructions for the delayed relearning phase of the experiment. During the relearning phase participants relearned only those items that were present on the 1-week retention test (the other half of the items was discarded). The relearning phase consisted of two alternating study and test cycles. During study cycles word pairs were presented one by one with 5 s per pair and during test cycles participants were given 5 s to type down a response.

A total of eight counterbalance conditions were used in the experiment. The two lists of word pairs were assigned to learning conditions in such a fashion that both lists appeared equally often in both conditions. Also, in the learning phase the order of learning conditions was counterbalanced. Finally, each half of a list appeared equally often on both retention tests.

Results

All data were analysed using repeated measures analysis of variance. Mauchly's test indicated that the assumption of sphericity had been violated for some of the data. In these cases, degrees of freedom were corrected using Greenhouse-Geiser estimates of sphericity.

Initial Practice Tests

Table 1 shows the mean proportion of correctly recalled target words and the mean response times on the three practice tests given during the initial learning phase of the experiment. The data from the initial practice tests were analysed using a 2 × 3 (Item difficulty × Test cycle) repeated measures ANOVA. Recall performance for related word pairs was higher compared to recall performance for the unrelated word pairs, $F(1, 23) = 34.55$, $p < .001$, $\eta_p^2 = .60$, indicating that item selection had occurred during testing. Also, there was a gradual increase in recall performance over the course of the three test cycles, $F(1.32, 30.45) = 15.68$, $p < .001$, $\eta_p^2 = .41$. The finding that recall performance can increase over the course of successive tests without corrective feedback is a well-documented phenomenon in research on learning and forgetting known as *the hypermnesic effect* (for a review see Payne, 1987). Follow-up contrasts showed that both the increase in test performance between T1 and T2, as well as the increase between T2 and T3 were significant ($F(1, 23) = 12.08$, $p < .01$, $\eta_p^2 = .34$, and $F(1, 23) = 10.14$, $p < .01$, $\eta_p^2 = .31$, respectively). The interaction between test cycle and item difficulty did not reach the level of significance, $F < 1$.

In addition to recall performance we also looked at the response times for target words that were correctly recalled on the practice tests. If the difficulty manipulation employed in the present study was successful, one would also expect faster response times for correctly recalled related word pairs compared to unrelated word pairs (MacLeod & Nelson, 1984). The response times for correctly recalled unrelated items were higher compared to the response times for the related items, $F(1, 22) = 18.66$, $p < .001$, $\eta_p^2 = .46$. Also, we found that response times tended to decrease as learning progressed, $F(1.52, 33.39) = 43.33$, $p < .001$, $\eta_p^2 = .66$. Follow-up contrasts showed that both the decrease in response time

Table 1

Mean Proportion of Correctly Recalled Target Words and Mean Response Times (in Milliseconds) on the Three Initial Practice Tests as a Function of Item Difficulty

Item difficulty and practice test	Proportion correct recall	Response time
Related		
T1	.57 (.28)	2,620 (368)
T2	.61 (.30)	2,351 (297)
T3	.62 (.30)	2,216 (354)
Unrelated		
T1	.35 (.29)	3,032 (543)
T2	.38 (.31)	2,537 (412)
T3	.41 (.31)	2,407 (401)

Note. Standard deviations are given in parentheses. T1, T2, and T3 refer to the first, second, and third practice tests, respectively.

between the first and the second test, as well as the decrease between the second and the third test were significant ($F(1, 23) = 41.40$, $p < .001$, $\eta_p^2 = .65$, and $F(1, 23) = 11.16$, $p < .01$, $\eta_p^2 = .34$, respectively). The interaction between test cycle and item difficulty did not reach the level of statistical significance $F(1.61, 35.33) = 3.35$, $p > .05$.

In sum, the response time data and the recall data provide converging evidence that the difficulty manipulation had been successful. Unrelated (difficult) items were less likely to be recalled on the practice tests compared to related (easy) items. Moreover, the average response time for successfully recalled target words was higher for the unrelated items than for the related items.

Retention

Figure 1 shows the proportion of correctly recalled target words on the 5-min and the 1-week retention tests as a function of learning condition and item difficulty. The data were analysed using a 2 × 2 × 2 (Item difficulty × Learning condition × Retention interval) repeated measures ANOVA. Like in the initial learning phase of the experiment, participants correctly recalled more related items than unrelated items, $F(1, 23) = 95.31$, $p < .001$, $\eta_p^2 = .81$. There also was a significant main effect of retention interval, $F(1, 23) = 141.97$, $p < .001$, $\eta_p^2 = .86$, indicating

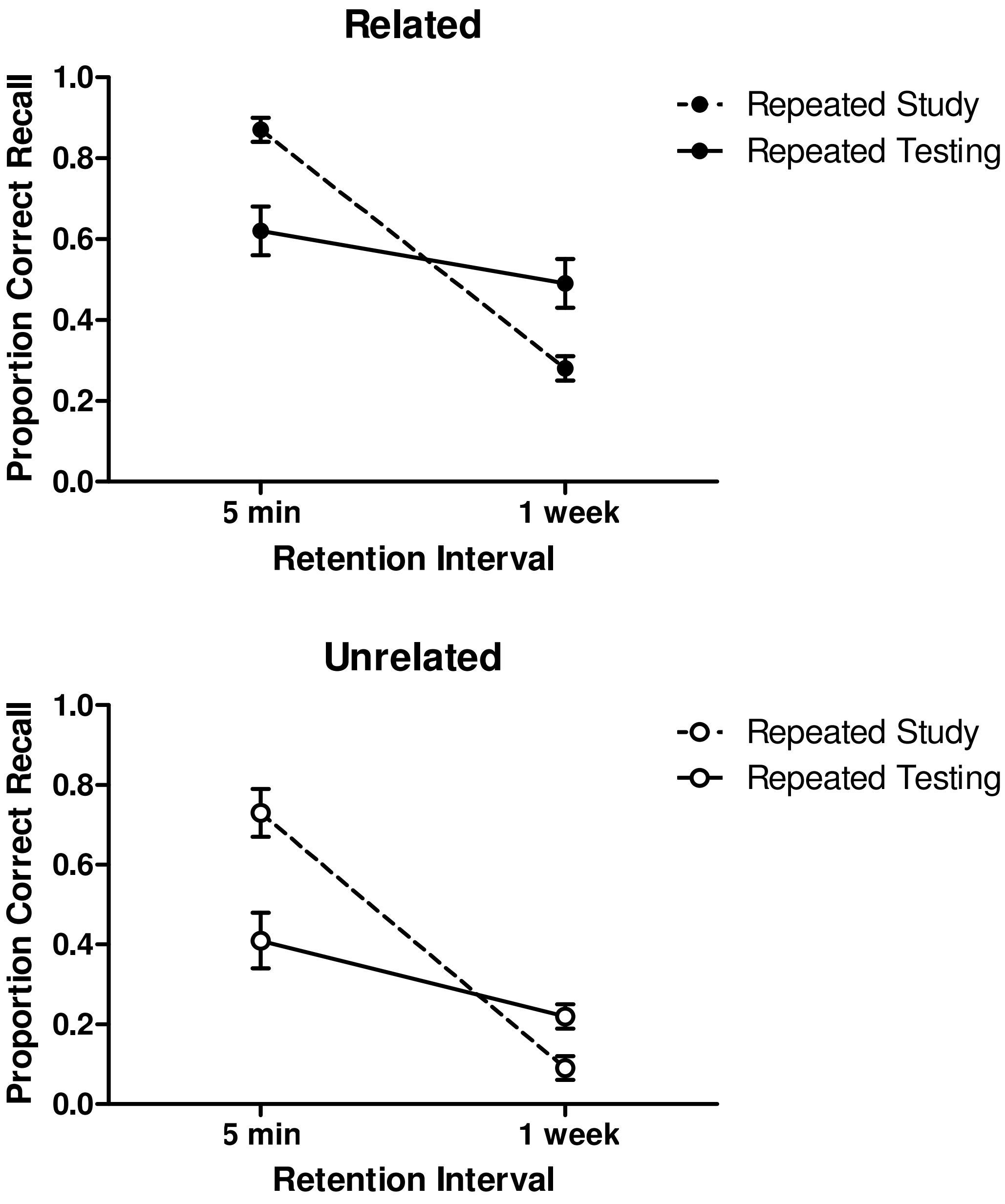

Figure 1. Proportion of correctly recalled target words on the 5-min and 1-week cued recall test as a function of item difficulty (related vs. unrelated) and learning condition (repeated study vs. repeated testing). Error bars represent standard errors of the means.

that forgetting occurred during the 1-week interval. Lastly, the main effect of learning condition was significant, $F(1, 23) = 4.66$, $p < .05$, $\eta_p^2 = .17$, indicating that, overall, recall performance was higher following repeated study compared to repeated testing. Importantly, the main effects of learning condition and retention interval were qualified by a significant interaction, indicating that there was a difference in rate of forgetting between the repeated study and the repeated testing condition, $F(1, 23) = 125.90$, $p < .001$, $\eta_p^2 = .85$. This interaction was unaffected by item difficulty, as the three-way interaction between learning condition, item difficulty, and retention interval did not reach the level of significance, $F < 1$.

Thus, repeated testing slowed down the rate of forgetting of both related and unrelated items, $F(1, 23) = 92.71$, $p < .001$, $\eta_p^2 = .80$, and $F(1, 23) = 67.81$, $p < .001$, $\eta_p^2 = .75$, respectively. For the related word pairs on the 5-min retention test, recall performance was higher in the repeated study condition (*M* = 87%, *SD* = 15%) compared to the repeated testing condition (*M* = 62%, *SD* = 28%), $t(23) = 5.36$, $p < .001$, $d = 1.28$. However, one week later, the repeated testing condition (*M* = 49%, *SD* = 29%) outperformed the repeated study condition (*M* = 28%, *SD* = 16%), $t(23) = 4.52$, $p < .001$, $d = 1.08$. The same patterns of results were obtained for the unrelated word pairs. On the 5-min retention test, recall performance was higher in the repeated study condition (*M* = 73%, *SD* = 29%) compared to the repeated testing condition (*M* = 41%, *SD* = 35%), $t(23) = 6.41$, $p < .001$, $d = 1.34$. On the 1-week retention test, the repeated testing condition (*M* = 22%, *SD* = 16%) outperformed the repeated study condition (*M* = 9%, *SD* = 13%), $t(23) = 4.15$, $p < .001$, $d = 0.86$.

Delayed Relearning

Figure 2 shows the proportion of correctly recalled target words on the tests given during the delayed relearning phase of the experiment as a function of learning condition and item difficulty. Note that T1 represents recall on the 1-week retention test. The relearning data were analysed using a 3 × 2 × 2 (Test cycle × Item difficulty × Learning condition) repeated measures ANOVA. There was a significant main effect of test cycle, $F(2, 46) = 306.62$, $p < .001$, $\eta_p^2 = .93$, simply indicating that recall increased over the course of the three tests. Follow-up contrasts showed that there was a significant increase in recall score from T1 to T2, $F(1, 23) = 250.45$, $p < .001$, $\eta_p^2 = .92$, as well as a significant increase from T2 to T3, $F(1, 23) = 39.11$, $p < .001$, $\eta_p^2 = .63$. Also, there was a significant main effect of item difficulty, $F(1, 23) = 147.84$, $p < .001$, $\eta_p^2 = .87$ indicating that more easy (related) items were successfully recalled compared to difficult (unrelated) items. Lastly, there was a significant main effect of learning condition, $F(1, 23) = 4.89$, $p < .05$, $\eta_p^2 = .18$. Overall, taking repeated tests one week prior to the relearning session resulted in better recall compared to the repeated study condition. However, as can be seen in Figure 2, this effect could be attributed entirely to the difference on the 1-week retention test given prior to relearning (T1). After just a single restudy cycle, the initial benefit of testing had disappeared and both learning conditions resulted in comparable recall performance for the remainder of the tests (T2 and T3) given during the 1-week session. This general observation was confirmed by a significant Learning condition × Test cycle interaction, $F(2, 46) = 21.62$, $p < .001$, $\eta_p^2 = .49$. Follow-up

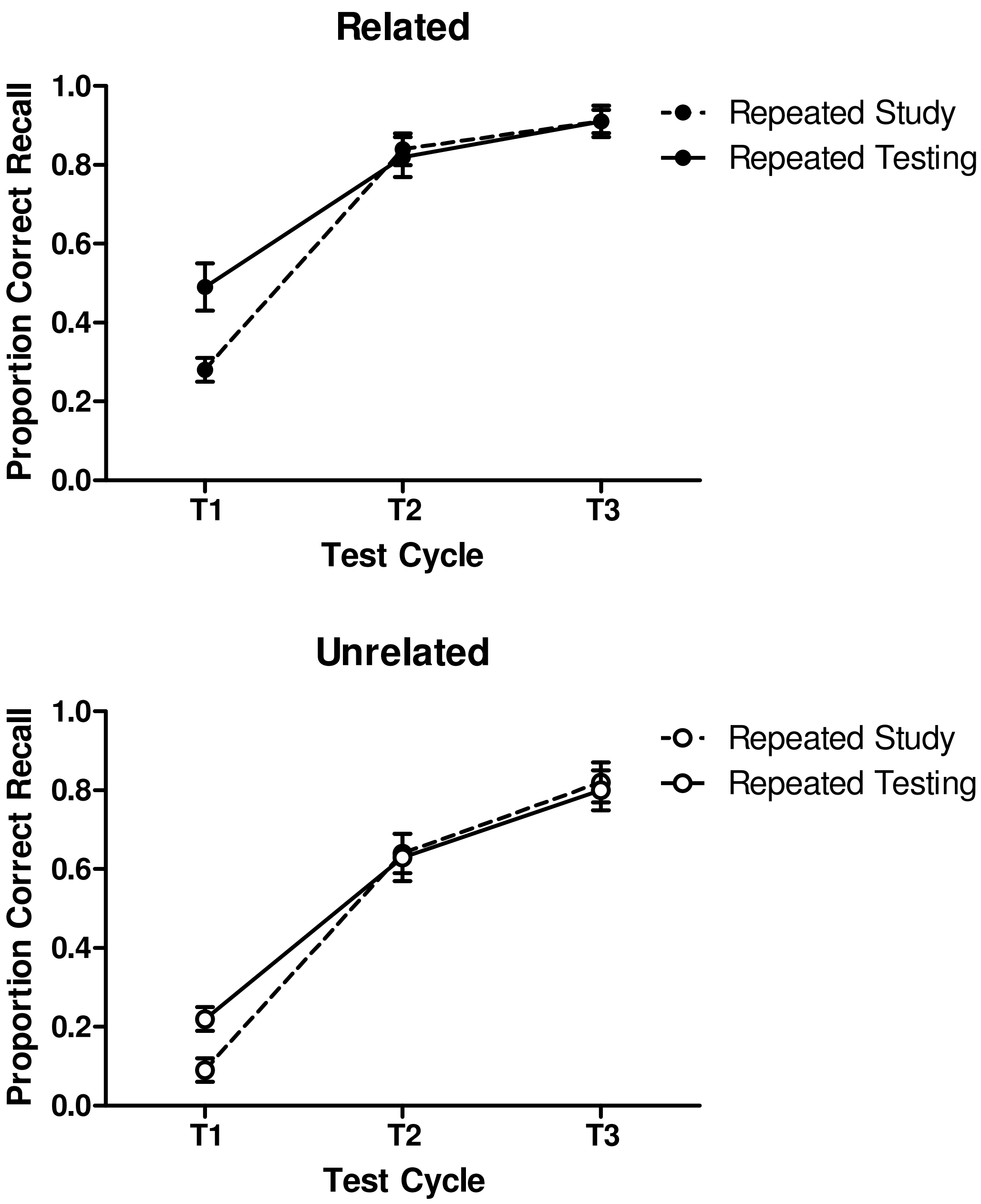

Figure 2. Proportion of correctly recalled target words on the three consecutive tests given during the 1-week relearning session as a function of item difficulty (related vs. unrelated) and learning condition (repeated study vs. repeated testing). Error bars represent standard errors of the means. Recall performance for related items in the repeated study condition was 28% (*SD* = 16%), 84% (*SD* = 18%), and 91% (*SD* = 13%) on T1, T2, and T3, respectively. For related items in the repeated testing condition, recall performance was 49% (*SD* = 29%), 82% (*SD* = 26%), and 91% (*SD* = 18%), respectively. For unrelated items in the repeated study condition, recall performance was 9% (*SD* = 13%), 64% (*SD* = 26%), and 82 % (*SD* = 24%), respectively. For unrelated items in the repeated test condition, recall performance was 22% (*SD* = 16%), 64% (*SD* = 29%), and 80% (*SD* = 23%), respectively.

contrasts showed that there was a significant Learning condition × Test cycle interaction for the first half of the relearning session (from T1 to T2), $F(1, 23) = 35.71$, $p < .001$, $\eta_p^2 = .61$, but not for the second half (from T2 to T3), $F < 1$.

Discussion

The first goal of the present study was to investigate the possible contribution of item selection to the testing effect. This issue has been neglected in research on the testing effect. However, especially under conditions of repeated testing, item selection could have an important contribution to the benefits of testing found after a delay. If item selection does indeed play a role in the testing effect than this would have implications both for theory and practice. As expected item selection occurred on the practice tests during the initial learning phase of the experiment. Related items were more likely to be successfully retrieved on the practice tests than unrelated items. Importantly, however, we did not find evidence for the idea that item difficulty would differentially affect the testing benefit found after a delay. That is, taking repeated tests during the initial learning phase of the experiment enhanced long-term retention of both difficult and easy items.

The focus of the present study was on the situation where participants are repeatedly tested with no intervening feedback. In a study by Karpicke (2009), in which item difficulty was also manipulated, participants learned easy and difficult foreign vocabulary word pairs in alternating study and test trials. This procedure generally results in high levels of recall performance compared to repeated testing without intervening study opportunities. For instance, in Experiment 1 of the Karpicke study, cumulative recall performance during initial learning was near 100% correct regardless of item difficulty. Karpicke found that testing enhanced retention for both difficult and easy items one week after initial learning. The present study extends these findings, by showing that this is true even when only a subset of items is retrieved repeatedly. Taken together, the results from both studies suggest that it is unlikely that the relative benefits of testing for long-term retention are affected by item difficulty. This finding does not seem to be in line with the retrieval effort hypothesis (e.g., Pyc & Rawson, 2009). Our results fit better with the bifurcation account of the testing effect (Halamish & Bjork, 2011; Kornell et al., 2011). Repeated testing without feedback can enhance long-term retention, regardless of item difficulty. However, since we looked at a restricted range of item difficulty levels in the present experiment, it should be noted that caution is warranted when generalizing our results. For future purposes it could be interesting to vary item difficulty at more than two levels and perhaps look at more extreme levels of item difficulty.

The second goal of the present study was to look at the relative benefits of repeated study and repeated testing on rate of relearning after a 1-week delay. In the first part of the relearning phase, we found that rate of relearning was faster for the repeated study condition compared to the repeated testing condition. This result is in line with the idea that some residual information is left in memory even when items cannot be consciously retrieved on a delayed recall test (Nelson, 1971). Also, our results support the assumption that target items that cannot be successfully retrieved on a delayed test are more likely to be close to threshold when these items have been repeatedly studied, than when these items have been tested, but not retrieved, during initial learning (Kornell et al., 2011). Importantly, however, the faster rate of relearning in the repeated study condition did not ultimately result in superior recall compared to the testing condition. Thus, although the disadvantage of repeated study on delayed recall is compensated for by faster delayed relearning, this does not make repeated study the more effective learning strategy.

In the present study, delayed relearning was assessed following conditions in which only a subset of items was initially learned. For future research, it could be interesting to investigate the effect of testing on delayed relearning after reaching relatively high criterion levels during initial learning. Also, it could be interesting to look at delayed relearning after very long retention intervals. Most research on learning and forgetting has focused on recall performance following relatively short retention intervals of days or weeks. However, in a real-life setting, educators generally aim to increase retention of information for periods far beyond these kinds of intervals. Clearly, it is unrealistic to expect students to achieve perfect recall years after initial learning has taken place. Consequently, one could argue that for educational purposes, rate of relearning following a delay is at least as important a factor to consider as recall performance. Accordingly, it has been noted that, when very long-term retention is concerned, the least educators can hope for is that forgotten information can be relearned relatively quickly (Nelson, 1971; Rawson & Dunlosky, 2011).

To conclude, the present study shows that repeated testing can improve retention of successfully retrieved items, regardless of item difficulty. For both easy and difficult items, we found a robust effect of testing on delayed recall performance. Furthermore, when provided with the opportunity to relearn following a delayed test, final relearning outcome did not appear to be determined by prior restudy or testing. Although rate of relearning was faster for previously restudied items at the beginning of the delayed relearning phase, at the end of the relearning phase both the repeated study and the repeated testing condition resulted in comparable recall performance. To put it more boldly, the results suggest that the enhanced retrievability of repeatedly tested items comes with little additional cost in terms of delayed relearning.

Chapter 7

Retention Beyond the Threshold: Test-Enhanced Relearning of Forgotten Information[*]

[*] Chapter 7 was published as: de Jonge, M., Tabbers, H. K., & Rikers, R. M. J. P. (2014). Retention beyond the threshold: Test-Enhanced relearning of forgotten information. *Journal of Cognitive Psychology*, *26*, 58-64.

Abstract

The effect of repeated testing on delayed relearning of paired associates was investigated. Participants learned two lists of Lithuanian-Dutch word pairs until reaching the criterion of one correct recall from long-term memory. In one condition, items subsequently received three post-retrieval study trials and in the other condition items received three post-retrieval test trials. Participants returned one week later for delayed recall and relearning. Post-retrieval test trials resulted in better delayed recall performance than post-retrieval study trials. Moreover, we found that items that were repeatedly studied or tested 1 week prior to relearning were relearned faster than a new set of similar (not previously presented) items. Most importantly, items were relearned faster when they had previously been learned under conditions of post-retrieval testing than items learned under conditions of post-retrieval study. Taken together, the results indicate that the benefits of repeated testing are not just limited to conscious recall on a delayed test. Repeated testing during initial learning is also a very effective strategy to enhance delayed relearning.

Numerous studies have shown that taking tests during learning can enhance delayed recall of successfully retrieved information and it has been emphasised in the literature that this finding could have important implications for educational practice (see Roediger & Karpicke, 2006a for a review of the literature). However, in a recent contribution to the field, two major limitations in testing effect literature have been identified (Rawson & Dunlosky, 2011). First of all, most testing effect studies have investigated the effect of testing following fixed amounts of retrieval practice instead of looking at learning to criterion. Secondly, studies have almost exclusively focused on the effect of testing during single session learning with delayed recall performance as the sole outcome variable instead of looking at multisession learning. In the present study we investigated the testing effect under conditions of learning to criterion. Most importantly, in addition to the effect of testing on delayed recall performance, we also looked at the effect of testing on delayed relearning.

The relearning method has a long history in research on learning and retention dating all the way back to the classic work by Hermann Ebbinghaus (1885). It has been emphasized in the literature that the relearning method is one of the most sensitive tools available for measuring retention (Macleod, 1988; Nelson, 1971, 1985). Surprisingly, however, the relearning method has been underutilized in memory research compared to more commonly used measurement tools like recall tests (MacLeod, 1988), even though these have severe limitations. For instance, a recall test is useful only for discriminating between recalled and unrecalled information. However, the fact that a certain item cannot be successfully recalled on a test does not mean that this particular item is no longer present in memory (Nelson, 1985). The memory strength for a seemingly forgotten item has just weakened to such an extent that it has temporarily become unrecallable. In that case, the item is said to be beneath the threshold for successful recall (Bahrick, 1967; Kornell, Bjork, & Garcia, 2011; Nelson, 1971). With some additional stimulation (e.g., re-exposure to the materials), forgotten items can be brought back to conscious memory with apparent ease. Indeed, research indicates that delayed relearning of forgotten items can be considerably faster compared to original learning (Bahrick, 1967; Ebbinghaus, 1885/1964) or to the learning of a new set of similar items (Macleod, 1988; Nelson, 1971, 1978).

Studies investigating the testing effect have almost exclusively focused on differences in retention between conditions for items that are above the threshold for successful retrieval. However, if retrieval practice results in the strengthening of memory traces for successfully recalled items, then one would expect that these benefits would extend beyond the recall threshold. Specifically, one would expect that forgotten items are still closer to the threshold for successful recall when these items were originally learned under conditions of repeated testing compared to items that were learned under conditions of repeated study. If this is the case, then prior testing should also facilitate relearning of forgotten items in the same way it facilitates delayed recall. On the other hand, there is also reason to question whether repeated testing will result in a delayed relearning benefit for forgotten items. For instance, the transfer appropriate processing account of the testing effect suggests that the often observed retention benefit for tested items might be the result of a greater degree of overlap between the processing engaged in during practice and the processing required by a delayed task. If retrieval is the process required by a delayed task (as is the case with a delayed retention test) then the transfer appropriate processing account predicts that practicing retrieval will aid later test performance most. However, when retrieval fails on a delayed test and additional study trials are necessary to bring items back to a state where they can be successfully retrieved, one might expect that more prior practice with study trials might aid subsequent relearning.

To sum up, it is still unclear whether repeated testing can also benefit delayed relearning of forgotten information. It has often been argued that students forget most of what they have learned relatively quickly after learning has taken place (e.g., Bahrick, 1979). Hence, when long-term retention is concerned, educators' main concern should be whether or not forgotten information can be relearned relatively quickly (Nelson, 1971; Rawson & Dunlosky, 2011). If repeated testing does not facilitate delayed relearning of forgotten information then the benefits for educational practice will also be limited. In the present study, we investigated the effect of repeated testing on delayed relearning using an initial encoding procedure comparable to the one used by Karpicke and Smith (2012). Items were first learned to a criterion of one correct retrieval from long-term memory. Subsequently, successfully retrieved items received either additional study or additional test trials. We expected repeated test trials to enhance delayed recall compared to repeated study trials (a classic testing effect). In addition, we expected to find faster relearning of items that were learned under conditions of repeated testing compared to items that were learned under conditions of repeated study.

Method

Participants

Nineteen undergraduate Psychology students at the Erasmus University in Rotterdam participated in partial fulfilment of course requirements. Data from one participant were excluded, because of experimenter error.

Materials

A total of 48 Lithuanian-Dutch word pairs (e.g., *pienas - melk* [*milk*]) were used in the experiment. The Lithuanian words were selected from a normative study on Lithuanian-English paired associates (Grimaldi, Pyc, & Rawson, 2010) and translated into Dutch for the purpose of the present experiment. The 48 word pairs were divided over three 16-item lists. The computer application E-prime (Psychology Software Tools, Pittsburgh, PA) was used to create and run the experiment.

Design and Procedure

We used a full within-subjects design in the present study. The experiment consisted of two sessions separated by a 1-week interval. The first session of the experiment consisted of an initial learning phase and the second session consisted of a relearning phase. During the initial learning phase, participants learned two lists of word pairs under two different learning conditions, post-retrieval study (PRS), and post-retrieval testing (PRT), respectively. A third list of word pairs was not used in the initial learning phase. The items on this list were presented as new items during the relearning phase of the experiment. A total of six counterbalance conditions were used to control for the effect of condition sequence as well as for the assignment of stimulus materials to conditions.

Initial Learning Phase

In both initial learning conditions, items were by default learned during alternating study and test blocks. Both conditions started out with an initial study block and a subsequent test block. During study blocks, word pairs were presented one at a time with a presentation rate of 5 s per pair. During test blocks, participants were shown only the Lithuanian word of a pair and they were given 5 s to type down the correct Dutch translation. No feedback was provided after giving a response. After the first study and test block, the procedure for the two learning conditions diverged. Individual items on a list would subsequently receive additional study and/or test trials as a function of both learning condition and previous test performance.

In the PRS condition, the procedure for items that had not been successfully retrieved during the previous test block remained the same. These items continued to be learned under the default conditions of alternating study and test trials. However, for those items that had been successfully retrieved, the learning procedure changed. Successfully retrieved items subsequently received three study trials during the next three learning blocks. After an individual item had received three post-retrieval study trials it was dropped from the list. This way, in the PRS condition, all items on a list were learned until every item had been successfully recalled once and had subsequently received three post-retrieval study trials.

The procedure for items in the PRT condition was similar to the procedure in the PRS condition. Items were by default learned under conditions of alternating study and test trials. Once a target item had been successfully retrieved, the learning procedure for this particular item was changed. However, in the PRT condition successfully retrieved items received three post-retrieval test trials (as opposed to study trials) and were then dropped from the list.

Items on a list were presented in a random order during learning blocks throughout the experiment. Also, to minimize the influence of short-term memory, learning blocks were separated by 1-min intervals. During these 1-minute intervals participants were asked to solve multiplication problems. Upon completion of the initial learning phase, participants were excused and asked to return for the second part of the experiment 1 week later.

Relearning Phase

One important difference with the initial learning phase of the experiment was that we used a mixed list in the relearning phase. This was done because we anticipated better recall performance on the 1-week test for the PRT condition compared to the PRS condition. As a consequence, we expected shorter to-be-relearned lists for the PRT condition. To control for possible effects of list-length, items from all conditions were intermixed. Also, in addition to items from the PRS and the PRT condition, the mixed list in the relearning phase also contained 16 new items to serve as a baseline control.

The relearning phase started out with a test block in order to identify those items that were below threshold for successful recall. Any item that was successfully recalled on the test was dropped from further study and testing. Consequently, only those items that had not been successfully recalled on the test recurred for further study and testing in the remainder of the relearning phase. For these items, participants received alternating study and test blocks until each item had reached the criterion of one successful retrieval from long-term memory. As soon as a target item had been successfully recalled, it was dropped from further study and testing. Like in the initial learning phase, participants worked on multiplication problems during 1-minute intervals in between learning blocks. The relearning phase ended as soon as all items had been successfully recalled.

Results

All data were analysed using repeated measures analysis of variance with one-tailed planned contrasts as follow-up. Mauchly's test indicated that the assumption of sphericity had been violated for some of the data. In these cases, degrees of freedom were corrected using Greenhouse-Geiser estimates of sphericity. Following Bakeman's (2005) recommendations for repeated measures analysis of variance, we used generalized eta squared (η^2_G) as a measure of the effect size for the results from the omnibus tests. For the results from the follow-up contrasts, Cohen's *d* is reported.

Initial Learning Phase

In the present study, the dependent variable of main interest was the number of trials to reach criterion. However, most previous research on the testing effect has focused on recall performance as a measure of both learning and retention. To allow for a more direct comparison between the present study and previous research, we will also report cumulative recall performance during the initial and delayed relearning phase of the experiment. Note that, for practical considerations (e.g. number of valid cases per participant), we only looked at cumulative recall on the first three test cycles. Figure 1 shows the cumulative proportion of correctly recalled items in the first three test cycles during the initial learning phase of the experiment. As can be seen in Figure 1, there was very little difference in recall performance across the two learning conditions. A 2 (learning condition) × 3 (test block) repeated measures ANOVA indicated that performance increased as learning progressed, $F(2, 22) = 269.59$, $p < .001$, $\eta^2_G = .71$. However, the main effect of learning condition and the Learning condition × Test cycle interaction did not reach the level of significance, both *Fs* < 1.

The average number of study trials to reach criterion per item in the initial learning phase was 1.94 ($SD = .38$) in the PRS condition, and 1.89 ($SD = .47$) in the PRT condition. Analysis revealed that the difference between the two conditions did not reach the level of significance, $t(17) = 0.42$, $p = .68$. In short, the trials to criterion data as well as the cumulative recall data indicate that the PRS and the PRT condition resulted in comparable rates of learning during the initial learning phase.

Relearning Phase

Figure 2 shows the cumulative proportion of correctly recalled items during the first three tests in the 1-week relearning phase of the experiment. As can be seen, overall cumulative recall was highest in the PRT condition, followed by the PRS and the NEW condition. The data were analysed using a 3 (learning condition) × 3 (test block) repeated measures ANOVA. There was a main effect of

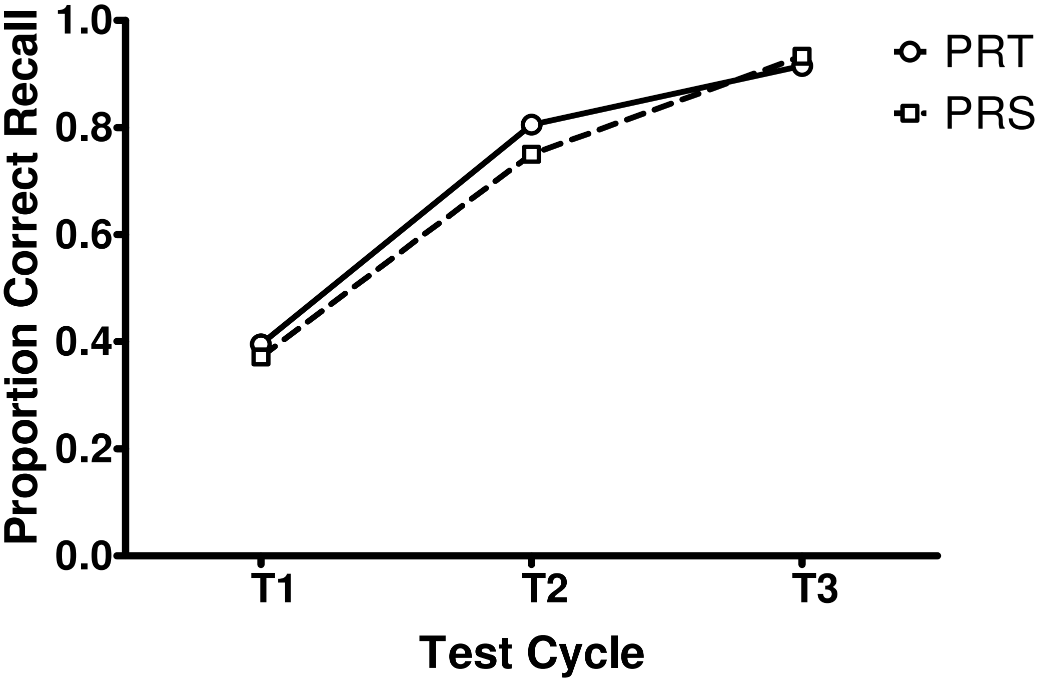

Figure 1. Cumulative proportion of correctly recalled target words on the first three test blocks in the initial learning phase as a function of learning condition.

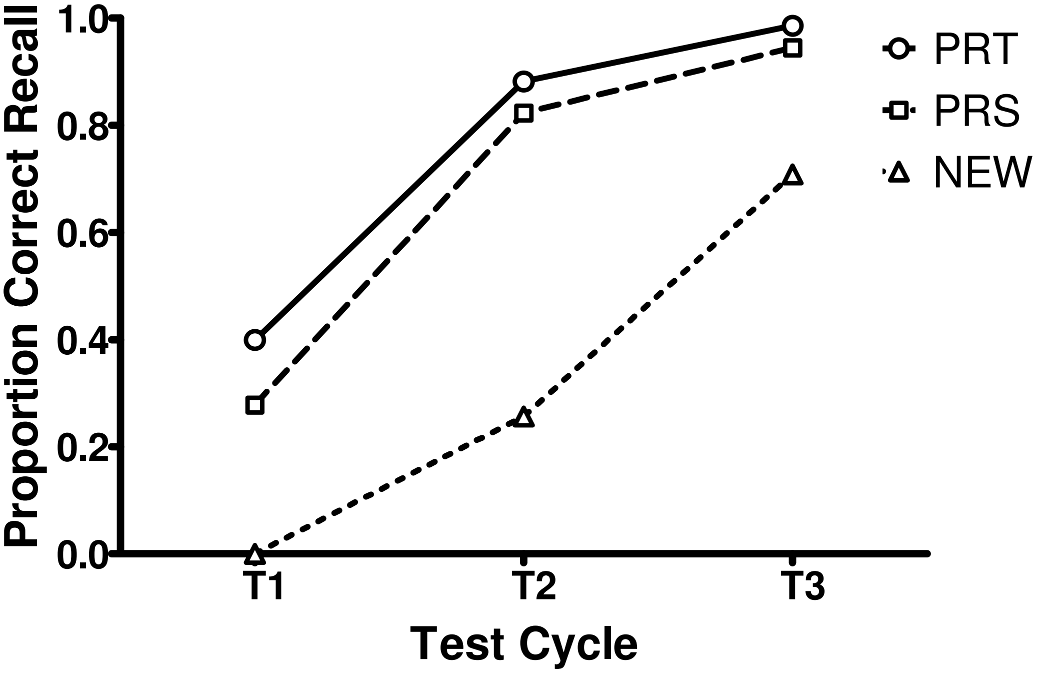

Figure 2. Cumulative proportion of correctly recalled target words on the first three test blocks in the 1-week relearning phase as a function of learning condition.

test cycle, $F(2, 34) = 611.48$, $p < .001$, $\eta^2_G = .77$, indicating that cumulative recall performance increased as learning progressed. Follow-up analysis showed that there was a significant increase from the first to the second test cycle, $t(17) = 24.11$, $p < .001$, $d = 5.68$, as well as from the second to the third test cycle, $t(17) = 13.25$, $p < .001$, $d = 3.12$. Secondly, there was a significant main effect of learning condition, $F(2, 34) = 204.23$, $p < .001$, $\eta^2_G = .62$. Follow-up analysis showed that cumulative recall in the relearning phase was higher in the PRT and the PRS condition compared to cumulative recall in the NEW condition , $t(17) = 18.69$, $p < .001$, $d = 4.41$. Also, cumulative recall was higher for the PRT condition compared to the PRS condition, $t(17) = 3.47$, $p < .005$, $d = .82$, replicating the general finding that repeated testing can enhance delayed recall performance (e.g., Karpicke & Roediger, 2007, 2008). Cumulative recall was higher in the PRT condition than in the PRS condition throughout the relearning phase, $t(17) = 2.65$, $p < .05$, $d = 0.63$, $t(17) = 1.93$, $p < .05$, $d = 0.48$, and $t(17) = 2.61$, $p < .05$, $d = 1.10$, for T1, T2, and T3 respectively.

The variable of main interest in the present study was the number of study trials it took to reach criterion. Overall, the mean number of study trials to criterion in the relearning phase of the experiment was 2.15 (*SD* = .52) for new items, 0.96 (*SD* = .40) for PRS items, and 0.74 (*SD* = .34) for PRT items. A repeated measures ANOVA revealed that the difference between conditions was significant, $F(2, 34) = 165.62$, $p < .001$, $\eta^2_G = .72$. Planned contrasts revealed that it took more trials to reach criterion for new items compared to PRS and PRT items, $t(17) = 15.12$, $p < .001$, $d = 3.84$. Most importantly, it took significantly more trials to reach criterion in the PRS condition compared to the PRT condition, $t(17) = 3.47$, $p < .005$, $d = 1.05$. Thus, the cumulative recall data and the trials to criterion data provide converging evidence indicating that the initial benefit in the PRT condition persisted throughout the relearning phase. That is, items that were repeatedly tested during initial learning, were relearned faster (i.e., in fewer trials) compared to items that were repeatedly studied.

Note that, in the analysis above, we took into account both the above-threshold (recalled) and below-threshold (forgotten) items. Another question is whether forgotten items, which could not be consciously retrieved on a recall test, would still benefit from having been tested repeatedly during initial learning. The average number of study trials to reach criterion in the relearning phase of the experiment was 2.15 (*SD* = .52) for new items, 1.29 (*SD* = .28) for PRS items, and 1.20 (*SD* = .22) for PRT items. There was a significant effect of learning condition, $F(1.29, 21.93) = 80.78$, $p < .001$, $\eta^2_G = .73$, indicating that there was a difference in the average number of study trials it took to reach criterion between the different learning conditions. Planned contrasts revealed that it took more trials to reach criterion for new items compared to forgotten items that had been previously studied or tested, $t(17) = 9.69$, $p < .001$, $d = 3.11$. Also, there was a relearning benefit for forgotten items that were learned under conditions of post-retrieval testing compared to forgotten items that were learned under conditions of post-retrieval study, $t(17) = 2.10$, $p < .05$, $d = 0.52$.

Discussion

In the present study we investigated the effect of repeated testing after learning to criterion on long-term retention of paired associates. We replicated the general finding that repeated testing enhances delayed recall performance. Moreover, we extended this finding by showing that the benefit of repeated testing persists throughout a delayed relearning phase. This finding is in contrast to findings from previous studies that did not use a procedure of learning to criterion during initial learning. For instance, de Jonge and Tabbers (2013) found that, when only a subset of items was repeatedly retrieved during initial learning, the benefit of repeated testing was limited to the first test given in a delayed relearning phase. After a single study block both the repeated testing and the repeated study condition showed comparable test performance and remained at comparable levels throughout the remainder of the relearning phase. The present study shows that, when all items in a set have been repeatedly retrieved during initial learning, the benefit of testing extends to delayed relearning. That is, a set of repeatedly tested items requires fewer trials to relearn compared to a set of items that has been repeatedly studied during initial learning.

For both forgotten items that were learned under conditions of post-retrieval testing as well as for forgotten items that were learned under conditions of post-retrieval study we observed a substantial amount of savings. In both conditions it took considerably fewer study trials to reach the criterion of one successful retrieval from long-term memory compared to the number of trials required to learn a list of new items. This result replicates earlier findings and is in line with the idea that forgetting is a decremental process rather than occurring in an all-or-none fashion and indicates that some residual information is still left in memory even when retrieval fails (Nelson, 1971). Interestingly, we also found that it took fewer trials to relearn forgotten items that had received post-retrieval test trials during initial learning compared to forgotten items that had been learned under conditions of post-retrieval study. These results suggest that the benefits of repeated testing are not limited to enhanced retrievability of a small subset of items on a later test. Repeated testing during initial learning seems to benefit memory strength in a more general way and even results in faster relearning of forgotten (unrecalled) items. Importantly, however, it should be noted that this relearning benefit was relatively small compared to some of the other effects reported in this manuscript. Furthermore, closer inspection of our data reveals that most of the forgotten items from the repeated testing condition and repeated study condition were brought back to an above-threshold state after just one restudy cycle during relearning. Thus, the observed relearning benefit for forgotten items in the repeated testing condition was based on a relatively small number of items, and some caution is warranted when interpreting these results.

Note, on the other hand, that the observed relearning advantage for forgotten items from the repeated testing condition in the present study is likely an underestimation. First of all, during the learning phase, all items in the repeated study condition also received test trials and were successfully recalled once. This was a necessary feature of the experimental design to make sure that items in the restudy condition were indeed stored into long-term memory. One would expect that the use of test trials in the restudy condition also improved item retention and as a consequence probably resulted in an overestimation of recall performance and rate of relearning relative to a situation with pure study trials. Secondly, one should also take into account the possible influence of item selection that might have affected list composition in the relearning phase of the present study. In the relearning phase, we mixed items from all conditions in one list (to control for possible list-length effects). Importantly, however, it has been noted in the literature that dropping items from a list as a function of prior test performance can result in lists of differential item difficulty (e.g., Underwood, Rehula, Keppel, 1962). That is, items that are relatively easy to recall are the first ones to be dropped from the list, leaving the more difficult items to make up the remaining to-be-relearned list. Since delayed recall performance on the first test was higher in the testing condition, one would expect that the remaining subset of to-be-relearned word pairs consisted mostly out of the more difficult items. Hence, if item selection did play a role in the present study, then this would have resulted in a disadvantage for the testing condition. Still, we found that relearning was faster for forgotten items in the repeated testing condition, compared to new items or forgotten items that were learned under conditions of repeated study. Future research should be directed at further investigating delayed relearning of forgotten information following post-retrieval test trials.

One way to interpret the findings from the present study is in terms of the new theory of disuse forwarded by Bjork and Bjork (1992). This framework distinguishes between two different types of memory strength, *storage strength*, and *retrieval strength*, respectively. The storage strength of a particular item in memory does not determine the retrievability of the item. Rather, the storage strength of an item simply relates to how well the item has been stored. Furthermore, it is assumed that storage strength, once it has been established, does not change over time. Retrieval strength, on the other hand, is assumed to deteriorate over time. If the retrieval strength for a particular item is high, this simply means that the item is recallable at the time (above threshold for successful retrieval). However, past results do not guarantee future performance. If the storage strength for an item is low then the retrieval strength for the item will quickly deteriorate and drop beneath the threshold for successful retrieval. One possible explanation for why repeated testing appears to slow down the rate of forgetting is that repeated testing results in relatively high levels of storage strength for successfully retrieved items. This could explain the general finding that, after a delay, more items are above the threshold for successful retrieval following conditions of repeated testing compared to conditions of repeated studying. The results from the present study concerning the delayed relearning of forgotten items could be interpreted in a similar fashion: greater storage strength for forgotten (unrecalled) items learned under conditions of repeated testing can also facilitate delayed relearning. Consequently, it takes fewer study trials to restore the retrieval strength of forgotten (unrecalled) tested items to a point where these items can be successfully retrieved on a cued recall test. Interestingly, the new theory of disuse has also been used to account for the finding that, within a single learning session, immediate relearning of items is faster following retrieval-induced forgetting (Storm, Bjork, & Bjork, 2008). The present study extends these findings by showing that prior retrieval practice can enhance relearning even after a relatively long (1-week) delay.

To conclude, the present study demonstrates that the benefits of repeated testing are not limited to conscious recall on a delayed test, but extend beyond the recall threshold. It has been argued that, in a real-life setting, the least educators can hope for is that forgotten information can be relearned quickly (Nelson, 1971; Rawson & Dunlosky, 2011). The present study shows that repeated testing can be an effective strategy to facilitate such delayed relearning.

Chapter 8

Differential Effects of Retrieval Practice on the Retention of Coherent and Incoherent Text Material[*]

[*] Chapter 8 was submitted for publication as: de Jonge, M., Tabbers, H. K., & Rikers, R. M. J. P. (2013). Differential effects of retrieval practice on the retention of coherent and incoherent text material.

Abstract

Research has shown that retrieval practice can enhance long-term retention of text material. In two experiments we investigated the effect of retrieval practice with a fill-in-the-blank test on the retention of complex text material. In Experiment 1, using a coherent text, we found no retention benefit of retrieval practice compared to a restudy (control) condition. In Experiment 2, text coherence was disrupted by scrambling the order of the sentences from the text. The material was subsequently presented as a list of facts as opposed to connected discourse. For the incoherent version of the text, retrieval practice slowed down the rate of forgetting compared to a restudy (control) condition. The results suggest that the connectedness of materials can play an important role in determining the magnitude of testing benefits for long-term retention. Retrieval practice with a completion test seems especially beneficial for unconnected materials and less so for highly structured materials.

In recent years there has been a resurgence of interest for the potential benefits of retrieval practice on long-term retention. Research has shown that taking tests during learning can have profound effects on later recall compared to less demanding learning strategies like repeated study (Roediger & Karpicke, 2006a, 2006b). The general findings are especially surprising, since repeated study will most often result in superior performance on a recall test given shortly after learning. However, this short-term benefit is not long lasting. Repeated study will generally result in a relatively fast rate of forgetting, while successful retrieval of information during learning slows down the rate of forgetting (Carpenter, Pashler, Wixted, & Vul, 2008; Wheeler, Ewers, & Buonanno, 2003). Consequently, testing generally results in superior recall performance after a relatively long retention interval. This so-called *retrieval practice effect* (also known as *the testing effect*) has been found with different types of materials, different types of tests and using a variety of retention interval conditions (Roediger & Karpicke, 2006a). Claims have been made that the testing effect is of critical importance for education, and these claims have been corroborated by studies replicating the general findings in actual classroom settings (e.g., Carpenter, Pashler, & Cepeda, 2009; McDaniel, Anderson, Derbish, & Morrisette, 2007).

The powerful effect of retrieval practice for simple verbal material has been consistently found using different types of tests (e.g., Carpenter, Pashler, & Vul, 2006; Carpenter et al., 2008; Karpicke & Roediger, 2007, 2008; Kuo & Hirshman, 1996; Pyc & Rawson, 2009; Toppino & Cohen, 2009; Wheeler et al., 2003). However, the positive effect of testing on retention appears to be less robust in studies using complex materials like texts. Especially those studies using test formats that are commonly used in education (i.e., short answer questions) have come up with somewhat inconsistent results. Some studies have found benefits of retrieval practice only when feedback was provided after taking a test, but not when feedback was withheld (e.g., Kang, McDermott, & Roediger, 2007; LaPorte & Voss, 1975). One reason why these studies might have failed to find a benefit of retrieval practice without feedback could be due to low initial retrieval on the practice test (Kang et al., 2007; Wenger, Thompson, & Bartling, 1980). For instance, in the Kang et al. (2007) study, recall on an initial practice test was only 54% correct. The authors hypothesized that giving corrective feedback could restore the effectiveness of the test. Accordingly, in Experiment 2, they found that testing with feedback enhanced 3-day recall performance relative to a restudy (control) condition.

Other studies have found benefits of retrieval practice over restudy even when no feedback was given to participants. For instance, Nungester and Duchastel (1982) found that taking a short answer test enhanced long-term recall performance for a factually oriented history passage. Also, in a more recent study by Hinze and Wiley (2011) found similar results using complex expository science texts and a fill-in-the-blank test. Interestingly, in their study positive effects of taking a test were found even though performance on the initial practice test was well below 50% correct. In Experiment 1 of their study, taking a fill-in-the-blank test enhanced recall performance on a similar test given two days later, compared to a restudy (control) condition. In Experiment 2, they found that taking a fill-in-the-blank test enhanced recall on a test given after a 1-week delay. However, in Experiment 3 of their study, taking a fill-in-the-blank test did not enhance recall on a subsequent multiple-choice test given two days later. This finding is especially surprising, since initial practice test performance was considerably higher in Experiment 3 (62% correct) compared to performance in the first two experiments (44%, and 45% correct, respectively). This indicates that the failure to obtain a retrieval practice benefit in Experiment 3 was not due to insufficient recall on the practice test. The authors suggest that the failure to obtain a retention benefit of retrieval practice in Experiment 3 of their study might be due to the change in test format on the final test. However, as they also note, other researchers have generally found evidence suggesting that taking a short answer test can facilitate later multiple choice test performance (Kang et al., 2007; McDaniel et al., 2007; Nungester & Duchastel, 1982). In other words, the absence of a retrieval practice benefit in Experiment 3 of the Hinze and Wiley (2011) study cannot be readily explained.

To sum up, the retrieval practice effect is a well-established phenomenon in the literature. However, the effect appears to be less consistent and less robust in studies using text material compared to studies using simple verbal materials. Interestingly, other researchers have made similar observations across different types of materials in the past. In very early studies, it was already noted that the benefit of testing varied considerably across different kinds of materials (e.g., Gates, 1917; Kühn, 1914). For instance, Kühn (1914) found that the benefit for nonsense syllables was quite large. However, for learning series of words the benefit was smaller, and for learning short verses testing was least beneficial. Kühn concluded that the relative advantage of testing appeared to increase as the to-be-learned materials became less meaningful. Gates (1917) obtained similar results for unconnected material (nonsense syllables) and connected material (biographies). He concluded that testing appeared to be most beneficial for unconnected material and less so for connected material.

In the present study, we aimed to investigate two possible explanations for the inconsistencies in testing effect studies using text materials. A first possibility is that the inconsistencies reported in the literature are simply the result of the way recall was assessed. In most testing studies using texts and short answer tests, recall was assessed only at a single retention interval. In the present study, we assessed recall at multiple retention intervals which enabled us to investigate the rate of forgetting. Secondly, we investigated the possibility that the connectedness of the to-be-learned materials might play an important role. In Experiment 1 we used a highly coherent text as to-be-learned materials, while in Experiment 2 the same information contained in the coherent text was presented as a list of facts rather than connected discourse.

Experiment 1

One possible explanation for the inconsistent results in testing effect studies using short answer questions could be the way recall performance was assessed. Testing effect studies using short answer questions have almost exclusively assessed recall after relatively long retention intervals of days or even weeks (e.g., Butler, 2010; Duchastel, 1981; Hinze & Wiley, 2011; Kang et al., 2007; LaPorte & Voss, 1975; Nungester & Duchastel, 1982). Assessing recall performance at a single point in time makes it impossible to directly investigate the course of forgetting. As noted earlier, one of the unique advantages of taking a recall test is that it slows down the rate of forgetting (Wheeler et al., 2003). Since testing effect studies using short answer questions have assessed recall only after relatively long intervals, we do not know how short answer tests might affect the course of forgetting. For instance, a benefit of testing found after a relatively long interval could also be the result of an initial difference between conditions which has simply persisted over the course of the retention interval. This possibility pertains especially to those studies using tests with corrective feedback during initial learning, because testing with feedback can also improve recall performance after a relatively short retention interval (e.g., Butler, Karpicke, & Roediger, 2008). On the other hand, it could also be the case that the absence of a testing effect found after a certain interval reflects the point in time where the respective forgetting functions following different conditions of practice crossover (e.g., Wheeler et al., 2003). In that case, there can be no apparent difference in recall performance after a relatively long interval even though the preceding courses of forgetting were different.

In sum, the conflicting results in testing effect studies using text materials and short answer tests could simply stem from the fact that recall was assessed solely after a single long retention interval. Perhaps the results from these studies would have been more consistent if the course of forgetting had been the subject of investigation. In Experiment 1 of the present study we investigated this possibility. Instead of looking at recall performance after a single long (1-week) retention interval, we also included a short (5-min) retention interval. If taking a short answer test improves the retention of text material, then the rate of forgetting over the course of the retention interval should be slower following a short answer test compared to a restudy (control) condition.

Method

Participants

Sixty-nine psychology students from the Erasmus University Rotterdam participated in partial fulfilment of course requirements. Five participants were excluded for failing to show up for the 1-week session of the experiment.

Materials

For the purposes of the present experiment a Dutch text about black holes was created. The text was 1070 words in length and consisted of 60 sentences. The information presented in the text was taken from several online sources (see Appendix). To obtain a rough estimate of readability for the black hole text, we used the sentence-to-sentence comparison feature on the Latent Semantic Analysis website (http://lsa.colorado.edu/). The average sentence-to-sentence cosine for a translated version of the black hole text was .39, indicating that the text was highly coherent (Foltz, Kintsch, & Landauer, 1998).

For testing purposes a short answer (fill-in-the-blank test) was created similar to the test used by Hinze and Wiley (2008). The test was created in such a fashion that it closely matched the restudy (control) condition. The test contained the exact same 60 sentences presented in the black hole text, but with information selectively omitted from it. Every single sentence contained one omission covering between one and three words in length. To get an estimate of prior knowledge, we asked 10 additional participants to answer the questions without having read the text prior to taking the test. Naturally, these participants did not participate in any of the other experiments using the black hole text. On average participants were able to correctly answer 12% (*SD* = 4%) of the questions. Table 1 shows a translated excerpt from the black hole text with corresponding fill-in-the-blank questions. E-prime (Psychology Software Tools, Pittsburgh, PA) was used to create and run the experiment.

Design and Procedure

A 2 x 2 between-subjects design was used with learning condition (restudy vs. testing) and retention interval (5 min vs. 1 week) as independent variables, and test score as dependent variable.[1] Participants were randomly assigned to learning conditions and retention interval conditions. Similar to the benchmark

[1] Of main interest in the present study was recall performance. However, we would like to point out that all participants also received an evaluation questionnaire about the materials used in the experiment and a transfer test given after completion of the final recall test. The evaluation and transfer data were collected for exploratory purposes. For the sake of brevity, the results are not reported here. However, interested readers can obtain the results from the first author upon request.

Table 1
Translated (from Dutch) excerpt from the Black Hole Text with Corresponding fill-in-the-blank Questions

Excerpt from the Text	Most black holes rotate, because the stars from which they are formed also rotate. Space outside a rotating black hole is dragged along with the black hole. The result is a sort of cosmic whirlpool where it is impossible for objects to remain stationary. This area, where everything is forced to move with the black hole, is called the ergosphere.
fill-in-the-blank Questions	Most black holes rotate, because the .... from which they are formed also rotate. Space outside a rotating black hole is .... with the black hole. The result is a sort of cosmic whirlpool where it is impossible for objects to .... This area, where everything is forced to move with the black hole, is called the ....

study by Roediger and Karpicke (2006b), time on task was equated for the different learning conditions. In the first session of the experiment, all participants first studied the text during a 15-min learning trial. The text was presented one sentence at a time in the middle of the computer screen and participants could proceed to the next sentence in the text by pressing the ENTER-key. This kind of sentence by sentence reading procedure is a commonly used procedure in research on text coherence and comprehension (see also Lorch & O'Brien, 1995). Note that, because study was self-paced, it was possible to read the sentences more than once. After the last sentence of the text had been studied, the text was presented again one sentence at a time. Participants continued to study the text in this manner until the total of 15 minutes study time had expired. At the bottom left of the screen participants received feedback about their progress (e.g., 3/60 indicated a participant was currently reading the third sentence out of 60 sentences) and at the bottom right of the screen the remaining time was displayed. Upon completion of the first 15-min block, instructions diverged. During the subsequent 15-min study block, one group of participants continued to study the text material, whereas the other group of participants received a 15-min fill-in-the-blank test. Participants in the testing condition were told that the text would again be presented to them, but that each sentence would now have some information omitted from it. They were told that they should try and complete the sentences by typing in the missing information using the keyboard. No corrective feedback was given during testing. As in the initial block, both restudy and testing were self-paced, so participants could go through the text or test more than once.

Following the learning phase, all participants worked on Sudoku puzzles for five minutes as a distractor task. Afterwards, half of the participants received a final fill-in-the-blank test identical to the one used in the learning phase of the experiment. The other half of participants received the final test one week later.

Results and Discussion

Scoring

The responses on the cued recall test were scored by awarding 1 point for every correct response, 0.5 points for partially correct responses, and 0 points for completely incorrect responses. For a small number of items paraphrases were possible. Paraphrased responses that contained the same meaning conveyed by the original text were scored as correct.

Learning Phase

For both conditions, we calculated the average number of study or test cycles during the initial learning phase (i.e., the mean number of sentences processed divided by the total number of sentences in the text). During the first block, participants in the restudy condition studied the text 2.74 times (*SD* = 1.01), and participants in the testing condition studied the text 2.45 times (*SD* = .79). The difference in number of study cycles did not reach the level of significance, $F(1, 62) = 1.69$, $p = .20$. During the second block, participants in the restudy condition studied the text 2.85 times (*SD* = .85), while participants in the testing condition went through the test 1.60 times (*SD* = .73). Analysis showed that, for the second block, the difference in number of cycles was significant, $F(1, 62) = 39.52$, $p < .001$, $\eta_p^2 = .39$, indicating that the fill-in-the-blank test was more time consuming compared to simply restudying the information. This finding is not surprising, and in line with the general idea that overt retrieval practice requires more time and effort compared to restudying (see also Roediger & Karpicke, 2006a). On average, participants in the testing condition scored 67% correct on the test.

Recall Performance

Figure 1 shows the mean proportion of correct recall for both learning conditions as a function of retention interval. Participants in the 5-min group outperformed the participants in the 1-week group (70% vs. 50%), $F(1, 60) = 27.48$, $p < .001$, η_p^2

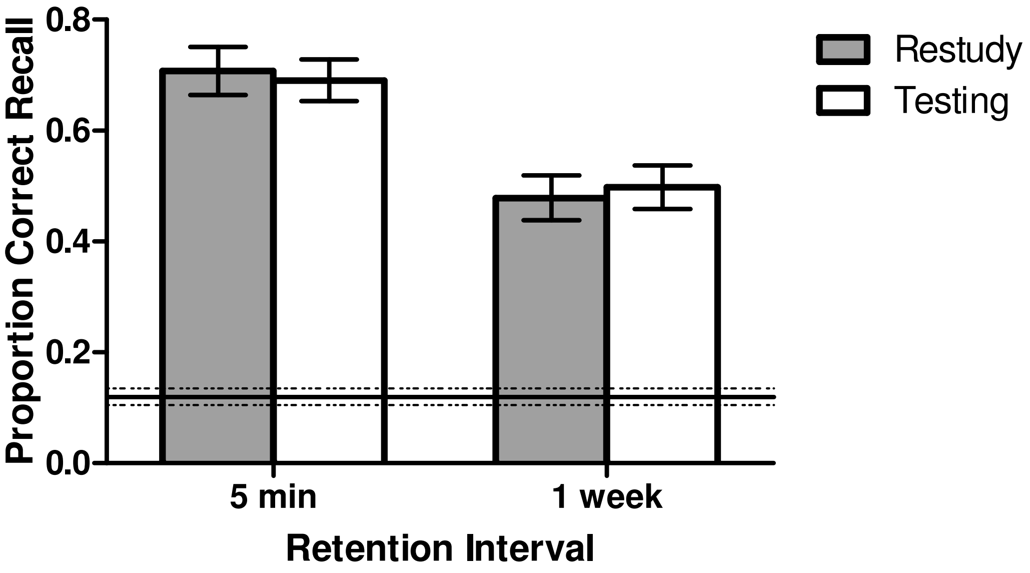

Figure 1. Proportion correct on the final recall test as a function of learning condition and retention interval in Experiment 1. The horizontal line represents baseline recall test performance for the coherent fill-in-the-blank test used in Experiment 1. Error bars represent standard errors of the means.

= .31, suggesting that forgetting occurred during the 1-week interval. However, there was hardly any difference between the restudy and the test conditions at both intervals. On the 5-min test, participants in the restudy condition correctly recalled 71%, and participants in the testing condition correctly recalled 69%. On the 1-week test, participants in the restudy condition correctly recalled 48%, and participants in the testing condition correctly recalled 50%. The main effect of learning condition and the learning condition × retention interval interaction did not reach the level of statistical significance (both *F*s < 1). Thus, we did not find a difference in the rate of forgetting between restudy and testing.

Experiment 2

The results from Experiment 1 extend those from previous studies. By looking at recall performance after multiple retention intervals rather than using a single long-term interval, we investigated the effect of taking a fill-in-the-blank test on the rate of forgetting. Importantly, however, we found no evidence for the idea that taking a fill-in-the-blank test can slow down the rate of forgetting. The results from the present study and those from previous studies (e.g., Kang et al., 2007; Hinze & Wiley, 2011) seem to suggest that the benefits of retrieval practice are less robust for complex text material. This could be related to some critical aspect of the materials used. One distinctive feature of text material is the highly structured and organized fashion by which information is presented. A text is not simply a list of facts that has been randomly put together, but rather it is a coherent set of ideas presented in a very particular and logical order. Studies have shown that text coherence can have profound effects on the retention of text material (Britton & Gülgöz, 1991; Kintsch, 1994). Especially when readers have little prior knowledge, text coherence is a very important factor determining learning from text (McNamara, 2001; McNamara & Kintsch, 1996).

The issue of text coherence has received very little consideration in research on the testing effect. Still, the notion that the organisation or connectedness of materials might attenuate the effect of retrieval practice is not entirely new. Earlier research has shown that the benefits of testing can vary considerably across different kinds of materials and it has been suggested that the connectedness of to-be-learned materials could play an important role determining the magnitude of retrieval practice benefits (e.g., Gates, 1917; Kühn, 1914). To investigate the possible role of coherence, we conducted a second experiment. In Experiment 2, we disrupted the connectedness of the text material used in Experiment 1 by presenting the information contained in the text as a list of randomly ordered facts rather than connected discourse.

Method

Participants

Seventy psychology students from the Erasmus University Rotterdam participated in partial fulfilment of course requirements. None of the participants had participated in Experiment 1. Data from five participants were excluded form analysis, because they failed to show up for the 1-week session of the experiment. Data from one participant were excluded for failing to follow basic instructions.

Materials

The coherence of the text used in Experiment 1 was disrupted by presenting the sentences in a scrambled order. In order to be comprehensible out of context, it was necessary to make some minor changes to the sentences taken from the black hole text. For instance, in some sentences an adverb was deleted (e.g., “So, black holes are…” was changed to “Black holes are…”). Also, in some sentences anaphoric references were replaced by their corresponding nouns (e.g., “they” was replaced with “black holes”). The average sentence-to-sentence cosine (http://lsa.colorado.edu/) of the scrambled text in Experiment 2 was significantly lower (*M* = 0.23, *SD* = 0.19) than the cosine of the text used in Experiment 1 (*M* = 0.39, *SD* = 0.22), *t*(116) = 4.16, *p* < .001, *d* = 0.77, indicating that the disruption of the text coherence had been successful. A fill-in-the-blank test was subsequently devised containing the exact same omissions as the test used in Experiment 1. The presentation order of items on the scrambled fill-in-the-blank test was kept constant throughout the experiment. As in Experiment 1, we asked 10 additional participants to answer the questions without having studied the materials prior to taking the test. Baseline test performance for the scrambled version of the test was similar to performance in Experiment 1. On average participants were able to correctly answer 11% (*SD* = 7%) of the questions.

Design and Procedure

As in Experiment 1, we used a 2 (learning condition) × 2 (retention interval) between-subjects design. The procedure was virtually identical to the one used in Experiment 1. The only important difference was the way we referred to the to-be-learned materials in the instructions. In the present experiment, participants were told that they would learn a list of facts about black holes.

Results and Discussion

Learning Phase

During the first block, participants in the restudy condition studied the list of facts 1.97 times (*SD* = .69) and participants in the testing condition studied the text 1.86 times (*SD* = .55). The difference in number of study cycles did not reach the level of significance, $F < 1$. During the second block, participants in the restudy condition studied the list of facts 2.29 times (*SD* = .94), while participants in the testing condition went through the test 1.45 times (*SD* = .44). As in Experiment 1, taking the test was more time consuming compared to simply restudying the list of facts, $F(1, 62) = 21.0$, $p < .001$, $\eta_p^2 = .25$. Participants in the testing condition scored 54% correct on the test.

Recall Performance

Figure 2 shows the mean proportion of correct recall for both learning conditions as a function of retention interval. There was a significant main effect of retention interval, $F(1, 60) = 6.42$, $p < .05$, $\eta_p^2 = .10$. Participants in the 5-min group recalled more on the final test (54%) compared to participants in the 1-week group (45%). The main effect for learning condition did not reach the level of significance, $F < 1$. Importantly, however, there was a significant learning condition × retention interval interaction, $F(1, 60) = 4.13$, $p < .05$, $\eta_p^2 = .06$. As can be seen in Figure 2, the restudy group showed a substantial amount of forgetting (31%). However, for the testing group there was virtually no decline in recall performance across the 1-week interval. Accordingly, follow-up analysis revealed that the effect of retention interval was significant for the restudy condition,

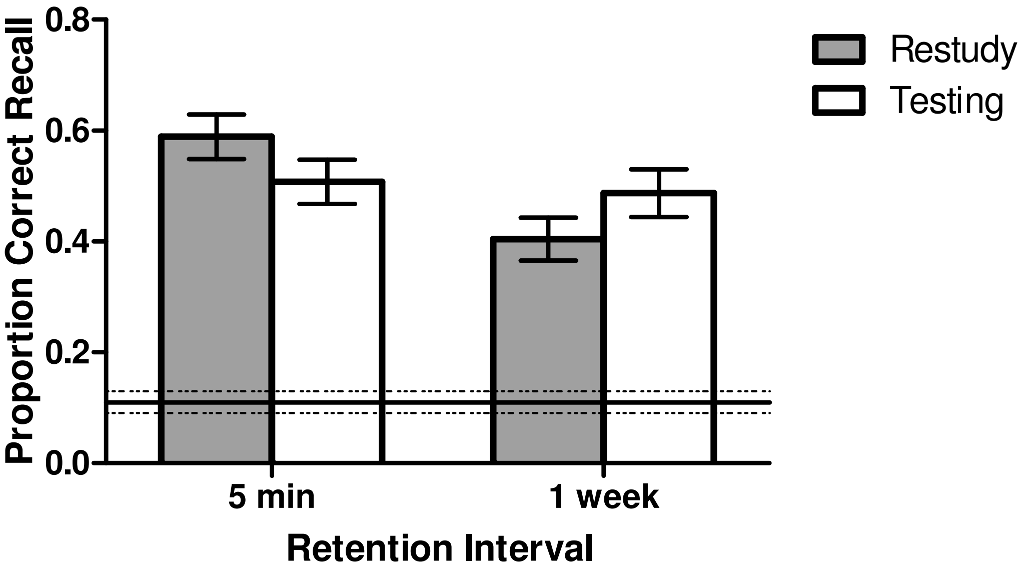

Figure 2. Proportion correct on the final recall test as a function of learning condition and retention interval in Experiment 2. The horizontal line represents baseline recall test performance for the scrambled fill-in-the-blank test used in Experiment 2. Error bars represent standard errors of the means.

$t(30) = 3.33$, $p < .001$, $d = 1.18$, but not for the testing condition ($t < 1$). Thus, for the incoherent materials used in Experiment 2, we observed a difference in rate of forgetting between the restudy (control) condition and the testing condition.

General Discussion

In the present study, we aimed to investigate two possible explanations for the inconsistencies in testing effect studies using text materials and completion tests. One possibility was related to the way recall was assessed in most previous studies. As noted, in most studies, recall was assessed after a single long-term retention interval. In the present study, we assessed recall at multiple retention intervals which enabled us to investigate the rate of forgetting. In Experiment 1, we found no retention benefit of retrieval practice compared to a restudy (control) condition for a highly coherent text. The testing group and the restudy group showed comparable rates of forgetting over the course of the 1-week interval. Thus, for the coherent text material used in Experiment 1, we found no evidence suggesting that taking a fill-in-the-blank test can slow down the rate of forgetting. However, in Experiment 2, when text coherence was disrupted, we found that retrieval practice effectively slowed down the rate of forgetting compared to a restudy (control) condition. Taken together, the results indicate that the benefits of retrieval practice can be dependent on the connectedness of the materials used.

Past research on text coherence has shown that the connectedness of material can have a powerful effect on later recall of text material (Britton & Gülgöz, 1991; Kintsch, 1994). Since we disrupted the coherence of the black hole text and presented the material as a list of facts in Experiment 2, one would expect that test scores would be lower in Experiment 2 compared to Experiment 1. Inspection of the retention test scores in Experiments 1 and 2 suggests that this was indeed the case. Averaged across conditions, participants in Experiment 2 performed worse compared to the participants in Experiment 1 on the retention test (50% vs. 59% correct), and also on the practice test (54% vs. 67% correct). As already noted, researchers have argued that retrieval practice can sometimes be ineffective when recall is relatively low on an initial practice test (e.g., Kang et al., 2007). However, in Experiment 2 of the present study, we found a benefit of retrieval practice over restudy even though recall performance on the initial practice test was considerably lower compared to performance in Experiment 1. Thus, it seems unlikely that the absence of a retrieval practice benefit in Experiment 1 was due to insufficient recall on the practice test.

In Experiment 2, using an incoherent list of facts, we found evidence suggesting that retrieval practice can slow down the rate of forgetting. It has been argued that tests appear to slow down the rate of forgetting because taking a practice test can result in stronger memory traces for successfully retrieved items compared non-recalled items or restudied items (Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011). One reason why tests might result in stronger memory traces is offered by the *elaborative retrieval hypothesis* (e.g., Carpenter, 2009; Carpenter & DeLosh, 2006). This hypothesis suggests that testing will result in more elaborate memory traces compared to passive restudy of information. Support for this hypothesis has been provided by studies showing that the effect of testing can get more pronounced as the amount of cue-support on the practice tests diminishes. For instance, in a study by Carpenter & DeLosh (2006), it was found that retrieving items with fewer letter cues was associated with better final recall test performance. One way to explain the results from the present study could be in light of the elaborative retrieval hypothesis. As already noted, in a coherent text, ideas are presented in a very particular logical order. It has been argued that the organizational structure of text materials can also serve as a retrieval cue to enhance later recall (Shimmerlick, 1978). Likewise, in the present study, the coherent context of the materials used in Experiment 1 might have functioned as a retrieval cue. One could argue that the test used in Experiment 1 might not have resulted in more elaborate processing relative to the processing already invited by the cue-support provided by the context of the text. However, for the isolated statements in Experiment 2, the absence of the supporting context of the text might have resulted in more elaborative processing on the test. Investigating this possible explanation could be a fruitful avenue to pursue in future research.

In the present study we investigated the retrieval practice effect using a short answer test. Clearly our conclusions are limited to the test format used. Also, importantly, some studies have found substantial memorial benefits for text material using more demanding test formats like free recall tests (e.g., Hinze & Wiley, 2011; Karpicke & Blunt, 2011; Roediger & Karpicke, 2006b). Interestingly, research suggests that taking a free recall test can also facilitates organizational processing of materials (Congleton & Rajaram, 2011; 2012; Zaromb & Roediger, 2010). On a free recall test, learners do not simply retrieve information from memory in an arbitrary order, but rather they retrieve the information in an organized fashion. For instance in the study by Zaromb and Roediger (2010), organizational processing was investigated in a testing effect paradigm. They found that taking a free recall test enhanced the retention of categorized lists of words. More importantly, they also found that prior testing improved category clustering, indicating that taking a free recall test might be associated with enhanced organizational processing. In the case of learning from text, organizational processing seems especially important. Perhaps a free recall test might be a more potent device for improving the retention of complex text material compared to a short answer test.

To conclude, our research shows that the benefit of retrieval practice can be dependent on the connectedness of the materials. Importantly, this does not mean that retrieval practice as a learning activity might not be useful for learning text materials. There is an overwhelming amount of support for the retrieval practice effect even when using complex materials like texts (e.g., Hinze & Wiley, 2011; Karpicke & Blunt, 2011; Nungester and Duchastel, 1982; Roediger & Karpicke, 2006b). Moreover, even those studies that, like the present one, have failed to find memorial benefits of testing under the strictest of control conditions (like review or restudy), certainly did not find any disadvantage of taking an intermediate test on long-term retention. However, our results do indicate that the effectiveness of retrieval practice can be dependent on the connectedness of the materials. Although more research is necessary to establish to which extent coherence plays a role in the testing effect, the results from the present study are promising and in line with earlier observations (e.g., Gates, 1917; Kühn, 1914). The present study represents a first step towards explaining the discrepancy between different kinds of materials by addressing the issue experimentally using material of differential coherence, but equal content. On the basis of our results, we have identified coherence as a possible factor determining the relative benefits of retrieval practice.

Chapter 9

Summary and Discussion

Of interest in the present thesis was the situation where students have a limited amount of time to learn by themselves a certain amount of information. The main question was how students might get the most out of such a limited amount of time. On the one hand, we were interested in learning strategies that can be used to make the initial learning of information more efficient. However, at the same time, we also considered the efficacy of learning in regard to the long-term retention of information. In the first part of the thesis we investigated the effect of study time distribution and the second part of the thesis focused on the potential benefits of retrieval practice. Below we briefly discuss the main findings and conclusions from the studies in the thesis and we will make some suggestions for future research.

Study Time Distribution

In the first part of the thesis, we investigated how study time should be distributed within a single (short) learning session to be optimally effective. First of all, we asked ourselves the question, given a limited amount of study time, in how many presentations should the available time be divided? Second, we asked ourselves the question whether the optimal distribution of study time would be dependent on the relative difficulty of the to-be-learned materials. Third and last, we were interested in how effective learners are at distributing study time during single-session learning episodes when they are allowed to self-pace.

In **Chapter 2**, Participants studied unrelated word pairs under different presentation rate conditions ranging from relatively fast (e.g., 16 presentations of 1 s per pair) to relatively slow (a single presentation of 16 s per pair). The total amount of study time per item was equated. We found a non-monotonic relationship between presentation rate and final cued recall performance. Both fast (e.g., 1 s) and slow (16 s) presentation rates resulted in poor immediate and delayed recall performance compared to intermediate (e.g., 4 s) presentation rates. We concluded that there appears to be a Goldilocks principle at work with regards to presentation rate: Presentation rates should be not too long, not too short, but just right.

In **Chapter 3**, we replicated and extended the findings from Chapter 2 using more ecologically valid materials. Participants (English speaking students) studied Dutch-English translation pairs under different presentation rate conditions. Like in the previous study, we found a non-monotonic relationship between presentation rate and cued recall performance. Again, both fast (e.g., 1 s) and slow (16 s) presentation rates resulted in poor immediate and delayed recall performance compared to intermediate (e.g., 4 s) presentation rates. Most importantly, we obtained a “Goldilocks” pattern of results regardless of language direction. For both translation directions (Dutch → English and English →

Dutch), performance was best for intermediate presentation rates and dropped off for short (1 s) and long (16 s) presentation rates. The results indicate that the effect of presentation rate is not just limited to artificial materials that are often used in a laboratory setting. Presentation rate is also an important factor to consider for the learning of educationally relevant materials like foreign vocabulary word pairs. Moreover, it seems that the optimal presentation rate during learning does not necessarily shift with difficulty of recall. Even though recall performance was worse for the English→Dutch direction compared to the Dutch→English direction, presentation rate had similar effects for both language directions.

Lastly, in **Chapter 4**, we compared a variety of fixed presentation rate conditions to a condition where participants were allowed to self-pace. First of all, in contrast to the constant presentation durations in the fixed conditions, we found that presentation rate varied considerably across study cycles when learners were allowed to self-pace. Self-paced learners showed a clear tendency to increase the pace as learning progressed. Participants started out with a reasonably slow presentation rate the first time through the list, but they eventually ended up with a relatively fast presentation rate by the last pass through the list. Secondly, for the fixed presentation rate conditions, we again replicated the "Goldilocks" pattern of results. Intermediate presentation rates resulted in better recall performance, compared to very fast and very slow presentation rates. Most importantly, however, we found that self-paced study resulted in better overall recall performance than most of the fixed presentation rates, with the exception of the 12 × 2 s condition which did not differ from the self-paced condition. Furthermore, in Experiment 2, we provided evidence for the idea that the opportunity to allocate study time as a function of item difficulty during self-paced study might be a crucial factor determining later test performance. When learners were allowed to self-pace, but total study time per item was equated, recall performance deteriorated. We concluded that learners can be quite proficient when it comes to allocating self-paced study time during multitrial learning.

The results in Chapter 2, 3, and 4 are unambiguous with regards to the effect of fixed presentation rate on learning and retention. In all three studies there was a substantial effect of presentation rate on later recall performance, with intermediate presentation rates resulting in superior test performance. The presentation rate effect appears to be very robust and generalizes across different types of simple verbal materials ranging from nonsense syllables paired with digits (e.g., Calfee & Anderson, 1971; Johnson, 1964) to more meaningful materials like unrelated word pairs (de Jonge, Tabbers, Pecher, & Zeelenberg, 2012), and foreign language vocabulary word pairs (de Jonge, Tabbers, Pecher, Jang, & Zeelenberg, 2013; Zeelenberg, de Jonge, Tabbers, & Pecher, 2013). Based on our research findings, we conclude that, for alphabetic language materials, a presentation rate round about 4 s seems a good rule of thumb for efficient and effective paired-associate learning.

One possible explanation for the presentation rate effect is the *effective study time hypothesis* (de Jonge et al., 2012). That is, some minimal amount of time might be necessary for learners to optimally form an association during study (Stubin, Heimer, & Tatz, 1970). However, at the same time, presentation durations beyond some optimal value might result in some form of deficient processing (Bugelski & McMahon, 1971). Also, theories on the distributed practice effect might, in part, explain the findings in the present thesis. For example, factors such as *encoding variability* and *study phase retrieval* may play a role in our findings. The encoding variability account of the distributed practice effect assumes that context fluctuates over time and that encoding materials in different contexts enhances memory performance (e.g., Glenberg, 1976; Melton, 1967). Encoding variability can explain why an intermediate presentation rate (e.g., 4 × 4 s) results in better performance than a slow presentation rate (e.g., 1 × 16 s). That is, at an intermediate presentation rate, learners will receive more repetitions resulting in more variable contextual elements encoded in the resulting memory trace. However, encoding variability cannot explain why a very fast (e.g., 16 × 1 s) presentation rate would result in inferior recall performance compared to more intermediate rates of presentation (at least not without making additional assumptions). Likewise, the study phase retrieval account could explain why performance in the 4 x 4 s condition is better than in the 1 × 16 s condition. More repetitions are bound to result in more potentially successful study phase retrievals. However, without making additional assumptions this account would also predict optimal performance for fast multi-repetition presentation rates (e.g., 16 × 1 s). One possibility would be to make the additional assumption that study phase retrieval takes a certain amount of time to be successful. With total study time held constant, a presentation duration of 1 s may provide many potential opportunities for retrieval of an earlier encoding of the same item, but very few of these retrieval opportunities may be successful. In that case, the study phase retrieval theory could also account for the deteriorating performance observed for fast presentation rates.

We believe that the results from these studies are not just interesting from a theoretical point of view, but they also hold substantial practical relevance. First of all, our results have important implications for research practice. In many experiments on learning and retention, where controlling for time on task is of the essence, the control condition to which an experimental manipulation is compared is a repeated study condition. For instance, this is often the case in research on the testing effect. Previous testing effect studies using simple verbal materials have compared testing conditions to a variety of restudy (control) conditions with presentation rates varying from 2 s (e.g., Zaromb & Roediger, 2010) up to 10 s (e.g., Carrier & Pashler, 1992). Most testing effect studies have, however, used intermediate (+/- 4 s) presentation rates for the restudy (control) condition. Still, when deciding upon the presentation rate of a restudy (control) condition, researchers are well-advised to keep in mind the profound effect presentation rate can have on learning and retention.

Secondly, the results from our studies also hold particular relevance for educational purposes. For instance, they could be used to enhance computerised foreign vocabulary learning. Recently, successful attempts have been made to optimize single-session foreign vocabulary learning using the ACT-R modelling system (Anderson & Lebiere, 1998) to schedule the spacing of repetitions of individual items (e.g., Pavlik & Anderson, 2008; van Rijn, van Maanen, & van Woudenberg, 2009). During the study trials in these experiments, items were presented with a fixed presentation rate comparable to the fixed rate found to be optimal in our studies. Importantly, however, in Chapter 4 we found that learners generally perform better when they are allowed to self-pace compared to studying with a fixed experimenter imposed presentation rate. This suggests that, for single-session foreign vocabulary learning, self-paced study might be more efficient. As noted, it has been argued that people often make suboptimal decisions during learning, because they do not have an accurate picture of the complexities of their own memory (Kornell & Bjork, 2009). Moreover, even when learners do have accurate metacognitions, it is unclear if they are able to put this knowledge to use (Son & Metcalfe, 2000). Notwithstanding, in our study, we found that learners were well able to allocate study time effectively during multitrial self-paced learning. Although the differential allocation of study time strategy employed by most learners might not be optimal to the point that they can fully compensate for the difficulty of individual items in a list, they are certainly not labouring in vain either. When learners are forced to indiscriminately use an equal amount of study time for all items in a list, their recall performance will deteriorate.

Note that one limitation of our study was that we did not look at the long-term efficacy of self-paced study trials. Thus, although our results indicate that self-paced study can be efficient, the question remains whether self-paced study is also preferred when long-term retention is concerned. The long-term efficacy of self-paced learning has not yet been investigated and this issue should be addressed in future research.

One interesting finding in Chapter 4 was the inclination of self-paced learners to increase the rate of presentations as learning progressed. Participants started out with a reasonably slow presentation rate, but ended up with a relatively fast rate by the last pass through the list. The experimenter-imposed presentation rates under investigation in the present thesis were limited to schedules using constant presentation rates. It would be interesting for future purposes to investigate whether fixed presentation rate schedules can be further optimized by mimicking the pattern of self-paced learners. That is, start out with a relatively slow fixed presentation rate for the first occurrences of items in a list, and increase the rate of subsequent repetitions as learning progresses. For one thing, in terms of study phase retrieval, one could argue that the first occurrence of an item, and the resulting memory representation after encoding, is of critical importance. If the original memory trace for an item is insufficient to induce study phase retrieval during a subsequent presentation, then no additional information will be added to the original trace (e.g., Raaijmakers, 2003). Thus, providing learners with ample study time during the first presentation of an item might facilitate study phase retrieval success on any subsequent presentation. However, repetitions of previously presented items need not be excessively long. A repetition of an item might be beneficial to the degree that it reminds the learner of the previous occurrence (successful study phase retrieval) and provides sufficient time to store additional information in the original trace.

On a related note, in a recent study, Benjamin and Tullis (2010) suggested a *model of reminding* as a theoretical framework to account for the distributed practice effect. This reminding theory implies that optimal learning is achieved by successfully balancing the benefits and costs of reminders during learning. The authors suggest that too long a lag leads to unlikely reminding, and too short a lag will result in impotent reminding. In short, they argue that there might be a “sweet spot” for the optimal duration of the lag between a presentation and a subsequent reminder. The data presented in the present thesis suggest that there might also be a “sweet spot” for the duration a reminder should have to be optimally efficient. For future purposes, it would be interesting to see how these two factors, spacing and pacing, might interact. For example, fast reminders might be most efficient for short inter-presentation lags, while reminders of longer duration might be most efficient for longer inter-presentation lags.

Retrieval Practice

In the first part of the thesis, we focused on the situation where the available amount of time is allocated exclusively to studying. In the second part of the thesis, we investigated the efficacy of providing learners with test trials in addition to study trials during single-session learning. It has been argued that time allocated to test trials during learning is generally well spent, even though it takes up time that might otherwise have been utilized for additional study (e.g., Nungester & Duchastel, 1989). In our studies on the retrieval practice effect, learning conditions where some portion of the available amount of time was reserved for testing were compared to learning (control) conditions where the full amount of available time was spent studying. We investigated the benefits of retrieval practice for enhancing the long-term retention of simple verbal materials, and more complex text material. In our studies, we assessed recall performance after both short and long intervals allowing us to assess the degree of forgetting. In addition, to further extend the approach taken in previous research, we also investigated the effect of practicing retrieval during initial learning on the delayed relearning of information.

In **Chapter 5**, we investigated the potential benefit of providing learners with test trials during the learning of foreign vocabulary word pairs. In two experiments, we assessed the rate of forgetting of word pairs learned under testing conditions and a restudy (control) condition across a 1-week, and a 4-week retention interval, respectively. In Experiment 1, we found that taking tests during learning slowed down the rate of forgetting over a 1-week interval compared to a restudy (control) condition. In Experiment 2, we replicated this finding and showed that, after a 4-week interval, the respective forgetting functions crossed over. On the 4-week final test, there was a substantial benefit of testing compared to a restudy (control) condition. Taken together, the results of our study provide a clear demonstration of the powerful effect retrieval practice can have on long-term retention. Also, in Chapter 5, we discussed two important issues that have not received a lot of attention in previous testing effect research. First of all, repeated testing might result in the overlearning of a small subset of relatively easy (successfully recalled) items (Thompson, Wenger, & Bartling, 1978). The long-term benefit observed for items that are repeatedly retrieved during initial learning, might in part reflect an item selection artifact. If the benefits of testing are largely limited to those items that are relatively easy to retrieve and do not extend to the more difficult items in a list, then the benefits for educational practice will also be limited. Second, we noted that testing effect studies have mostly focused on recall performance as an outcome variable and we advocated looking at other measures of retention that might hold more practical relevance (i.e., relearning). With the discussion of these important, yet neglected, issues we also set the stage for Chapter 6 and 7.

In **Chapter 6**, we investigated the effect of repeated testing on item selection, retention, and delayed relearning of paired associates. Participants learned mixed word pair lists containing easy (related) and difficult (unrelated) word pairs under a repeated study and a repeated testing condition. During the initial learning phase of the experiment, we found that more related word pairs were successfully recalled on the practice tests compared to unrelated word pairs (i.e., item selection occurred during repeated testing). Importantly, however, long-term retention benefits were found for tested items, regardless of item difficulty. For both easy (related) and difficult (unrelated) word pairs the repeated testing condition outperformed the repeated study condition on the 1-week retention test. These results suggest that the retention benefit following conditions of repeated testing cannot be attributed to mere item selection. Secondly, as noted, in Chapter 6, we also looked at the effect of repeated testing on delayed relearning. We found that, in the beginning of the delayed relearning phase, relearning was faster for previously restudied items compared to previously tested items. After a single restudy cycle, the initial benefit of the repeated testing condition had evaporated, and both the restudy and the repeated testing condition performed about equally well on the remainder of the tests given during the relearning phase. These results suggest that repeated testing can enhance delayed recall performance with little additional cost in terms of delayed relearning.

In Chapter 6, we focused on the situation where only a subset of information was encoded and repeatedly retrieved during initial learning. In **Chapter 7**, we further examined the effect of repeated testing on delayed relearning using a learning-to-criterion procedure before introducing the experimental manipulation (repeated study vs. repeated testing). All items were first learned to the criterion of one successful retrieval from long-term memory and all items subsequently received three post-retrieval study or test trials. One week after initial learning, participants returned for delayed recall and relearning. We found that post-retrieval test trials resulted in better retention test performance than post-retrieval study trials. Also, we found that items from both learning conditions (post-retrieval study and post-retrieval testing) were relearned faster than a new set of similar (not previously presented) items. Most importantly, items were relearned faster when they had previously been learned under conditions of post-retrieval testing than items learned under conditions of post-retrieval study. These results show that the benefits of repeated testing are not just limited to conscious recall on a delayed test. Repeated testing during initial learning is also a very effective strategy to enhance delayed relearning.

Lastly, in **Chapter 8**, we argued that the positive effect of retrieval practice might be less robust for text materials. That is, studies using educationally relevant test formats (e.g., short answer questions) have come up with somewhat conflicting findings (e.g., Hinze & Wiley, 2011; Kang, McDermott, & Roediger, 2007; LaPorte & Voss, 1975; Nungester & Duchastel, 1982). Interestingly, in very early studies on the retrieval practice effect, it was already noted that the benefits of testing appeared to be less pronounced for more meaningful materials (e.g., Gates, 1917; Kühn, 1914). In the two experiments in Chapter 8, we investigated the effect of fill-in-the-blank retrieval practice on the retention of complex text material. In Experiment 1, using a coherent text, we found no retention benefit of retrieval practice compared to a restudy (control) condition.

However, in Experiment 2, when text coherence was disrupted we found that retrieval practice slowed down the rate of forgetting compared to a restudy (control) condition. The combined results suggest that the connectedness of materials might play an important role in determining the magnitude of testing benefits for long-term retention. Retrieval practice with a completion test seems especially beneficial for unconnected materials and less so for highly structured materials.

Taken together, our studies on the retrieval practice effect indicate that testing during learning can enhance the long-term retention of simple verbal materials. First of all, in Chapter 5, 6, and 7, we replicated the general finding that testing can have a substantial effect on the delayed recall of paired associates. Second, our results in Chapter 6 suggest that the effect of retrieval practice does not appear to be dependent on the difficulty of the to-be-learned materials. Third, in Chapter 7, we found that the benefit of retrieval practice is not just limited to enhanced recall performance on a delayed test, but also extends to delayed relearning. In short, our studies indicate that, for simple verbal materials, the effect of retrieval practice is quite robust. Interestingly, however, our results in Chapter 8 suggest that the benefits of retrieval practice might be less robust for more complex materials like coherent texts.

The results from our studies on the retrieval practice effect have substantial implications for theory. For instance, the combined results of Chapter 6 and 7 provide support for the bifurcation account of the testing effect. The bifurcation account suggests that repeated testing can bifurcate the distribution of item strengths on a target list, whereas repeated study will not result in a bifurcated distribution (Halamish & Bjork, 2011; Kornell, Bjork, & Garcia 2011). It is assumed that repeated testing divides item distributions into weak (unrecalled) and strong (recalled) items (e.g., Kornell, 2011). That is, on a first practice test, only a subset of items is successfully retrieved. These items are strengthened during subsequent test trials, whereas previously unrecalled items do not receive further practice and will weaken as a consequence. However, under conditions of repeated study, all items in a set are practiced continuously. As a consequence, all restudied items will get strengthened, yet to a lesser extent than the successfully retrieved items. Thus, when only a subset of items is retrieved during initial learning, the bifurcation model would predict that previously restudied items that are forgotten over time will be to be closer to the threshold for successful retrieval compared to weak (unrecalled) tested items. Consequently, one would expect rate of relearning after a delay to be faster for items that were learned under conditions of repeated study. The results of Chapter 6 supported this prediction. After a 1-week delay, rate of relearning was faster for repeatedly studied items compared to repeatedly tested items. That is, although there was an initial advantage of repeated testing on the 1-week retention test, after just one restudy cycle the repeated study condition had caught up with the repeated testing condition and both conditions remained at comparable levels during the remainder of the relearning phase.

So far, we have considered what the bifurcation account of the testing effect would predict when only a subset of items is repeatedly retrieved during initial learning. However, when all items are repeatedly retrieved during initial learning, the distribution of items strengths is not assumed to be bifurcated. Rather, one would expect that, repeatedly tested items are strengthened more than items that were repeatedly studied. Thus, under these circumstances, the bifurcation account would predict that an initial benefit of repeated testing should persist during relearning rather than evaporate early on in the relearning phase. The results of Chapter 7 supported this prediction. That is, we found that the benefit of testing persisted across the first three test cycles in the relearning phase and it took considerably less trials to reach criterion for items that were repeatedly tested during initial learning compared to repeatedly studied items or new items.

Lastly, our findings in Chapter 8 provide a challenge for theoretical accounts of the retrieval practice effect. That is, a solid theory should be able to account for our finding that the effects of retrieval practice might be moderated by the connectedness of the to-be-learned materials. As noted, it has been argued that test taking can be beneficial for learners because taking a practice test can result in stronger memory traces for successfully retrieved items (Halamish & Bjork, 2011; Kornel et al., 2011). One possible reason why tests can result in stronger memory traces is because taking a test can result in more elaborative processing compared to passive restudy of information. One way to explain the differential effects of retrieval practice for the coherent and incoherent materials in our study could be in light of *the elaborative retrieval hypothesis* (e.g., Carpenter & DeLosh, 2006). As noted, the organizational structure of a coherent text can also serve as a strong retrieval cue to enhance later recall performance (Shimmerlick, 1978). However, when text coherence is disrupted, the absence of a supporting context might result in more elaborative processing on a practice test relative to repeated study. This is one possible explanation why retrieval practice appears to be especially beneficial for unconnected materials and less so for coherent materials.

One thing that should be noted is that the results of Chapter 8 are in need of replication and extension. One weakness of our study was that the coherence of the materials was manipulated between experiments rather than within a single experimental design. Additional experimentation is essential for further strengthening our position that coherence is a possible moderator of the retrieval practice effect, especially since such a moderating effect would have far-reaching implications both for theory and practice. Secondly, in our study, both the coherence of the text materials and the practice test were disrupted in Experiment 2. Thus, we cannot be sure whether the coherence of the text or the coherence of the test was responsible for the discrepancy in results between Experiment 1 and Experiment 2. Future research should address this issue by manipulating both independently.

The results from our studies on the retrieval practice effect also hold substantial practical relevance. First of all, in line with previous findings (Karpicke & Roediger, 2007, 2008), the results from Chapter 5 indicate that repeated retrieval during learning is beneficial for long-term retention. The results from Chapter 7 provide further evidence for this finding. That is, in both experimental conditions in our study in Chapter 7, participants received a number of test trials. The benefit of the post-retrieval testing condition over the post-retrieval study condition indicates that, even when an item is learned to the degree that it can be successfully retrieved on a test, it is still beneficial to continue practicing retrieval. On a related note, researchers have recently addressed the question of how much retrieval practice is recommended for efficient and durable learning (Rawson & Dunlosky, 2011). Based on their findings, Rawson and Dunlosky recommended practicing to three correct recalls during initial learning.

The results from our research on the effect of retrieval practice on delayed relearning extend those from previous studies. As noted, in most circumstances, learned information will become unrecallable shortly after learning has taken place. Although it seems desirable to pursue ways of having students achieve perfect recall even after relatively long retention intervals, this might not be a particularly realistic or even remotely feasible goal. Likewise, it has been argued that, in an actual educational setting, when long-term retention is concerned, the least educators can hope for is that forgotten information can be relearned relatively quickly (Nelson, 1971; Rawson & Dunlosky, 2011). Thus, focusing solely on delayed recall performance as an outcome variable in experiments on learning and retention might limit the practical relevance for educational purposes. In this respect, classical memory researchers might have had it right all along. When Hermann Ebbinghaus (1885/1964) first undertook his groundbreaking study on the retention of (his own) memories, he did not choose recall performance as his measure of interest. Rather he was interested in how much practice it would take him to bring forgotten memories back to conscious recollection during relearning. Relearning might have fallen from grace as the predominant measure of retention in memory research, but if there ever was a time to reembrace the relearning method, then the time is now. In this day and age, with vast amounts of information readily available on the internet, learners might be less and less inclined to rely on conscious recollection of information. In practice, especially for more complex educationally relevant materials, students might not be able to remember everything by heart. When they need access to previously studied information later on, they would hopefully not just rely on what little they can recall. Preferably they would choose to look back at the original source of information and, in that case, they will have benefitted most from original learning, if it results in the rapid reinstatement of what was once learned, but then forgotten. The results of our study in Chapter 7 indicate that repeated retrieval practice can be an effective strategy to facilitate the delayed relearning of foreign vocabulary word pairs. For future purposes, it would be interesting to see whether our findings also extend to more complex materials (e.g., science discourse).

One important difference between the studies in Chapter 6 and 7 is that in Chapter 6 we also manipulated the difficulty of the materials used. We used mixed lists containing both easy (related) and difficult (unrelated) word pairs and this enabled us to assess the relative benefits of testing for materials of differential difficulty. Although Chapter 7 did not include such an a priori manipulation, we did look into the possible role of item difficulty in relearning a posteriori.[1] Based on a median split on the number of study trials to reach criterion for items presented as new during the relearning phase of the experiment, items were divided into 25 easy and 23 difficult word pairs. Although some caution is warranted when interpreting the results from this exploratory analysis, we think the results are interesting and we believe they could have important implications for the relearning method in general. The results of the additional analysis suggested that there was a benefit of initial repeated retrieval practice on delayed relearning of forgotten difficult items, but not so much for forgotten easy items. One possible reason for this finding could be that, during the relearning phase, most of the easy items were so close to threshold that a single 5 s study trial was sufficient to bring them back into conscious memory and allow retrieval on the subsequent test trial. This might have been the case for easy items, regardless of prior encoding condition. Thus, the results from our exploratory analysis suggest that, for easy items, the relearning method used in Chapter 7 might not have been particularly sensitive and that there might be some room for improvement. A recommendation for

[1] A subsequent item-wise analysis was conducted to explore the possible role of item difficulty in Chapter 7. Data were analysed using a mixed ANOVA with item difficulty (easy vs. difficult) as between-subjects variable, learning condition (repeated study vs. repeated testing) as within-subjects variable, and number of study trials to criterion as dependent variable. Interestingly, there was a significant Item difficulty × Learning condition interaction, $F(1, 45) = 4.85$, $p < .05$, $\eta_p^2 = .10$. Follow-up analysis revealed that, for unrecalled easy items, there was no significant difference in the number of study trials it took to reach criterion between the two learning conditions, $t(24) = .28$, $p = .78$. However, for unrecalled difficult items, it took significantly more trials to reach criterion in the repeated study condition ($M = 1.46$, $SD = .31$) compared to the repeated testing condition, $t(21) = 2.88$, $p < .01$, $d = 0.62$.

future studies could be to use a faster (e.g., 1 s) presentation rate for study trials during delayed relearning. Using a relatively fast presentation rate during delayed relearning could perhaps result in a more fine-grained measure of retention compared to the 5 s presentation rate used in Chapter 7.

Lastly, in Experiment 1 of Chapter 8, we did not find a retention benefit of retrieval practice for coherent text material. Only when the coherence of the text was disrupted in Experiment 2, did we obtain a more typical pattern of results suggesting that retrieval practice slowed down the rate of forgetting. Together, these results indicate that the benefit of retrieval practice might be less robust for connected materials like coherent texts. Thus, our results suggest that there might be limits to the usefulness of the testing effect for educational purposes. Akin to the *material appropriate processing framework* (McDaniel, Einstein, Dunay, & Cobb, 1986) our findings suggest that the benefits of retrieval practice might be largely dependent on the nature of the to-be-learned materials of interest. The more connected the materials are, the less learners might benefit from intermediate practice tests.

However, it should be noted that, for the coherent text used in Experiment 1, we did not find any disadvantage of test-taking either. Thus, in terms of efficiency and efficacy, our results at least indicate that learners might have little to lose when allocating time to retrieval practice rather than using all available time for study. Furthermore, in contrast to our results, some other studies investigating the effect of retrieval practice on the retention of text materials suggest that learners have much to gain from practicing retrieval during learning (e.g., Hinze & Wiley, 2011; Karpicke & Blunt, 2011; Nungester & Duchastel, 1982; Roediger & Karpicke, 2006b). Thus, notwithstanding our own findings, it might still be preferred to practice retrieval during learning even for complex materials like science discourse. Also, it should be noted that, the results from our study are limited to the test format used (a fill-in-the blank test). Using more demanding test formats like free recall (e.g., Roediger & Karpicke, 2006b) or providing learners with feedback during testing (e.g., Kang et al., 2007) might be more potent ways of improving the learning and retention of complex text materials compared to the test used in our study.

Final remarks

In the introduction to this thesis it was tongue-in-cheekishly argued that, in some respects, learning might rightfully regarded as a complete waste of time. However, in all seriousness, learning is a quintessential part of life and, in most respects, learning can be a very valuable and rewarding endeavor. Thus, strategies that can be used for improving the efficiency and efficacy of learning should be of interest to many a person. We investigated two strategies that hold potential for enhancing not only the initial learning, but also the subsequent retention of information. The findings in the present thesis suggest that learners can profit greatly from the right set of conditions. We found that, when time to learn is limited, it matters a great deal how one puts to use the available amount of time. First of all, the pacing of study trials during learning can have a profound effect on both immediate and delayed recall performance. Based on our findings we recommend using a self-paced study procedure or, if circumstances demand so, a fixed presentation rate of around 4 s for the learning of simple verbal materials (e.g., foreign language vocabulary). Secondly, we found that allocating some portion of the available amount of total study time to be used for retrieval practice can have a substantial positive effect on long-term retention. In line with other recent findings (e.g., Karpicke & Roediger, 2007, 2008), our results suggest that, even when an item is already learned to the degree that it can be successfully retrieved from long-term memory, continued retrieval practice is still recommended. Importantly, extending previous findings, we found that the benefits of repeated retrieval practice are not just limited to enhanced recall of a small subset of items on a delayed test. Repeated testing during initial learning can also be a very effective strategy for enhancing the delayed relearning of all items in a set. We conclude with an Ebbinghausian word of comfort for those learners who often feel the fruit of their work falls short of their efforts. Even when performance on a delayed test suggests that the original learning of information might have been a waste of time in terms of recall, the original learning might still have been beneficial in the sense that it can enhance the delayed reinstatement of what was learned, but then forgotten.

Nederlandse Samenvatting

Summary in Dutch

Dit proefschrift gaat over de situatie waarin studenten een beperkte hoeveelheid tijd tot hun beschikking hebben om zelfstandig een bepaalde hoeveelheid informatie te leren. De centrale vraag was hoe studenten het maximale uit deze tijd zouden kunnen halen. Hierbij waren we vooral geïnteresseerd in leerstrategieën die het initiële leren efficiënter zouden kunnen maken. Daarnaast hebben we ook de effectiviteit van het leren met betrekking tot het onthouden van informatie over een langere periode bekeken. Het eerste gedeelte van dit proefschrift richtte zich op het effect van studietijdverdeling en het tweede gedeelte richtte zich op de mogelijke voordelen van *retrieval practice* (oefenen door het ophalen van informatie uit het geheugen). Hieronder worden de belangrijkste resultaten en conclusies van dit proefschrift besproken.

Studietijdverdeling

In het eerste gedeelte van dit proefschrift werd onderzocht hoe men in een enkele (korte) leersessie de studietijd het beste kan verdelen over de te bestuderen items. Ten eerste vroegen wij ons af in hoeveel aanbiedingen de beschikbare tijd per item zou moeten worden opgedeeld, gegeven dat er een vaste studietijd is. Ten tweede vroegen wij ons af of de optimale verdeling van studietijd wellicht afhankelijk is van de moeilijkheid van het te leren materiaal. Tot slot waren we geïnteresseerd in hoe effectief de beschikbare studietijd wordt gebruikt wanneer de lerende zelf het aanbiedingstempo kan bepalen in een leersessie.

In **Hoofdstuk 2** bestudeerden proefpersonen ongerelateerde woordparen onder verschillende aanbiedingstempo's, variërend van relatief snel (bijvoorbeeld 16 aanbiedingen van 1 seconde per woordpaar) tot relatief langzaam (een enkele aanbieding van 16 seconden per woordpaar). De totale hoeveelheid studietijd per item werd constant gehouden. We vonden een niet-monotone relatie tussen het aanbiedingstempo en de uiteindelijke prestatie op een geheugentest. Snelle (1 seconde per woordpaar) en langzame (16 seconden per woordpaar) aanbiedingstempo's resulteerden in een slechte prestatie op een latere geheugentest (na 5 minuten en na 2 dagen) vergeleken met de tussenliggende aanbiedingstempo's (zoals 4 seconden per woordpaar). De conclusie was dat er een zogenaamd 'Goldilocks'-principe (naar het bekende sprookje van Goudlokje) lijkt te zijn wat betreft aanbiedingstempo: Het aanbiedingstempo moet niet te lang, niet te kort, maar precies goed zijn.

In **Hoofdstuk 3** werden de bevindingen uit Hoofdstuk 2 gerepliceerd met meer ecologisch valide studiemateriaal. Proefpersonen (studenten uit de VS) bestudeerden Engels-Nederlands woordparen met verschillende aanbiedingstempo's. Net als in onze eerdere studie vonden we een niet-monotone relatie tussen aanbiedingstempo en de latere prestatie op een geheugentest. De snelle (1 seconde) en de langzame (16 seconden) aanbiedingstempo's resulteerden wederom in minder goede prestaties op een latere geheugentest (na 5 minuten en na 2 dagen) vergeleken met tussenliggende (4 seconden) aanbiedingstempo's. Nog belangrijker, de resultaten vertoonden een "Goldilocks"-patroon ongeacht de vertaalrichting van de woordparen. Voor beide vertaalrichtingen (Nederlands → Engels en Engels → Nederlands) vonden we dat de geheugenprestatie van proefpersonen beter was voor de tussenliggende aanbiedingstempo's vergeleken met snelle en langzame aanbiedingstempo's. Deze resultaten suggereren dat het aanbiedingstempo een belangrijke factor kan zijn bij het leren van materiaal dat relevant is voor de onderwijspraktijk, zoals woordjes in een nieuwe taal. Bovendien lijkt het er op dat het optimale aanbiedingstempo niet verschuift naarmate de taak lastiger wordt. De algehele geheugenprestatie was weliswaar minder goed in de Engels → Nederlands vertaalrichting dan in de Nederlands → Engels vertaalrichting, maar het optimale aanbiedingstempo was vergelijkbaar voor beide vertaalrichtingen.

Tot slot werden in **Hoofdstuk 4** verschillende vaste aanbiedingstempo's vergeleken met een conditie waarin proefpersonen zelf het tempo mochten bepalen (een *self-paced* conditie). Ten eerste vonden we dat het tempo in de self-paced conditie niet constant was over de verschillende studiecycli. In de self-paced conditie hadden proefpersonen sterk de neiging om het tempo te verhogen naarmate het leerproces vorderde. De eerste keer dat ze door de lijst met woordparen gingen, hielden de proefpersonen een redelijk langzaam tempo aan, maar de laatste keer dat ze door de lijst gingen, werd er een relatief hoog tempo gehanteerd.

Voor de condities met een vaststaand aanbiedingstempo repliceerden we wederom het 'Goldilocks' patroon. Tussenliggende aanbiedingstempo's resulteerden in een betere prestatie op de geheugentaak vergeleken met de meer extreme aanbiedingstempo's. Nog belangrijker, we vonden dat self-paced studeren over het algemeen resulteerde in betere prestaties op een latere geheugentaak vergeleken met condities met een vaststaand aanbiedingstempo. De enige uitzondering was de 12 × 2 s conditie, waarin de geheugenprestatie vergelijkbaar was met die in de self-paced conditie. In Experiment 2 van Hoofdstuk 4 onderzochten we een verklaring voor het voordeel van de self-paced conditie over de vaste aanbiedingstempo's. De resultaten van Experiment 2 ondersteunden het idee dat de mogelijkheid om meer of minder studietijd toe te wijzen aan verschillende items tijdens het self-paced studeren cruciaal kan zijn voor een betere prestatie op een latere geheugentest. Als het de lerende namelijk niet was toegestaan om de totale hoeveelheid studietijd per item te variëren, dan verslechterde de prestatie op een latere geheugentest. We concludeerden dat studenten over het algemeen vrij bekwaam zijn in het toewijzen van studietijd aan items tijdens herhaald bestuderen in een enkele leersessie.

Samenvattend, de resultaten in Hoofdstuk 2, 3, en 4 zijn eenduidig wat betreft het effect van aanbiedingstempo op het leren van woordparen. In alle drie de studies vonden we aanzienlijke effecten van aanbiedingstempo op latere geheugenprestaties, waarbij de tussenliggende aanbiedingstempo's tot de beste prestaties leidden. Het effect van aanbiedingstempo lijkt erg robuust te zijn en generaliseert over verschillende typen simpel verbaal materiaal, variërend van onzinsyllaben gepaard met cijfers (Calfee & Anderson, 1971; Johnson, 1964) tot meer betekenisvol materiaal zoals ongerelateerde woordparen (de Jonge, Tabbers, Pecher, & Zeelenberg, 2012), en het leren van woordjes in een nieuwe taal (de Jonge, Tabbers, Pecher, Jang, & Zeelenberg, 2013; Zeelenberg, de Jonge, Tabbers, & Pecher, 2013). Op basis van onze bevindingen concludeerden we dat, voor het leren van woordjes in alfabetische talen, een constant aanbiedingstempo van rond de 4 seconden een goede vuistregel lijkt te zijn. De resultaten in Hoofdstuk 4 suggereren echter dat een self-paced procedure wellicht nog efficiënter is voor het leren van woordjes in een vreemde taal tijdens een enkele leersessie. Eén noemenswaardige beperking van onze studie in Hoofdstuk 4 was echter dat we niet naar de langetermijneffectiviteit van self-paced studeren hebben gekeken. Dus hoewel onze resultaten uitwijzen dat self-paced studeren efficiënt kan zijn, blijft het de vraag of zelf het tempo bepalen ook de voorkeur verdient wanneer het om langetermijnretentie gaat. De effectiviteit van self-paced leren op de lange termijn is nog niet onderzocht en dit punt moet in toekomstig onderzoek worden behandeld.

Retrieval Practice

Het eerste gedeelte van dit proefschrift ging over de situatie waarin de beschikbare hoeveelheid tijd exclusief werd gebruikt voor het herhaald bestuderen van informatie. In het tweede gedeelte van dit proefschrift werd onderzocht hoe effectief het is om naast studietrials ook nog testtrials aan te bieden tijdens een enkele leersessie. Eerdere studies suggereren dat tijd gespendeerd aan het ophalen van eerder geleerde informatie over het algemeen goed besteed is, zelfs wanneer er hierdoor minder tijd overblijft voor verdere studie (Nungester & Duchastel, 1989). In onze studies naar dit zogeheten *retrieval practice effect* werden condities waarin een gedeelte van de beschikbare tijd gereserveerd was voor *retrieval practice* (door te toetsen) vergeleken met (controle-)condities waarin alle beschikbare tijd gebruikt werd voor het herhaald bestuderen van informatie. We onderzochten het effect van retrieval practice op de langetermijnretentie van simpel verbaal materiaal, en op de langetermijnretentie van complexer tekstmateriaal. In onze studies werden geheugenprestaties gemeten na zowel een kort interval (in de orde van minuten) als een langer interval (in de orde van dagen of weken). Dit maakte het mogelijk om de mate van het vergeten te onderzoeken in de verschillende leercondities. Tot slot hebben we ook onderzocht wat voor effect retrieval practice tijdens een eerste leerfase heeft op hoe snel vergeten informatie opnieuw geleerd kan worden tijdens een latere herleerfase.

In **Hoofdstuk 5** onderzochten we de mogelijke voordelen die retrieval practice kan bieden bij het leren van woordjes in een vreemde taal. In twee experimenten vergeleken we de mate van vergeten voor een conditie waarin woordparen herhaaldelijk werden getest met een (controle-)conditie waarin woordparen herhaaldelijk werden bestudeerd. De mate van vergeten werd bekeken over een interval van 1 en 4 weken. In Experiment 1 vonden we dat er in de leercondities met tests relatief minder snel werd vergeten over een interval van 1 week dan in de conditie waarin herhaald werd gestudeerd. In Experiment 2 werd deze bevinding gerepliceerd en we vonden dat de vergeetfuncties elkaar na 4 weken zelfs kruisten. Vier weken na het initiële leren was er dus een substantieel voordeel van testen ten opzichte van herhaald studeren. De resultaten van ons onderzoek demonstreren eens te meer dat testen een sterk effect kan hebben op de langetermijnretentie van informatie.

In Hoofdstuk 5 werden tevens twee belangrijke kwesties besproken die tot nog toe weinig aandacht hebben gekregen in eerder onderzoek naar het effect van testen. Op de eerste plaats zou het zo kunnen zijn dat het herhaald testen slechts resulteert in het versterken en goed leren van een klein gedeelte van relatief makkelijke (succesvol uit het geheugen opghaalde) items (zie ook Thompson, Wenger, & Bartling, 1978). Het langetermijnvoordeel voor herhaaldelijk geteste items zou dus wellicht verklaard kunnen worden als een itemselectie-effect. Als het voordeel van testen niet opgaat voor de moeilijkere items in een lijst en beperkt is tot de makkelijkere items dan zullen de voordelen voor onderwijsdoeleinden ook beperkt zijn. Ten tweede merkten we op dat eerdere studies naar het effect van testen vooral hebben gekeken naar geheugenprestaties (aantal goed op een latere test) en we pleitten ervoor om ook naar andere retentiematen te kijken die wellicht relevanter zijn voor de praktijk (bijvoorbeeld hoelang het duurt om iets later weer opnieuw te leren). Deze twee kwesties vormden de basis voor Hoofdstuk 6 en 7.

In **Hoofdstuk 6** hebben we het effect van herhaald testen op itemselectie, retentie, en herleren onderzocht. In ons onderzoek leerden proefpersonen gemengde lijsten met makkelijke (gerelateerde) en moeilijke (ongerelateerde) woordparen. In de ene conditie werden de items herhaaldelijk getest en in de andere conditie werden de items herhaaldelijk bestudeerd. Een week na het initiële leren keerden de proefpersonen terug voor een geheugentest en een herleerfase. Tijdens de initiële leerfase van het experiment vonden we dat de geheugenprestatie op de oefentests beter was voor de gerelateerde woordparen dan voor de ongerelateerde woordparen (oftewel, er was sprake van itemselectie op de oefentests). Op de geheugentest na een week vonden wij echter een langetermijnvoordeel van herhaald testen, ongeacht de moeilijkheid van de items. Voor zowel makkelijke (gerelateerde) als moeilijke (ongerelateerde) woordparen presteerde de conditie waarin herhaaldelijk werd getest beter dan de conditie waarin herhaaldelijk werd gestudeerd. Deze resultaten suggereren dat het retentievoordeel van herhaald testen niet alleen het resultaat is van itemselectie. In de herleerfase die volgde na de geheugentest, vonden we verder dat het leren op zich sneller verliep voor items die eerder herhaald bestudeerd waren dan voor items die eerder herhaald getest waren. Na een enkele herstudeercyclus was het initiële voordeel van herhaald testen zelfs verdwenen. Maar beide condities presteerden vervolgens vergelijkbaar op de resterende tests in de herleerfase. Onze resultaten laten zien dat het herhaald getest worden tijden het initiële leren een positief effect heeft op latere geheugenprestaties en dat het geen substantiële nadelen oplevert voor het later opnieuw leren van informatie.

Hoofdstuk 6 was gericht op de situatie waar slechts een gedeelte van de informatie werd opgeslagen en herhaaldelijk werd opgehaald tijdens een initiële leerfase. In **Hoofdstuk 7** hebben we gebruik gemaakt van een leren-tot-een-criterium-procedure, om het effect van herhaald testen op herleren verder te onderzoeken. Alle items werden eerst geleerd tot het criterium van één keer succesvol ophalen uit het langetermijngeheugen en vervolgens kregen alle items nog drie (*post-retrieval*) studie- of testtrials. Een week na het initiële leren keerden de proefpersonen terug voor een geheugentest en een herleerfase. Post-retrieval testtrials resulteerden in betere prestaties op de geheugentest dan post-retrieval studietrials. Ook vonden we dat de items van beiden experimentele condities (post-retrieval studie en post-retrieval test) sneller opnieuw geleerd werden dan een nieuwe set items die niet eerder was bestudeerd. Nog belangrijker, we vonden dat items het snelst opnieuw geleerd werden wanneer ze in eerste instantie geleerd waren met testtrials dan met studietrials. Deze resultaten laten zien dat de voordelen van herhaald testen niet beperkt zijn tot het bewust ophalen uit het geheugen op een latere test. Herhaald testen kan ook een erg effectieve strategie zijn om herleren te faciliteren.

Tot slot werd in **Hoofdstuk 8** beargumenteerd dat het positieve effect van retrieval practice wellicht minder robuust is voor het leren van teksten. Eerdere studies naar het effect van testen op het leren van teksten waarin toetsvormen gebruikt werden die relevant zijn voor het onderwijs (zoals korte antwoordvragen), hebben nogal wisselvallige resultaten opgeleverd (Hinze & Wiley, 2011; Kang, McDermott, & Roediger, 2007; LaPorte & Voss, 1975; Nungester & Duchastel, 1982). Interessant is dat in heel vroege studies naar het retrieval practice effect al werd opgemerkt dat het voordeel van testen opvallend minder tot uitdrukking lijkt te komen, naarmate het geteste materiaal betekenisvoller wordt (Gates, 1917; Kühn, 1914). In de twee experimenten in Hoofdstuk 8 onderzochten we het effect van retrieval practice met een *fill-in-the-blank* test op de langetermijnretentie van een complexe tekst. In Experiment 1 vonden we geen voordeel van retrieval practice ten opzichte van een herstudieconditie voor het leren van een samenhangende tekst. Beide condities resulteerden in vergelijkbare prestaties op een geheugentest na 5 minuten en 1 week na het leren. In Experiment 2 werd de samenhang van de tekst verstoord. De zinnen in de tekst werden door elkaar gehusseld en er werd aan de proefpersonen verteld dat zij een lijst met feitjes zouden gaan leren. Voor de lijst met feitjes vonden we dat er, gedurende een interval van 1 week, minder werd vergeten in de retrieval practice conditie dan in de herstudieconditie. Samengenomen suggereren de resultaten van Experiment 1 en 2 dat het langetermijnvoordeel van testen afhankelijk zou kunnen zijn van de samenhang van het te leren materiaal. Retrieval practice met een fill-in-the-blank test lijkt een groter voordeel op te leveren, naarmate het te bestuderen materiaal minder samenhangend is.

Samenvattend, onze studies naar het retrieval practice effect laten zien dat testen tijdens het leren de langetermijnretentie van simpel verbaal materiaal kan verbeteren. Ten eerste repliceerden we in Hoofdstuk 5, 6, en 7 de algemene bevinding dat testen een substantieel effect kan hebben op een latere geheugentest. Ten tweede suggereren de resultaten in Hoofdstuk 6 dat het effect van retrieval practice niet zozeer afhankelijk is van de moeilijkheid van het te leren materiaal. Ten derde vonden we in Hoofdstuk 7 dat het voordeel van retrieval practice niet beperkt is tot een verbeterede prestatie op een latere geheugentest, maar dat retrieval practice ook het later herleren van informatie kan faciliteren. Kortom, onze studies laten zien dat het effect van retrieval practice robuust is voor simpel verbaal materiaal. Interessant is echter dat de resultaten in Hoofdstuk 8 suggereren dat het voordeel van retrieval practice minder robuust is voor complexer materiaal zoals een samenhangende tekst. Deze bevinding suggereert dat de praktische voordelen van testen voor de praktijk wellicht beperkt zijn.

References

Anderson, J. R. (1995). Learning and memory: An integrated approach. New York: Wiley.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought.* Mahwah, NJ: Erlbaum.

Archer, E. J. (1960). A re-evaluation of the meaningfulness of all possible CVC trigrams. *Psychological Monographs, 74*, 1-23.

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

Averell, L., & Heathcote, A. (2011). The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology, 55*, 25-35.

Bahrick, H. P. (1967). Relearning and the measurement of retention. *Journal of Verbal Learning and Verbal Behavior*, *6*, 89-94.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experiment Psychology: General, 108*, 296-308.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*, 379-384.

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.

Begg, I. M., Martin, L. A., & Needham, D. R. (1992). Memory monitoring: How useful is self-knowledge about memory? *European Journal of Cognitive Psychology, 4*, 195-218.

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology, 61*, 228-247.

Bjork, R. A. (1988). Retrieval practice and the maintenance of knowledge. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues. Vol. 1: Memory in everyday life* (pp. 396–401). New York: Wiley.

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

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

Britton, B. K., & Gülgöz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology, 83*, 329-345.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 977-990.

Bugelski, B. R. (1962). Presentation time, total time, and mediation in paired-associate learning. *Journal of Experimental Psychology, 63*, 409-412.

Bugelski, B. R., Kidd, E., & Segman, J. (1968). Image as a mediator in one-trial paired-associate learning. *Journal of Experimental Psychology, 76*, 69-73.

Bugelski, B. R., & McMahon, M. L. (1971). The total time hypothesis: A reply to Stubin, Heimer, and Tatz. *Journal of Experimental Psychology, 90*, 165-166.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1118-1133.

Butler, A. C., Karpicke, J. D., & Roediger III, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 918-928.

Calfee, R. C. (1968). Interpresentation effects in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior, 7*, 1030-1036.

Calfee, R. C., & Anderson, R. (1971). Presentation rate effects in paired-associate learning. *Journal of Experimental Psychology, 88*, 239-245.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 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 and Cognition, 34*, 268-276.

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., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin and Review, 13*, 826-830.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory and Cognition, 36*, 438-448.

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

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

Congleton, A. R., & Rajaram, S. (2011). The influence of learning methods on collaboration: Prior repeated retrieval enhances retrieval organization, abolishes collaborative inhibition, and promotes post-collaborative memory. *Journal of Experimental Psychology: General, 140*, 535-551.

Cooper, E. H., & Pantle, A. J. (1967). The total-time hypothesis in verbal learning. *Psychological Bulletin, 68*, 221-234.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268-294.

Cull, W. L., d'Anna, C. A., Hill, E. J., Zechmeister, E. B., & Hall, J. W. (1991). When are optimal rates of presentation optimal (for learning)? *Bulletin of the Psychonomic Society, 29*, 48-50.

de Deyne, S., & Storms, G. (2008). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods, 40*, 198-205.

de Jonge, M. & Tabbers, H. K. (2013). Repeated testing, item selection, and relearning: The benefits of testing outweigh the costs. *Experimental Psychology, 60,* 206-212.

de Jonge, M., Tabbers, H. K., Pecher, D., Jang, Y., & Zeelenberg, R. (2013). *The efficacy of self-paced study in multitrial learning*. Manuscript submitted for publication.

de Jonge, M., Tabbers, H.K., Pecher, D., & Zeelenberg, R. (2012). The effect of study time distribution on learning and retention: A Goldilocks principle for presentation rate. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 405-412.

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation, 53,* 63-147.

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist, 43*, 627-634.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Human memory* (pp. 197–236). San Diego, CA: Academic Press.

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

Dunlosky, J., & Hertzog, C. (1998). Training programs to improve learning in later adulthood: Helping older adults educate themselves. In D. J. Hacker, J. Dunlosky & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 249-275). Mahwah, NJ: Erlbaum.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). New York: Dover. (Original work published 1885)

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25*, 285-307.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*, 1-104.

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

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

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.

Greeno, J. G. (1964). Paired-associate learning with massed and distributed repetitions of items. *Journal of Experimental Psychology, 67*, 286-295.

Grimaldi, P. J., Pyc, M. A., & Rawson, K. A. (2010). Normative multitrial recall performance, metacognitive judgments, and retrieval latencies for Lithuanian-English paired associates. *Behavior Research Methods, 42*, 634-642.

Groot, P. J. M. (2000). Computer assisted second language vocabulary acquisition. *Language Learning and Technology, 4*, 60-81.

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 801-812.

Heim, A. W., Watts, K. P., Bower, I. B., & Hawton, K. E. (1966). Learning and retention of word-pairs with varying degrees of association. *The Quarterly Journal of Experimental Psychology, 18*, 193-205.

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using short answer tests. *Memory, 19*, 290-304.

Izawa, C. (1966). Reinforcement-test sequences in paired-associate learning. *Psychological Reports, 18,* 879 –919.

Jang, Y., Wixted, J. T., Pecher, P., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability. *The Quarterly Journal of Experimental Psychology, 65*, 962-975.

Johnson, N. F. (1964). The functional relationship between amount learned and frequency vs. rate vs. total time of exposure of verbal materials. *Journal of Verbal Learning and Verbal Behavior, 3*, 502-504.

Jönsson, F. U., Hedner, M., & Olsson, M. J. (2012). The testing effect as a function of explicit testing instructions and judgments of learning. *Experimental Psychology, 59*, 251-257.

Kang, S. H. K., McDermott, K. B., & Roediger III, H. L. (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. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*, 469-486.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*, 772-775.

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

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

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67*, 17-29.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for dutch word frequency based on film subtitles. *Behavior Research Methods, 42*, 643-650.

Kimball, D. R., & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory and Cognition, 31*, 918-929.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49*, 294-303.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135*, 36-69.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin and Review, 14*, 219-224.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138*, 449-468.

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

Kühn, A. (1914). Über Einprägung durch Lesen und durch Rezitieren. *Zeitschrift für Psychologie, 68*, 396-481.

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

LaPorte, R. E., & Voss, J. F. (1975). Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology, 67*, 259-266.

Lewis, J. R. (1989). Pairs of Latin squares to counterbalance sequential effects and pairing of conditions and stimuli. In: *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 1223-1227).

Lockhart, R. S. (2002). Levels of processing, transfer-appropriate processing, and the concept of robust encoding. *Memory, 10*, 397-403.

Loftus, G. R. (1985). Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11, 397*-406.

Lorch, R. F., & O'Brien, E. J. (Eds.) (1995). *Sources of coherence in reading*. Hillsdale, NJ: Erlbaum.

MacLeod, C. M. (1988). Forgotten but not gone: Savings for pictures and words in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 195-212.

MacLeod, C. M., & Nelson, T. O. (1984). Response latency and response accuracy as measures of memory. *Acta Psychologica, 57*, 215-235.

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.

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective? *Journal of Experimental Psychology: General, 122*, 47-60.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory and Cognition, 18*, 196-204.

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., Einstein, G. O., Dunay, P. K., & Cobb, R. E. (1986). Encoding difficulty and memory: Toward a unifying theory. *Journal of Memory and Language, 25*, 645-656.

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

McDaniel, M. A., Roediger III, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin and Review, 14*, 200-206.

McGuire, W. J. (1961). A multiprocess model for paired-associate learning. *Journal of Experimental Psychology, 62*, 335-347.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55*, 51-62.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes, 22*, 247-288.

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

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

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. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*, 463-477.

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

Murdock, B. B. (1960). The immediate retention of unrelated words. *Journal of Experimental Psychology, 60*, 222-234.

Naveh-Benjamin, M., & Kilb, A. (2012). How the measurement of memory processes can affect memory performance: The case of remember/know judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 194-203.

Nelson, T. O. (1971). Savings and forgetting from long-term memory. *Journal of Verbal Learning and Verbal Behavior, 10*, 568-576.

Nelson, T. O. (1978). Detecting small amounts of information in memory: Savings for nonrecognized items. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 453-468.

Nelson, T. O. (1985). Ebbinghaus's contribution to the measurement of retention: Savings during relearning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 472-479.

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

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

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied, 14*, 101-117.

Payne, D. G. (1987). Hypermnesia and reminiscence in recall: A historical and empirical review. *Psychological Bulletin, 101*, 5-27.

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

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.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*, 335.

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: Application of the SAM model. *Cognitive Science, 27*, 431-452.

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

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology, 97*, 70-80.

Roediger III, H. L. (2013). Applying cognitive psychology to education: Translational educational science. *Psychological Science in the Public Interest, Supplement, 14*, 1-3.

Roediger, III, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17*, 382-395.

Roediger III, H. L., Balota, D. A., & Watson, J. M. (2001). Spreading activation and arousal of false memories. In H. L. Roediger, J. S. Nairne, I. Neath & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 95–115). Washington, DC: American Psychological Association.

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

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

Roediger III, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition, 1*, 242-248.

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.

Schneider, V. I., Healy, A. F., & Bourne, L. E. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention and transfer. *Journal of Memory and Language, 46,* 419-440.

Shimmerlik, S. M. (1978). Organization theory and memory for prose: A review of the literature. *Review of Educational Research, 48*, 103-120.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 204-221.

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science, 3*, 315–316.

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

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

Stubin, E. J., Heimer, W. I., & Tatz, S. J. (1970). Total time and presentation time in paired-associate learning. *Journal of Experimental Psychology, 84*, 308-310.

Thiede, K. W. (1999). The importance of monitoring and self-regulation during multitrial learning. *Psychonomic Bulletin and Review, 6*, 662-667.

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., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology, 56*, 252-257.

Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language, 64*, 109-118.

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

Underwood, B. J., Rehula, R., & Keppel, G. (1962). Item-selection in paired-associate learning. *American Journal of Psychology, 75*, 353-371.

van Rijn, H., van Maanen, L., & van Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. In *Proceedings of the 9th International Conference on Cognitive Modeling*.

Wenger, S. K., Thompson, C. P., & Bartling, C. A. (1980). Recall facilitates subsequent recognition. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 135-144.

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

Wickelgren, W. A. (1974). Single-trace fragility theory of memory dynamics. *Memory and Cognition, 2*, 775-780.

Wixted, J. T. (2004). On common ground: Jost's (1897) law of forgetting and Ribot's (1881) law of retrograde amnesia. *Psychological Review, 111*, 864-879.

Wixted, J. T., & Carpenter, S. K. (2007). The Wickelgren power law and the Ebbinghaus savings function. *Psychological Science, 18*, 133-134.

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science, 2*, 409-415.

Zaromb, F. M., & Roediger III, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory and Cognition, 38*, 995-1008.

Zeelenberg, R., de Jonge, M., Tabbers, H. K., & Pecher, D. (2013). *The effect of presentation rate on foreign language vocabulary learning*. Manuscript submitted for publication.

Dankwoord

Acknowledgements in Dutch

Aan de totstandkoming van dit proefschrift hebben velen, zij het direct dan wel indirect, een bijdrage geleverd. Een aantal mensen in het bijzonder wil ik hiervoor hartelijk danken.

Om te beginnen wil ik mijn promotor, prof. dr. Rikers, en copromotor, dr. Tabbers bedanken. Remy, hartelijk dank voor alle goede gesprekken over onderzoek, onderwijs, en niet te vergeten het vaderschap (soms met enigszins bloeddoorlopen oogjes vanwege een korte nacht). Ook wil ik je in het bijzonder bedanken voor je frisse blik en het meedenken over de structuur van het proefschrift. Je dwong me om ook goed naar het grotere plaatje te blijven kijken en vooral bij de afronding van mijn proefschrift was jouw inhoudelijke feedback vaak doorslaggevend.

Huib, op de eerste plaats hartelijk dank voor de begeleiding, alle goede gesprekken, en je betrokkenheid. Je zorgde er met je kritische blik altijd voor dat ik net nog wat langer en beter over de dingen nadacht en daar heb ik veel van geleerd. Daarnaast wil ik je danken voor je aanstekelijke enthousiasme en ongebreidelde optimisme. Je bent zonder twijfel een van de best gehumeurde personen die ik ken. Je grapte tijdens mijn project weleens dat het er eigenlijk wel bij hoort om als promovendus op enig moment in een soort van crisis te belanden. Dat ik een dergelijke ervaring heb moeten missen, is zeker voor een groot deel aan jou te wijten. Ik heb tijdens mijn promotie met veel plezier met je samengewerkt en ik ben enorm blij dat onze samenwerking nu ook nog eens een vervolg heeft gekregen.

Verder gaat mijn dank uit naar dr. Zeelenberg en dr. Pecher. René, Diane, hartelijk dank voor de plezierige en vruchtbare samenwerking door de jaren heen. Jullie waren bij een groot gedeelte van het onderzoek betrokken en jullie hebben dan ook een belangrijke bijdrage geleverd aan de totstandkoming van dit proefschrift. In barre tijden waren jullie, bij wijze van spreken, mijn baken in het met valkuilen bezaaide mijnenveld dat metacognitie heet (heb ik deze parel van een zin toch nog in het proefschrift gekregen!). Eén van de absolute hoogtepunten tijdens mijn promotietraject was zonder twijfel dat ik met jullie mee mocht naar San Diego om daar onderzoek te doen (de officiële lezing) en een hele reeks aan speciale biertjes te drinken (de officieuze lezing). Ik denk nog regelmatig met veel plezier terug aan deze reis en zal het hopelijk nooit vergeten.

De leden van de kleine commissie, prof.dr. Tamara van Gog, prof.dr. Liesbeth Kester, en dr. Peter Verkoeijen dank ik voor het kritisch lezen en het beoordelen van het proefschrift en het deelnemen aan de oppositie. Tevens dank ik de leden van de grote commissie, prof.dr. Jeroen Raaijmakers, en prof.dr. Rolf Zwaan voor hun bereidheid met mij van gedachten te wisselen over het proefschrift.

Alle collega's van Cognition and Learning, Memory Lab, en de O&O groep (Samantha, Noortje, Gerdien, Jan, Nicole, Gabriela, Lisa, Kiki, Bruno, Tim, Michel, Jim, Charly, Tamara, Sofie, Martine, Gino, Anique, Vincent, Jacqueline, Kim, Lysanne, Wim, Lydia, Lisette, et al.) bedankt voor de discussies, de feedback, en de gezelligheid. In het bijzonder wil ik iedereen van de O&O pub-groep bedanken voor het kritisch lezen en reviewen van enkele vroege versies van hoofdstukken die later in aangepaste vorm in dit proefschrift zouden komen. Een aantal van deze manuscripten zijn inmiddels ook geaccepteerd voor publicatie in een wetenschappelijk tijdschrift en ik ben er van overtuigd dat al jullie constructieve op- en aanmerkingen mijn kansen in het publicatiecasino aanzienlijk hebben vergroot.

Verder dank ik alle collega's met wie ik door de jaren heen een kamer heb gedeeld. Anita, Maartje, Marianne, Marien, Ali, Anna, en mijn nieuwe roomie, Kimberley, bedankt voor de gezelligheid.

Het ondersteunend personeel van het Erasmus Behavioral Lab dank ik voor alle hulp, tips, en uitleg. Daarnaast dank ik de goede mensen van het secretariaat voor alle hulp en ondersteuning.

Het Erasmus Trustfonds wil ik bedanken voor de medefinanciering van verschillende congresbezoeken en mijn bezoek aan UCSD.

I thank the Department of Psychology at UCSD. In particular I thank Dave Huber for his hospitality and allowing us to test participants in his laboratory. Also, I thank Yoonhee Jang for her valuable contribution to our study on multitrial self-paced learning (Chapter 4).

Mijn paranimfen en goede vrienden, Rutger Balvers en Nick Verhoeven. Heren, hartelijk dank voor de rugdekking. Verder dank ik Tim Pelgrim. We begonnen ooit gelijktijdig aan onze masterstage bij Rene en Diane (onderzoek naar het effect van testen). De weledele kunst van het programmeren in Eprime heb ik voor een groot gedeelte van jou af mogen kijken.

Ook wil ik mijn familie bedanken. Allereerst mijn ouders, Peter en Addi, bedankt voor alle steun en liefde. Jullie zijn geweldig. Bedankt ook voor het om de week oppassen, Felice en Max zijn dol op hun opa en oma. Guido en Suzan, jullie zijn ook fantastisch. En dan niet alleen als broer en zus maar ook als oom en tante.

Tot slot, mijn gezinnetje. Lieve Jolan, ik hou van je en ben heel blij dat ik met je samen mag zijn. Bedankt voor alle steun en de vrijheid die je me hebt gegeven (als ik bijvoorbeeld een paar weken naar San Diego wilde gaan). Felice en Max, jullie hebben mijn leven veranderd. Dankzij jullie begint elke dag met een vrolijke noot.

Curriculum Vitae and Publications

Curriculum Vitae

Mario de Jonge was born in Vlissingen, the Netherlands, on January 7th, 1980. He completed secondary education in 1999 at the Sint-Laurenscollege in Rotterdam. In 2002, he started studying Psychology at the Erasmus University Rotterdam. He obtained his Bachelor's degree in 2007 and his Master's degree in 2008. In 2009 he started working as a Ph.D. student at the department of Psychology, Erasmus University Rotterdam, of which the present thesis is the result. As a Ph.D. student, he was engaged in teaching a variety of practical and theoretical courses, mainly in methodology and statistics and he supervised several research projects of Bachelor students.

Publications

de Jonge, M., Tabbers, H. K., Pecher, D., & Zeelenberg, R. (2012). The effect of study time distribution on learning and retention: A Goldilocks principle for presentation rate. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 405-412.

de Jonge, M., & Tabbers, H. K. (2013). Repeated testing, item selection, and relearning: The benefits of testing outweigh the costs. *Experimental Psychology, 60*, 206-212.

de Jonge, M., Tabbers, H. K., & Rikers, R. M. J. P. (2014). Retention beyond the threshold: Test-enhanced relearning of forgotten information. *Journal of Cognitive Psychology, 26*, 58-64.

Submitted manuscripts

de Jonge, M., Tabbers, H. K., Pecher, D., Jang, Y., & Zeelenberg, R. (2014). *The efficacy of self-paced study in multitrial learning*. Manuscript submitted for publication.

Zeelenberg, R., de Jonge, M., Tabbers, H. K., & Pecher, D. (2014). *The effect of presentation rate on foreign language vocabulary learning*. Manuscript submitted for publication.