The Influence of Sleep-Associated Memory Consolidation on Beneficial Effects of Retrieval Practice

Inaugural-Dissertation zur Erlangung der Doktorwürde
der Philosophischen Fakultät II
(Psychologie, Pädagogik und Sportwissenschaft)
der Universität Regensburg

vorgelegt von

Christoph Holterman

aus München

Regensburg 2016

Erstgutachter: Prof. Dr. Karl-Heinz T. Bäuml
Zweitgutachter: Prof. Dr. Klaus W. Lange

Acknowledgements

The present work would not have been possible without all the support I received over the years.

First I want to thank Prof. Karl-Heinz Bäuml who initiated this project. His scientific experience and advise have been essential for planning the investigations, analyzing the results and summarizing them in this thesis. His guidance and ideas have always been extraordinarily helpful on the way.

Moreover, I particularly want to thank my colleagues Lena Abel, Alp Aslan, Ina Dobler, Margit Frimberger, Oliver Kliegl, Bernhard Pastötter, Julia Rupprecht, Andreas Schlichting, and Petra Witzmann. You were a great company and always there for me whenever I needed your help.

I am also very thankful that Valerie Haller, Anna Karl, Carla Nottberg, David Schnell, and Franziska Welker helped me with data collection. Without you this project would have never been possible.

Last but not least I want to thank my family and friends for their unlimited support. I owe to my parents all the possibilities I have in my life and am very grateful for that.

I want to dedicate this work to you Larissa and Matilda. You mean everything to me.

Preface

How to memorize information sustainably to remember it over a long period of time has always been a question of core interest for human beings. Both researchers and educators have ever been striving to find suitable possibilities to improve learning strategies in order to boost mnemonic performance. Their goal was to identify the most effective ways of memorizing, enabling learners to improve quality of learning without having to increase study time. Amongst the first researchers emphasizing the importance of active repetition of learning material was William James who wrote in his *Principles of Psychology*(1890):

> *"A curious peculiarity of our memory is that things are impressed better by active than by passive repetition".*
> (adopted from Roediger & Karpicke, 2006a)

James observed that learning quality could be improved if students actively retrieved to-be-studied information during the learning phase. Instead of rereading information over and over again in order to memorize it, his suggestion was that they frequently should test themselves. Bearing in mind that at the end of the 19th century learning strategies applied in schools were often restricted to rote learning and teacher-centered repeated studying (see e.g. Cuban, 1993) one can imagine that he was quite ahead of his time. In fact, traditional ways of learning are built on a strict spearation between study and test. While study (i.e. exposure to the learning material) is used to acquire information, testing (i.e. active retrieval of the learning material) is typically only used to measure learning success at a later point of time.

In contrast to this tradition, research has now accumulated a lot of evidence pointing out more effective ways to learn. Many studies have over the last decades investigated the capability of retrieval practice cycles (test cycles) during learning to enhance long-term memory. Results convincingly show that retrieval practice can benefit memories to a large extent even if no feedback is provided. Additionally, they reveal that it is way more powerful in doing so than a comparable amount of restudy cycles - an effect that has been named the testing effect (see Roediger & Butler, 2011; Roediger & Karpicke, 2006a). Interestingly, this effect has been found to increase with retention interval between learning and final test and has even been found to persist after a time interval of one week (Roediger & Karpicke, 2006b). However, despite numerous studies and an overwhelming empirical support, this knowledge has yet to be brought to appliance in schools and universities. This might not only be due to the above mentioned deviating educational tradition, with tests being only used to assess learning performance. It might even be due to the fact that this effect is not very easy to grasp with intuition. In fact, it might be counterintuitive at first glance that retrieval practice can enhance memories even though study material is exposed to the learner by a lesser degree than e.g. during restudy. Thus, evidence on the testing effect can clearly provide interesting information about how we can learn effectively and about the importance of active learning instead of mere repetition of learning material.

Not only retrieval practice but also sleep has been found to be very beneficial for long-term memories. After initial study, memories need to be stabilized in a consolidation process in order to be recalled after a longer period of time. Based on empirical evidence, research has mainly focused on the conducive effects of sleep on consolidation processes. Indeed, sleep after learning typically results in better retention than a delay of comparable duration that is spent awake. This pattern of results has been replicated in numerous studies over the years and has become known as the sleep effect (see Diekelmann, Wilhelm, & Born, 2009; Rasch & Born, 2013). While the first studies on the relationship between sleep and memory ascribed this effect to a passive shelter that protects memories during sleep, recent research has found evidence for an active system consolidation (see Diekelmann & Born, 2010). Instead of being only an idle and disconnected state of mind, sleep

is now considered to imply active neuronal processes that can be associated to memory consolidation, and thus to better retention due to less time-dependent forgetting. Speaking in favor of such an active process, there are some studies showing that sleep can be selectively beneficial for some types of memory and less so for other types (e.g. Drosopoulos, Schulze, Fischer, & Born, 2007; Payne, Stickgold, Swanberg, & Kensinger, 2008; Wilhelm et al., 2011).

The present work aims to find out more about the influence of sleep on the testing effect. Thus far, testing effects have been investigated over delays with varying duration. In fact, as mentioned above, several studies suggest that testing effects increase with length of retention interval (Roediger & Karpicke, 2006b; Toppino & Cohen, 2009; Wheeler, Ewers, & Buonanno, 2003). Bearing in mind the beneficial effects of sleep on memory consolidation and the fact that longer delays but not shorter delays typically include one or several sleep intervals one might assume that sleep is a factor that fosters such facilitated retrieval-induced mnemonic benefits. In contrast to this assumption, studies on sleep-associated memory consolidation consistently point out that sleep effects occur primarily if learning and sleep are applied with a certain temporal proximity (see e.g. Gais, Lucas, & Born, 2006). Consistently, as previous studies on testing effects after longer delays usually did not control for length of time interval between learning and sleep, it has still to be investigated whether sleep actually might be a causal factor leading to enhanced testing effects after longer delays or if, alternatively, it might be that sleep does not influence the size of the testing effect or even decreases it. In fact, according to a study by Drosopoulos et al. (2007), mainly items with low memory strength (i.e. restudied items) but less items with high memory strength (i.e. items subjected to retrieval practice) should profit from sleep. Consistently, this would predict a decreased testing effect after sleep. To systematically investigate the influence of sleep on the testing effect, sleep and wake intervals following initial learning were controlled for in the present work.

Based on different theoretical frameworks of the testing effect, predictions about a possible relationship between testing effects and sleep are very diverging. According to the distribution-based bifurcation model (Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011) retrieval practice but not restudy is supposed to

lead to a strengthening of items that places them high above a certain recall threshold. Interestingly, as this recall threshold is already exceeded, further sleep-induced strengthening of items should consequently not result in significant benefits in a final memory test. This implies that learning material should mainly profit from sleep after restudy but less so after retrieval practice - i.e. testing effects should be reduced after sleep versus wake delay.

While the above-mentioned framework is basically strength-related, there are even other accounts focusing more on the cognitive processes resulting in testing effects. Generally, such accounts do not contradict the bifurcation model in any way but offer an explanation for how memories are strengthened by retrieval practice. The elaborative retrieval hypothesis (Carpenter & Delosh, 2006; Pyc & Rawson, 2009) assumes that testing effects are the result of effortful retrieval processes during the learning phase that lead to deeper processing of learning material through the activation of information semantically related to the target items (e.g. Carpenter, 2009). Consistently, this hypothesis would rather predict unaltered testing effects after sleep delay. Alternatively, it might even predict increased testing effects after sleep as sleep but not wake intervals have been found to activate semantic networks around target information (e.g. Darsaud et al., 2011) which might add to beneficial elaborative processes commenced by retrieval practice. However, reduced testing effects after sleep delay would not be easily allegeable with the elaborative retrieval hypothesis.

A third theoretical framework is offered by the episodic context account (Karpicke, Lehman, & Aue, 2014). Following this account, retrieval practice but not restudy is supposed to lead to a reinstatement of the the original learning context, updating the memory representation with information from the new temporal context during retrieval practice. Consistently, this is supposed to result in a reduced size of the search set during final test, fostering the testing effect. This account does not imply any direct predictions about a possible influence of sleep on the testing effect but in a study by Cairney, Durrant, Musgrove, and Lewis (2011) memory recall after sleep is suggested to depend less on contextual cues. Thus, less reliance on contextual cues and a reactivation of memories during sleep supposed e.g. by Rasch, Büchel, Gais, and Born (2007) might reduce the

testing effect. So, contrasting predictions of the elaborative retrieval hypothesis, the episodic context account does not predict greater testing effects after sleep delay. In fact, testing effects are either not affected or decreased, according to this account. Overall, the results of the present work might even be an interesting component of the research on theories about the factors underlying the testing effect.

In addition to reducing general time-dependent forgetting, both retrieval practice (e.g. Halamish & Bjork, 2011) and sleep (e.g. Ellenbogen, Hulbert, Stickgold, Dinges, & Thompson-Schill, 2006; but see Deliens et al., 2013) have been found to reduce memories' susceptibility to the detrimental effects of retroactive interference, which arise from additional study material that is applied after the learning of the original target material. However, interference effects on the testing effect and on the sleep effect have so far not been investigated conjointly in one study. The present work aims to not only investigate time-dependent forgetting but also interference-induced forgetting in relation to retrieval practice and sleep. Therefore, memory performance for varying learning material was tested after retrieval practice versus restudy over varying levels of practice and after a sleep versus wake delay either followed or not followed by interference induction. So the results will not only provide information about the influence of sleep on the testing effect but also about the capability of retrieval practice and sleep to reduce interference susceptibility.

Both testing and sleep have been repeatedly found to promote long-term retention of mnemonic information. However, these empirically supported benefits are yet to be brought to appliance as tools to improve learning e.g. in educational settings as in schools and universities. As mentioned before, there is often still a strict disposition of restudy being used during learning and testing solely being a measuring method for learning success. Similarly, even knowledge and implementation of the beneficial mnemonic effects of sleep is still quite scarce. Sleep is usually regarded a passive resting state of the body that does not actively contribute to memory consolidation but at the utmost serves as a shelter from interfering information that accumulates during wake periods. So both testing and sleep tend to be underestimated in their potential to function as tools to promote

effective learning. Thus, the results of the present work are not only of theoretical interest but can even provide useful information about how memory is affected by the combination of these tools and how this knowledge could be brought to practical appliance. If, for example, sleep would contribute to an increased testing effect this could be applied in educational settings to teach learners to boost their memories to an even higher degree than can be achieved by either testing or sleep. On the other hand, if it would lead to reduced testing effects, sleep could provide a possibility to reduce the gap between restudy and testing, i.e. making restudy a more promising study method in comparison to retrieval practice if learning is followed by a sleep interval compared to a wake interval. Thus, results of this work might even provide interesting information about a possible combinated appliance of both testing and sleep as memory modifiers in educational contexts.

Contents

Parts of the present thesis are published as:

Bäuml, K.-H. T., Holterman, C., & Abel, M. (2014). Sleep can reduce the testing effect - it enhances recall of restudied items but can leave recall of retrieved items unaffected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 1568-1581.

Abstract

One of the most powerful tools to promote memories is active retrieval of to-be-remembered information. Retrieval practice during the learning phase can improve memory performance and reduce time-dependent forgetting, compared to simple restudy. Several studies indicate that there are robust testing effects after longer delays, which are naturally filled both with periods of sleep and wakefulness. However, sleep delays have been found to affect memories differentially compared to wake delays. Research on the sleep effect shows that sleep can strengthen and stabilize memories resulting in mnemonic benefits. The present work aimed to investigate a possible influence of sleep on the testing effect. In a series of 4 experiments testing effects after wake versus sleep delays were analyzed using categorized item material (Experiments 1 and 2) as well as paired associates (Experiments 3 and 4). Following initial study, participants were asked to restudy the items or to engage in active retrieval practice. After a 12-h delay of either diurnal wakefulness or nocturnal sleep, subjects were asked to retrieve the items in a final recall test. Effects were investigated both in the presence (Experiment 2 and 4) and in the absence (Experiments 1 and 3) of retroactive interference. The results showed that there were reliable testing effects after wake delays, while testing effects were reduced or even eliminated after sleep delays. This pattern of results arose because only restudied items could profit reliably from sleep, while items did not show sleep-related benefits after retrieval practice. Additionally, testing effects were increased, while sleep effects being unaltered, in the presence versus absence of retroactive interference. Implications of these results are discussed on the basis of empirical and theoretical work on the testing effect and on sleep-associated memory consolidation.

Chapter 1

Background

1.1 The Testing Effect - Empirical Findings

Schools and other educational facilities are commonly established to pursue the goal of supporting students in the learning of new information that migth be of relevance for their future life. Traditionally, in an attempt to strengthen memories for previously encoded information, teachers employ different kinds of repetition methods, usually comprising restudy of the initially acquired learning material. After a certain period of time (delay) there is a test, which demands of students to actively retrieve the previously learned information from their memories. Typically, this test has the sole function to estimate students' knowledge about a certain topic that has been part of the lessons. Thus, the test of the learning material is commonly not considered part of the learning process but is simply used as an indicator for progress in learning.

Benefits of testing memories

In a survey Karpicke, Butler, and Roediger (2009) asked students about their preferred learning strategies. First, they were supposed to tell about their favorite strategy in a free report, while the second question was a forced report between the strategies of repeated study versus retrieval practice (self-testing). Results of the survey clearly showed that restudy was ranked highest by most participants, while retrieval practice was only rarely mentioned as preferred strategy. Moreover, of those who named retrieval practice as their favorite strategy, most specified that they used it to get feedback about their learning status rather than using it to improve memory performance (see even Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Kornell & Bjork, 2007). Consistently, when asking participants to predict their final recall performance, they usually overestimate their performance after restudy while underestimating their performance after retrieval practice (Roediger & Karpicke, 2006b). These and other studies suggest that common knowledge about the power of testing memories is only very fractional today. Hence, they point out the most probable reason for the infrequent use of this study strategy, possibly leading students to learn less effectively not fully

utilizing their memory's capacity.

Indeed, numerous studies indicate that tests can be effectively used during the learning process to benefit memories. In fact, evidence on the testing effect (see e.g. Roediger & Butler, 2011) indicates that retrieval-practice (test) cycles during the learning phase are, compared to restudy cycles, very powerful in reducing time-dependent forgetting (e.g. Hogan & Kintsch, 1971). This cannot only be found in laboratory studies but has even been investigated in educational settings (e.g. McDaniel, Anderson, Derbish, & Morrisette, 2007). Moreover, testing effects have proved to arise both in the presence and absence of corrective feedback and might not rely on successful retrieval during practice trials (e.g. Kornell, Hays, & Bjork, 2009). Effects have even been found after longer delays between learning phase and final test and seem to grow larger with the duration of such a time interval (e.g. Roediger & Karpicke, 2006b). In addition to its proven merits in reducing time-dependent forgetting, retrieval practice has even been found to reduce detrimental effects of both proactive interference (PI; e.g. Szpunar, McDermott, & Roediger, 2008) and retroactive interference (RI; e.g. Halamish & Bjork, 2011). Moreover it does not only affect practiced material itself but can lead to improved transfer of knowledge to new contexts (e.g. Butler, 2010) and can impede (e.g. Anderson, Bjork, & Bjork, 1994) or facilitate (e.g. Chan, 2009) retrieval of related but untested material depending on the conditions applied (for an overview over benefits of testing memories, see Roediger, Putnam, & Smith, 2011).

Testing can reduce time-dependent forgetting

Indeed, in contrast to most students' study habits, a long-standing tradition of research on the testing effect constantly emphasizes the benefits of active retrieval from memory (e.g. Allen, Mahler, & Estes, 1969; Duchastel & Nungester, 1982; Gates, 1917; Hogan & Kintsch, 1971). Already over 100 years ago Abbott (1909) mentioned benefits of recall over more pronounced time of actual perception (i.e. restudy). Still, despite increased numbers of publications on the issue over the last decades, the phenomenon remains fairly unnoticed outside memory research and,

thus, is only scarcely used in educational settings though its benefits have even been reported there (see below). Typical studies on the testing effect compare a retrieval-practice condition with a restudy condition. Retrieval practice usually implies that, after initial study of certain information, participants are asked to actively retrieve the information, while during restudy, they are simply repeatedly provided with the intially studied material (see e.g. Roediger & Karpicke, 2006b). Using a variety of learning materials like word lists (Hogan & Kintsch, 1971), paired associates (Carrier & Pashler, 1992) and pictures (Wheeler & Roediger, 1992), constant robust testing effects have been reported (for a review of recent literature see Roediger & Butler, 2011). Moreover, benefits of testing are not restricted to any specific type of test and have been found using e.g. free recall tests (Carpenter & Delosh, 2006), cued recall tests (Carpenter, Pashler, & Vul, 2006) and multiple-choice tests (Kang, McDermott, & Roediger, 2007). Additionally, not only advantages over restudy have been reported but testing seems to be even superior in reducing time-dependent forgetting compared to elaborative, and often recommended, study strategies as concept mapping (Karpicke & Blunt, 2011).

As mentioned before, several studies have even investigated the testing effect in an educational setting. Naturally, findings on the benefits of retrieval practice in the laboratory have sparked interest in its potential to improve learning of students in school. But it was not until some years ago that systematic investigations were waged in the classroom. In a study by McDaniel et al. (2007) a group of college students enrolled in a course took part in weekly tests on the topics taught or were presented target material for additional reading. Results in a cumulative final test clearly showed that taking tests improved students' memory for target material compared to restudy of the material. Several other studies found similar testing effects when investigating the capability of retrieval practice in educational settings (e.g. Butler & Roediger, 2007; Larsen, Butler, & Roediger, 2008; Spitzer, 1939). Thus, apart from findings in the laboratory, benefits of testing generalize to settings resembling classical learning scenarios at schools. This underscores the importance of an application in educational settings to improve students' learning skills. Similarly, also extending findings in the laboratory, some studies used learning materials more convenient for investigating everyday purposes. For

instance, Carpenter and Pashler (2007) asked participants to engage in visuospatial map learning either through restudy of the maps or computer-based tests with one feature of the maps missing. Even here subjects' map drawings after delay showed that retrieval practice is the more powerful method to enhance memory. Thus, the testing effect is a phenomenon that should be used more often in every day life to improve effective learning.

One might wonder if testing effects are restricted to retrieval practice with consecutive feedback. Indeed, in the absence of feedback subjects might stick to incorrect answers which then might impair memory performance in the final recall test. While restudy provides an opportunity to be repeatedly exposed to the original learning material, retrieval practice without feedback might lead to deteriorated results. In fact, providing feedback after retrieval practice has been found to result in greater memory performance (e.g. Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Kulhavy & Stock, 1989). However, numerous studies provide evidence that even retrieval practice without feedback leads to reliable testing effects (see e.g. Karpicke & Roediger, 2008; Roediger & Karpicke, 2006b). Still, especially multiple choice tests pose the threat of lures being regarded as correct answers and studies show that choosing such a wrong alternative leads to repeated mistakes during comprehensive final tests (Butler, Marsh, Goode, & Roediger, 2006; Marsh, Roediger, Bjork, & Bjork, 2007). Thus, particularly in multiple-choice testing, feedback can be an effective way to reduce the negative effects of misleading response options (Butler & Roediger, 2008). While researchers agree on the fact that correct responses during retrieval practice can promote memory performance, recent research even points out a positive effect of incorrect responses on later memory retrieval. Kornell et al. (2009) examined testing effects using material that ensured unsuccessful retrieval attempts during the practice phase followed by corrective feedback. Results showed that such unsuccessful retrieval attempts can foster memory performance in a delayed final test. Thus, even tests that pose a certain challenge, and not only errorless testing, might be beneficial for later recall (see even Richland, Kornell, & Kao, 2009).

Typical studies on the testing effect consist of different phases. Participants are initially asked to study the learning material and then to either restudy it or

practice retrieval of it. Eventually, after a certain delay, they are asked to recall the material in a final test. In some recent studies the length of intermediate delay was varied to investigate whether length of delay can moderate the testing effect. Roediger and Karpicke (2006b) asked participants to study prose passages and, afterwards, let them either restudy or engage in retrieval practice of the passages. This practice phase was then followed by varying time intervals depending on the delay condition. Delay lasted either for five minutes, two days, or a whole week before final test was administered. Results indicate that testing effects depend on the time spent between practice phase and final test. Indeed, restudy was found to be superior to retrieval practice in reducing time-dependent forgetting over five minutes. In contrast, typical testing effects were only observed after longer delays of two days or one week. Roediger and Karpicke (2006b) used prose passages as learning material and free recall during the practice phase, which resembles typical learning in educational settings. However, this leaves room for speculations on learning strategies used by participants. Therefore, Toppino and Cohen (2009) strived to replicate the finding of a test-delay interaction under more controlled conditions. Using paired associates and a cued-recall test during the practice phase, they found a testing effect after a delay of two days but, in line with the results by Roediger and Karpicke (2006b), an inverted testing effect after only few minutes. Thus, a test-delay interaction with benefits of restudy over retrieval practice after short delays but benefits of retrieval practice over restudy after longer delays was consistently evident in both studies (for similar results, see Wheeler et al., 2003; for theoretical considerations on the test-delay interaction, see below). The short-term advantage of restudy over retrieval practice might be one of the reasons for the fact that retrieval practice is often underestimated as a memory enhancer and, therefore, is only rarely used in educational contexts (see above).

Retrieval practice and interference

While research on the classical testing effect has been taking place over many decades and findings are manifold (see above), research of recent years has extended

classical findings by new insights going beyond earlier explored effects of retrieval practice. Instead of focusing on the capability of testing to reduce time-dependent memory for practiced material, Szpunar et al. (2008) investigated the effects on recall of subsequently learned new material. Memory for learned target information can be impeded by previously studied material - an effect typically refered to as proactive interference (PI; see e.g. Postman & Keppel, 1977; Underwood, 1957; Wixted & Rohrer, 1993). Szpunar et al. (2008) asked participants to study several lists of items, each list being followed by retrieval practice or restudy of the list or by a distractor task. Critical for the study was recall of the last list learned (List 5), which determined the degree of disadvantageous effects of PI. Interestingly, results indicated that testing previously learned material can insulate against the buildup of proactive interference, protecting subsequently learned target information from the negative influences of PI (for similar results see Pastötter, Schicker, Niedernhuber, & Bäuml, 2011; Weinstein, McDermott, & Szpunar, 2011).

As mentioned above, retrieval practice of certain material can reduce the detrimental effects of proactive interference (PI) arising from this material on subsequently learned information (Pastötter et al., 2011; Szpunar et al., 2008; Weinstein et al., 2011). Another form of interference that can be a major cause of forgetting is retroactive interference (RI; see e.g. McGeoch & McDonald, 1931; Underwood, 1948). This type of interference arises when, after studying certain target information, additional related information is encoded. In a classical study by Barnes and Underwood (1959), participants were asked to learn a target list of word pairs (AB) which was followed by a similar consecutive learning list (AC), with identical stimulus items (A) but different response items (C) (see also Müller & Pilzecker, 1900). Results revealed that, compared to a control condition, presence of the additional learning list reliably reduced recall performance for the target list. According to the cue-overload principle, items that are related to the same cue (A) compete with each other for access to conscious awareness (e.g. Watkins, 1979; Watkins & Watkins, 1976). As a consequence of this competition for memory resources, interference is supposed to arise (see Anderson & Neely, 1996). More recent theoretical reasoning extends these considerations

about the origin of interference and the resulting forgetting. Wixted (2004) argues that retroactive interference largely acts on memory traces that have not yet been consolidated in memory (see even Wixted, 2005). Consolidation processes gradually stabilize memory traces over a period of time after encoding (for further information about memory consolidation, see below and e.g. McGaugh, 2000). This point of view is supported by the classical finding of a temporal gradient of retroactive interference, meaning that interference effects are the more pronounced the earlier interfering material is learned after the encoding of target information (e.g. Müller & Pilzecker, 1900; Skaggs, 1925).

Due to its prominent role in theories about forgetting, retroactive interference has also been frequently investigated in combination with retrieval practice. In a study by Halamish and Bjork (2011) participants were asked to study a target list of paired associates (AB). In line with typical studies on the testing effect, this phase was either followed by restudy or by retrieval-practice cycles of the target list. To induce retroactive interference, participants were subsequently asked to study another list of paired associates with identical stimulus items (A) but new response items (C; AB - AC paradigm; see above). Results showed that testing effects were enhanced in the presence of retroactive interference. In other words, retrieval practice could protect target material from subsequently studied interfering material, i.e. it reduced items' susceptibility to retroactive interference (for similar results, see Abel & Bäuml, 2014; Potts & Shanks, 2012). Thus, retrieval practice has been found to reduce effects of proactive interference accruing from studied material, impairing subsequently learned material (Szpunar et al., 2008) and to reduce susceptibility to retroactive interference of material on previously studied target material. Additionally a recent study by Kliegl and Bäuml (2016) shows that retrieval practice but not restudy can insulate memories against intralist interference. Together, this indicates that retrieval practice may help to distinguish target information from interfering information (Halamish & Bjork, 2011).

Effects of testing going beyond practiced material

Over the years, beyond benefits for tested material, effects of retrieval practice on memory have sparked interest, resulting in numerous studies investigating the issue. In a seminal article Anderson et al. (1994) described a phenomenon called retrieval-induced forgetting (RIF) which they explored using the retrieval-practice paradigm. Participants were initially asked to study categorized item material (e.g. Fruit - Apple, Fruit - Orange, Tree - Hickory, Tree - Elm) and to consecutively actively retrieve parts of the items from parts of the categories in a retrieval-practice phase, being provided with the category name and a fragment of the correspondent item (e.g. Fruit - Ap___). Hence, after this phase, there were three types of items: Practiced items (RP+), unpracticed items from practiced categories (e.g. Fruit - Orange, RP-), and unpracticed items from unpracticed categories (NRP) that served as control items. The results in a final test showed a typical benefit of retrieval practice for practiced items (RP+) over control items (NRP). However, going beyond this classical testing effect, the most astonishing finding was that unpracticed items from practiced categories (RP-) showed diminished recall compared to control items. Thus, there were benefits of retrieval practice for practiced items but costs for items from the same semantic categories that were not practiced during the retrieval-practice phase (retrieval-induced forgetting). This phenomenon is often proposed to be the result of inhibitory processes active during retrieval practice (see e.g. Anderson, 2003; Bäuml, Pastötter, & Hanslmayr, 2010). To overcome interference between competing items, unpracticed items (RP-) are supposed to be inhibited through executive control and, thus, reduced in memory strength to enable successful retrieval of practiced items (RP+). Consequently, retrieval-induced forgetting is proposed to play an important role in mnemonic functions of our everyday life supporting memory for important information at the expense of less relevant infomation.

However research of the last years points out that retrieval practice does not only initiate a self-limiting process reducing recall of related but untested material but can even be self-facilitating. Indeed, using prose material Chan,

McDermott, and Roediger (2006) could show that retrieval practice of a subset of initially studied prose passages could benefit recall of another subset that was not tested after 24 hours (retrieval-induced facilitation). At first glance this finding seems to contradict evidence on retrieval-induced forgetting. However, in a follow-up study Chan (2009) could identify boundary conditions of both retrieval-induced forgetting and retrieval-induced facilitation. They varied both degree of integration of the studied passages as well as the delay following the learning phase. Results showed that facilitation occured when material was highly integrated and the final test was delayed by 24 hours. In contrast, when integration was disrupted and the final test occured already after 20 minutes, retrieval-induced forgetting could be observed. Using a different but related paradigm Bäuml and Samenieh (2010) showed that retrieval-practice of certain item material could facilitate recall for related intially studied but untested material that subjects previously had been asked to forget (directed forgetting; e.g. Bjork, 1989) while it induced forgetting if subjects had previously been asked to keep remembering the material (two faces of memory retrieval). In accordance with this, retrieval practice after a context change has been found to result in retrieval-induced facilitation while retrieval-induced forgetting has been found to be evident in the absence of such a context change (Bäuml & Samenieh, 2012; Bäuml & Schlichting, 2014). Thus, whether retrieval-induced forgetting or retrieval-induced facilitation can be observed seems to depend on the specific set of conditions and the paradigm used. The above described studies clearly demonstrate that even memory material that was not subjected to retrieval practice might be affected by it, which goes beyond classical findings on the testing effect.

Overall, research has accumulated numerous studies providing empirical evidence for the power of retrieval practice in reducing time-dependent forgetting. Moreover, recent investigations have even pointed out the capability of testing to prevent detrimental effects of proactive and retroactive interference, stabilizing target memories and segregating them from other interfering information. Going beyond benefits for practiced material, enhanced transfer of learning has been associated with retrieval practice and taking tests during the learning phase has been found to affect related but unpracticed material. However, effects of

testing memories are still widely underestimated and retrieval practice is mainly regarded and used as an evaluation method, rather than to improve the learning process. This is not only in contrast to compelling findings of testing effects in the laboratory, but even more so to applied studies showing reliable effects even in educational settings. Taking all of this into consideration, active retrieval clearly benefits memories in many respects (for an overview, see Roediger & Butler, 2011; Roediger et al., 2011). In addition to this vast amount of empirical evidence, research has undertaken attempts to even tackle its theoretical background. The following paragraphs shall give an insight into considerations about the origin and causes of the testing effect.

1.2 The Testing Effect - Theoretical Background

Although knowledge about the testing effect has been around for quite some years, insights about the origin of retrieval-related benefits are still quite sparse (see Roediger & Karpicke, 2006a). The following pages shall give an overview over the most prominent theoretical accounts providing explanations for the background of the testing effect and empirical findings related to it.

Overlearning and the theoretical framework of the bifurcation model

Early research on the testing effect focused on the fact that subjects are exposed to the learning material for a longer time if they engage in retrieval practice which results in a greater amount of processing (amount-of-processing acount, see e.g. Kolers, 1973; Slamecka & Katsaiti, 1987; Thompson, Wenger, & Bartling, 1978). Basically, according to this account, tested material is supposed to be subject of plain overlearning, leading to better recall performance in the final delayed test. The idea of additional exposure accounting for the effect of testing might originate in the fact that earlier studies usually compared a retrieval-practice condition with a no-practice condition (control condition) (e.g. Spitzer, 1939). Thus, in the face of a lacking restudy control condition, benefits of retrieval practice might easily be attributed to additional exposure time. In the presence of such a control condition, this explanation of the testing effect seems unlikely though, as restudied material likewise is exposed further after initial study. In fact, restudied material is even exposed intactly for a longer period of time than material subjected to retrieval practice, where typically only fragments of the material are presented during the practice phase. Further evidence against this exposure-based account comes from studies on the test-delay interaction. While restudy seems to be superior to retrieval practice after short delays between learning phase and final test, this pattern is reversed, resulting in a reliable testing effect after longer delays (Roediger & Karpicke, 2006b; Toppino & Cohen, 2009). Hence, apparently, overlearning

is taking place during restudy cycles but benefits of this are only short-dated and no longer evident after longer time intervals. If testing effects were due to additional exposure of the learning material, this test-delay interaction would not be apparent. Thus, explanations based on additional exposure do not offer a satisfactory theoretical basis accounting for benefits of retrieval practice (see also Glover, 1989).

A recent theoretical framework takes a different approach to explain benefits of testing memories and to account for several empirical findings and boundary conditions of the testing effect. According to the distribution based bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011), memory strength distribution of learning material becomes bifurcated through retrieval practice. Core assumptions of this model are that memory strength of studied items is normally distributed on a continuum and that material is recalled in a final test only if it lies above a certain recall threshold; i.e. the recall test does not measure an item's memory strength directly but items are recalled correctly only if their memory strength is above recall threshold (Kornell et al., 2011). Also critical to the model is that it only applies for situations of retrieval practice without feedback. Before initial study, according to the model, all items are supposed to be normally distributed on a memory strength distribution. In conjunction with intial study, all items are then supposed to be strengthened equally, graphically leading to a shift of the whole normal-distribution curve on the memory-strength axis. However, during the following practice phase restudy and retrieval practice are supposed to influence memory-strength distribution in very different ways. While restudy leads to a further strengthening of all restudied items, retrieval practice creates a bifurcated distribution. Successfully retrieved items are strengthened to a greater degree than restudied items, while incorrectly or not retrieved items remain at the same memory-strength level as before the practice phase (for a graphical illustration of the bifurcation model, see Figure 1, see also Halamish & Bjork, 2011; Kornell et al., 2011). These core assumptions of this model can explain why time-dependent forgetting might be reduced after retrieval practice. As items that were successfully retrieved during the practice phase are strenghtened to a higher degree than restudied items, more of these items remain above recall threshold

after a certain delay, resulting in the classical testing effect.

Going beyond reduced time-dependent forgetting, the bifurcation model can even provide an explanation for several other findings related to the testing effect (Halamish & Bjork, 2011). Depending on the recall threshold at final test, it predicts very different results (see Figure 1). Basically, according to the model, difficulty of the final test should moderate the size of the testing effect. Final-test difficulty should lead to a higher recall threshold, resulting in greater testing effects as mainly items that were successfully retrieved during retrieval practice might pass this higher threshold. In contrast, restudied items, being overall lower in memory strength, would only pass a higher threshold, if at all, by a smaller margin. On the other hand, if test difficulty was lower, resulting in a lower recall threshold, considerably more restudied items would be above threshold, while mainly just the successfully retrieved proportion of items subjected to retrieval practice would make it beyond threshold. As the final recall test does not measure item-based memory strength but just the mean amount of items above recall threshold, such conditions might even result in an inverted testing effect, i.e. better recall for restudied items than for tested items (see Figure 1). Thus, the bifurcation model predicts an interaction between final-test difficulty and the size of the testing effect (see Halamish & Bjork, 2011).

One possibility to test this assumption systematically is to apply differing final test formats. Halamish and Bjork (2011) asked participants to study a list of paired associates in one initial study cycle which was then followed either by restudy or retrieval-practice cycles. Final test format was manipulated so that participants engaged either in a cued-recall test with cue item and fragment of the target item (easy) or only cue item (intermediate) presented, or in a free recall test of the target items (difficult). In line with predictions made by the bifurcation model, results show that only difficult final test formats lead to a testing effect, while easier formats might even result in inverted testing effects (for similar results, see Hogan & Kintsch, 1971; Kang et al., 2007; but see Glover, 1989). Moreover, as mentioned above, many studies have reported a test-delay interaction, i.e. testing effects to be absent (or even reversed) after short delays between practice phase and final test but testing effects to be present after longer delays (Roediger &

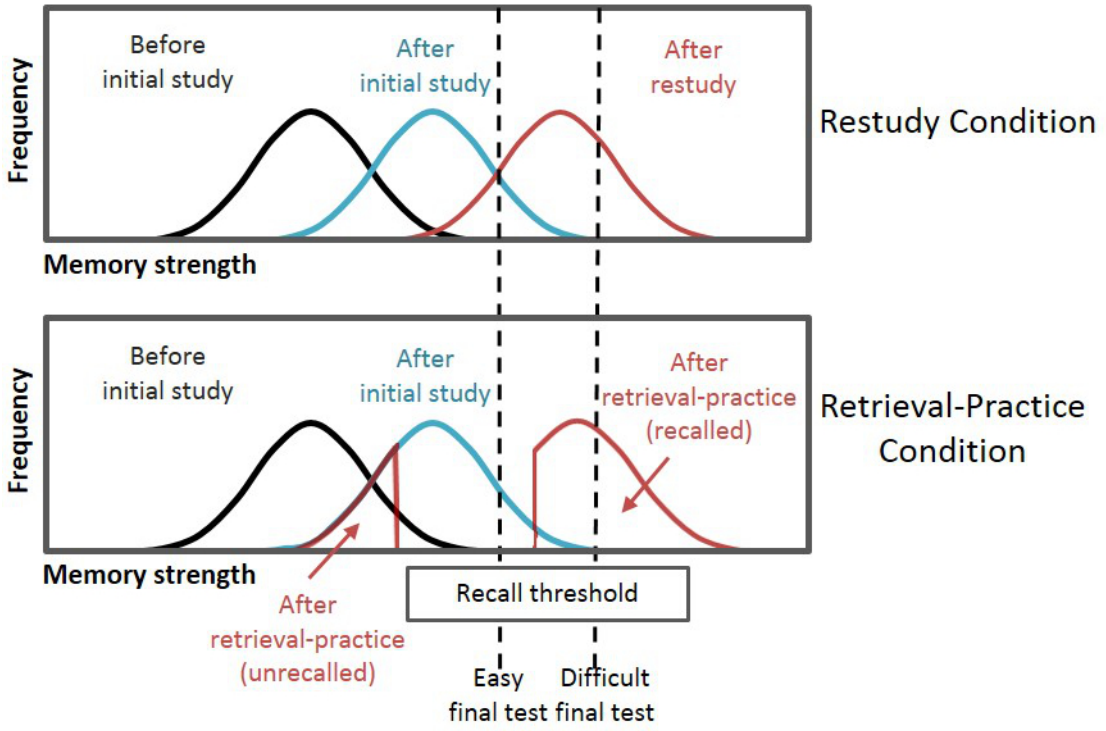

Figure 1: Memory strength distributions in the restudy- (upper panel) and the retrieval-practice condition (lower panel), based on the bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011). In conjunction with intial study, all items are supposed to be strengthened equally (blue curves). During the following practice phase restudy leads to a further strengthening of all restudied items while retrieval practice creates a bifurcated distribution. Successfully retrieved items are strengthened to a greater degree than restudied items, while incorrectly or not retrieved items remain at the same memory-strength level as before the practice phase. This implies that difficulty of final test moderates the size of the testing effect. Easier final tests (left vertical dotted line) create a lower recall threshold than more difficult final tests (right vertical dotted line). Successfully retrieved but not restudied items might pass a higher recall threshold by a great margin leading to a testing effect. In contrast, restudied items might pass a lower recall threshold by a greater margin while, after retrieval practice, only the successfully retrieved proportion of items would be recalled. This might result in testing effects to be present in difficult final tests and absent in easier final tests (here testing effects might even be inverted with benefits of restudy over retrieval practice).

Karpicke, 2006b; Toppino & Cohen, 2009; Wheeler et al., 2003). It has been suggested earlier that this interaction is the result of restudy supporting memory strength and retrieval practice supporting item retrievability (see e.g. Bjork, 1975). The bifurcation model offers a simpler explanation for the test-delay interaction as longer delays but not shorter delays might result in increased final test difficulty (see Halamish & Bjork, 2011). At shorter final-test delays, many restudied items might still be above recall threshold, possibly resulting in advantages of restudy over retrieval practice during final test. On the other hand, if delay is prolonged, mainly the items that were successfully retrieved during retrieval practice remain above threshold, while most of the restudied items should have already crossed below threshold, resulting in a classical testing effect. Under such conditions, it is critical which type of practice (restudy versus retrieval practice) produces a greater number of items that remain above recall threshold.

A third factor that migth moderate difficulty of the final test is the presence or absence of retroactive interference (see above). Several studies show that retrieval practice but not restudy can reduce items' susceptibility to retroactive interference, resulting in enhanced testing effects in the presence of retroactive interference compared to its absence (Abel & Bäuml, 2014; Potts & Shanks, 2012). These results raise questions about the origin of such an insulating effect of retrieval practice. To investigate them in the light of predictions of the bifurcation model Halamish and Bjork (2011, Exp. 3) induced retroactive interference directly before the final test. Results revealed once more that final-test difficulty can moderate the testing effect, as testing effects were evident in the presence of retroactive interference but reduced or even numerically reversed in its absence. According to Halamish and Bjork (2011) this is caused by the fact that mainly restudied items suffer from diminishing effects of interference while items that were subject to successful retrieval practice are high above recall threshold, resulting in better performance in the final test. Thus, even the finding of retrieval practice protecting memories from retroactive interference is in line with the bifurcation model. Overall, this shows that empirical evidence supporting the bifurcation model is manifold. According to the model final-test difficulty should be a moderating factor of the testing effect. This has been supported by studies investigating

final-test format of varying difficulty and final-test delay of varying length, as well as studies exploring testing effects in the presence and absence of retroactive interference (RI).

Overall, the bifurcation model is a general framework that provides an account for a certain pattern of results without specifying cognitive processes underlying benefits of retrieval practice over restudy. It offers a strength-related explanation for the influence of the three most prominent moderating factors of retrieval-based mnemonic benefits - test delay, test format, and interference. In contrast to process-related accounts (see below), it does not examine the deep structure behind how and why this strengthening occurs and, thus, does not in any way contradict assumptions made by these accounts.

The elaborative-retrieval hypothesis and the mediator-effectiveness hypothesis

Other research has focused more on the unique features of testing that might result in retrieval-related benefits. One advantage retrieval of learning material has over restudy is that it better resembles conditions during the final test. Thus, the transfer-appropriate-processing account states that benefits of testing are related to matching test conditions between initial retrieval practice and the final cumulative test (e.g. Morris, Bransford, & Franks, 1977). In other words, the same skills are supposed to underly retrieval practice and later final tests. This should give tested material an advantage over restudied material that lacks such "pre-experiencing" of later test conditions. This implies that if test format during retrieval practice and final test is matched, testing effects are supposed to be greatest. Though there is some evidence supporting this claim (e.g. McDaniel & Fisher, 1991; McDaniel, Kowitz, & Dunay, 1989), more recent research challenges such an assumption. In a study by Carpenter and Delosh (2006) differing test formats during retrieval practice and final test were paired with each other and results were investigated systematically. Results show that matching test formats are not driving the testing effect. Contradicting the transfer-appropriate-processing account, retention was not highest when format

of retrieval practice and final test were of the same type. Similar evidence comes from a study by Glover (1989) who found final-test retention to be independent of matching test formats between retrieval practice and test. Thus, transfer-appropriate processing cannot fully account for the benefits of testing memories.

Consequently, neither enhanced exposure to the learning material (see above), nor transfer-appropriate processing can satisfactory explain retrieval-induced mnemonic benefits. Hence, some process feature of retrieval practice itself might evoke the effect. Interestingly, results of both the study by Carpenter and Delosh (2006) and the study by Glover (1989) revealed, that retention in a final recall test was best when participants were asked to retrieve the item material by means of a free-recall task and less pronounced when retrieval practice involved easier test formats. These findings are in line with the desirable-difficulty framework (e.g. Bjork, 1994, 1999) according to which difficult learning processes can enhance long-term retention, even though initial learning is slowed. According to Roediger and Karpicke (2006b) retrieval practice compared to restudy creates such desirable difficulties. Indeed, retrieval practice requires more elaborative effort which is supposed to foster deeper processing of materials (e.g. Gardiner, Craik, & Bleasdale, 1973) and consequently greater strengthening. This lies at the basis of the elaborative-retrieval account (see e.g. Carpenter & Delosh, 2006) and has been repeatedly confirmed in empirical studies (e.g. Carpenter, 2009; Jacoby, 1978; Pyc & Rawson, 2009). Consistently, in addition to aggravated test formats, several findings indicate that benefits of testing are even increased if retrieval practice is rendered more difficult by other means. This involves for example conditions of interference induction prior to retrieval practice (Cuddy & Jacoby, 1982) or prolonged retention intervals between initial study and retrieval practice (Landauer & Eldridge, 1967; Whitten & Bjork, 1977).

To further investigate the nature of such elaborated processing Carpenter (2009) asked participants to study cue-target pairs that were semantically weak (e.g. Basket-Bread) or strong associated (e.g. Toast-Bread). In a subsequent practice phase they were asked to either restudy the item pairs (Toast-Bread) or engaged in retrieval practice in a cued-recall test (Toast-_____?). Results in a

delayed final test revealed a classical testing effect, i.e. better recall after retrieval practice than after restudy. Moreover, a comparison between weakly and strongly associated item pairs showed that they were recalled at a similar rate after restudy. In contrast, after retrieval practice, weakly associated item pairs were more likely to be recalled than strongly related pairs. This underscores the role of elaborative processes in the formation of the testing effect, as weak association requires greater retrieval effort. In addition, these results give an insight in how elaboration might lead to deeper processing of mnemonic information. It might be that greater effort during retrieval practice initiates the activation of a broad semantic network. Such spreading activation (e.g. Collins & Loftus, 1975) might be fostered by employing weaker cues and less by stronger cues. Indeed, while stronger cues during the practice phase imply easy and fast recall, weaker cues might initiate a beneficial process leading to the build-up of a semantic network facilitating recall during the final test by the means of the establishment of multiple retrieval routes.

A related concept is addressed by the mediator-effectiveness hypothesis which states that effortful retrieval practice creates more effective mediators during the learning phase than restudy. Mediators are words, phrases, or concepts linking a cue with a target item. In a study by Pyc and Rawson (2010) participants were asked to study paired associates and to generate keyword mediators between cue and target item before the subsequent practice phase. Results reveal that recall of mediators was reliably increased after retrieval practice versus restudy. Moreover, providing such mediators during final test enhances recall after restudy but not retrieval practice, indicating that mediators might be automatically activated during final tests after retrieval practice. According to Pyc and Rawson (2010), retrieval practice enables subjects to choose more effective mediators as unsuccessful retrieval could initiate a process that might support shifting from less effective to more effective mediators. This is in line and related with research suggesting that active retrieval but not restudy enhances the integration of target memories in existing semantic networks multiplying the amount of possible retrieval routes during the final test (see above; see even McDaniel & Masson, 1985). Indeed, mainly the effort during retrieval processes seems to be relevant and extra semantic information is activated even during incorrect attempts to retrieve

the target information, leading to multiple available retrieval cues during delayed final tests (Soraci et al., 1999, 1994) or a choice of more effective mediators(Pyc & Rawson, 2010). Overall, research supports the idea that retrieval practice leads to better retention as it is more effortful than simple restudy, which results in deeper and more elaborative processing of the learning material.

The episodic context account

The most recent account on the testing effect, the episodic context account, uses a different approach to try to explain benefits of retrieval practice over restudy (Karpicke et al., 2014). The term context, which is central to this account, has earlier been applied for several different factors that usally shift during the course of the memory process and affect delayed recall performance. Probably the most familiar of them is the external context, which is related to the surroundings a subject experiences during the encoding and later retrieval phase. This type of context has been traditionally manipulated in studies by changing exterior variables as e.g. asking participants to change their physical location between encoding and retrieval of memory material (see, e.g. Davies, 1986) or by shifting perceptual variables as e.g. visual or auditory characteristics (see, e.g. S. M. Smith, 1985). Moreover the term internal context has been used to describe a subjects internal mental state at a certain point of time (e.g. Bower, 1981). Generally, studies manipulating the context between encoding phase and subsequent recall phase usually point out the important role it plays for successful retrieval processes. Indeed, relatively coherent contexts beetween encoding and retrieval phase have been found to lead to better recall performance than if context is purposefully shifted and therefore incoherent (see Godden & Baddeley, 1975).

The episodic context account focuses on the temporal context, which is a representation of external and internal context supposed to slowly shift with the passage of time (Howard & Kahana, 2002). Different models have been employed to describe this form of context change. While some of them assume that this shift over time occurs in a random manner (e.g. Lehman & Malmberg, 2013), others link it to constant retrieval processes on the basis of contextual cues (e.g.

Howard & Kahana, 2002). Regardless of this differentiation, the episodic context account assumes that contexts drift over time and that contexts of events that occur in temporal proximity are more similar than contexts of events that occur with increasing delay. Thus, according to this account, temporal context cues will presumably be more coherent with the encoding context after shorter versus longer delays. That means that with increasing delay more and more features of the encoding context have to be reinstated in order to successfully retrieve an item. Core assumption of the episodic context account is that retrieval practice leads to an update of the temporal context that exceeds contextual updating through restudy. When a subject is supposed to actively retrieve a certain item from the original set of encoded items after a delay, this demands that some features of the encoding context are reactivated, while less reactivation is needed during restudy. The episodic context account assumes that retrieval creates an updated context representation that contains a mixture of features from the prior (study) context and from the current context in which retrieval practice occurs.

When trying to retrieve a certain item during a recall test, one assumes that subjects are doing so on the basis of a certain search set of items. This search set contains all items to be considered candidate items (Raaijmakers & Shiffrin, 1981). Recall performance at the final test is supposed to rely on retrieval cues that can lead to a reduction of the size of the search set, i.e. that can help to identify the candidate item by contrasting it to and reducing the number of competing items. Consistently, as retrieval practice is supposed to reinstate parts of the original encoding context and to update it with features of the retrieval context, the search process during the final recall test can become more efficient. Indeed, the search set might be reduced to items fitting both the context during encoding and the context during retrieval practice, facilitating item recall during final test. Thus, according to the episodic context account (Karpicke et al., 2014) retrieval practice is benefitting recall performance by reinstating the original encoding context, updating the memory representation with information from the new temporal context during retrieval practice, finally resulting in a restricton of the search set during final test. Thus, the present account offers a context-based explanation for the advantage of retrieval practice over restudy.

The episodic context account is related to some accounts that have been proposed to explain the spacing effect, i.e. better retention after spaced than after massed learning (see e.g. Delaney, Verkoeijen, & Spirgel, 2010). According to contextual variability accounts of the spacing effect, the occurrence of a studied item in two different contexts produces a varied set of retrieval cues facilitating successful retrieval in a later test. Contextual drift is supposed to result in greater contextual variability if items are learned in a spaced and not in a massed fashion. In fact, the greater the distance between two presentations during the initial study phase, the more contextual variability is supposed to occur (e.g. Glenberg, 1979). Study-phase retrieval accounts propose that spacing effects occur as information is retrieved during additional presentation in the study phase. Thus, additional information is added to the original memory trace when the prior presentation is retrieved (e.g. Greene, 1989). Spacing effects are assumed to occur when such retrieval tasks require reliance on contextual cues (see Kahana & Greene, 1993). The episodic context account extends these ideas as it refers to an intentional rather than incidental reinstatement of the study context through retrieval practice. The amount of contextual updating is likely to be greater for intentionally retrieved items than it is for incidentally retrieved items (see Karpicke et al., 2014). Recent results by Bäuml and Dobler (2015) indicate that retrieval practice might be more effective in context retrieval than restudy, supporting these assumptions made by the episodic context account.

Evidence for this account of the testing effect comes from several empirical studies. As mentioned above Carpenter and Delosh (2006) examined testing effects after retrieval practice of varying test formats. They consistently found free-recall tests to produce the best retention of item material. When examining these results from the perspective of the episodic context account they can be reasonably accounted for. One might assume that test formats leading to relatively easy retrieval might not require much context reinstatement or updating. Thus, easier retrieval practice would not lead to a noteworthy reduction of the search set (see Karpicke et al., 2014), implying less pronounced testing effects than after free recall. Similarly, even the results of the study by Carpenter (2009) can be accounted for by differences in context reinstatement during retrieval practice. As

mentioned earlier, they show that weakly associated item pairs (e.g. Basket-Bread) profit more from testing than strongly associated item pairs (e.g. Toast-Bread) do. One can imagine that retrieval from a weak semantic cue affords more context reinstatement than does retrieval from a stronger semantic cue. Strong semantic cues imply easy access to the target items while weaker cues need to be complemented by additional context information to lead to successful retrieval. Thus, according to Karpicke et al. (2014) it is not difficulty of retrieval practice per se that facilitates testing effects but the extend to which context reinstatement is required.

Additional evidence comes from a recent study by Lehman, Smith, and Karpicke (2014) which contrasted retrieval practice against semantic elaboration. Subjects were asked to study five lists and then freely recall the last list. In the control condition, subjects studied and performed a distractor task between lists. In the retrieval practice condition, subjects studied and then tried to recall each list after studying it in a free recall task. Finally, in the elaboration condition the subjects were shown the items and instructed to generate semantic associates for each word after studying each list to test for the semantic elaboration account (Carpenter, 2009). Results show that only retrieval practice could enhance recall of the final list and reduce intrusions from prior lists while semantic elaboration reduced recall and increased number of intrusions compared to the control condition. Additionally, when examining retrieval dynamics by analyzing cumulative recall, they noticed that retrieval practice but not semantic elaboration led to an earlier and more rapid approach to asymtote compared to control condition. This indicates that retrieval practice leads to a restriction of the search set while semantic elaboration rather extends it (see Bäuml & Kliegl, 2013; Wixted & Rohrer, 1994). Moreover, this underscores the importance of context reinstatement for the testing effect while challenging the assumption that semantic elaboration might lie at the core of beneficial effects of retrieval practice. Additionally, also supporting the episodic context account, results of a recent study by Kliegl and Bäuml (2016) indicate that retrieval creates distinct context features for individual retrieved items, reducing intralist interference.

Summary and evaluation of the different accounts

Over time several different accounts have been brought forward to explain the beneficial effects of retrieval practice over restudy (testing effect). According to the bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011), retrieval practice results in a bifurcation of item strength between items successfully retrieved during retrieval practice and those not retrieved. In contrast, all restudied items are strengthened but at a smaller margin. This assumption combined with the idea of a recall threshold can explain the finding of reduced time-dependent forgetting after retrieval practice and can even account for and is supported by several other empirical findings related to the testing effect (see above). The bifurcation model offers a strength-related framework for the testing effect. Another account that focuses more on the nature of cognitive processes active during retrieval practice is the elaborative-retrieval account (Carpenter, 2009; Carpenter & Delosh, 2006). Core assumption of this account is that testing but not restudy leads to an elaborative retrieval process that facilitates the buildup of semantic mnemonic networks around the target items resulting in better retention during final recall. A related concept, the mediator-effectiveness account (Pyc & Rawson, 2010), assumes that retrieval practice supports the activation of more effective mediators between cues and target items alleviating recall during final test. Finally, the most recent account, the episodic context account (Karpicke et al., 2014), links the testing effect to effects of context reinstatement and updating during retrieval practice, resulting in a reduced search set of items.

Overall, when evaluating the different accounts of the testing effect that are described above, one might classify them into two different categories. The bifurcation model offers a strength-related explanation that does assume a greater strengthening induced by retrieval practice than restudy. It does not offer any further assumption for how this strengthening is induced. In contrast, both the elaborative-retrieval hypothesis (and even the mediator effectiveness account) and the episodic context account are based on an explanational framework for how retrieval practice as opposed to restudy might benefit later memory recall. Thus, the bifurcation model does not contradict these accounts in any way as

they are process-related. The assumptions made by the bifurcation model are supported by empirical findings based on studies varying difficulty of the final recall test. Thus, it offers an explanation for the test-delay interaction (Roediger & Karpicke, 2006b; Toppino & Cohen, 2009; Wheeler et al., 2003). Additionally it can account for variations in the size of the testing effect based on different final test formats (Halamish & Bjork, 2011; Hogan & Kintsch, 1971; Kang et al., 2007) and the presence or absence of retroactive interference (Abel & Bäuml, 2014; Potts & Shanks, 2012). In contrast, studies supporting the elaborative retrieval account and the mediator effectiveness account focus more on characteristics of the initial retrieval-practice phase. Core assumption is that effortful and elaborative retrieval fosters recall performance (Carpenter, 2009) and leads to more effective mediators(Pyc & Rawson, 2010) linking cues with target items. This is challenged by the episodic context account that offers a more context based explanation for such findings. Additional evidence for this account comes from a study by Lehman et al. (2014) who contrasted retrieval practice against semantic elaboration (see above).

1.3 Sleep-Associated Memory Consolidation

The function and relevance of sleep is an often discussed issue that has always fascinated researchers (see e.g. Foster, 1901; Siegel, 2005). During sleep, the organism is in a state of reduced responsiveness to external stimuli which is accompanied by a loss of consciousness. Sleep deprivation and even sleep disruption typically result in severe cognitive constraints (e.g. Killgore, 2010) and animals deprived from sleep for a longer period eventually die (Rechtschaffen & Bergmann, 1995). Thus, even though sleep might seem to be a quite uninteresting state from the outside, one should not underestimate its importance for human and animal organisms. Over the years sleep has been associated with different functions, as the restoration of energy resources and repairing of cell tissue (Mamelak, 1997), thermoregulation (Rechtschaffen & Bergmann, 1995), metabolic regulations (Knutson, Spiegel, Penev, & Van Cauter, 2007), and immune functions (Lange, Dimitrov, & Born, 2010). However, despite of their importance, these functions might arguably also be achieved without the loss of consciousness that is affiliated with sleep. Indeed, from an evolutionary perspective it seems quite odd and inadaptive to expose oneself to all kinds of threats during a very long period of inattention. This led researchers to, besides peripheral functions, consider sleep's importance for brain functions. Research has specifically identified a relationship between sleep and memory and over the past decades there have been numerous studies on the beneficial effects of sleep on memory (sleep effect; for an overview, see Diekelmann et al., 2009; Rasch & Born, 2013). In particular, sleep has been associated with memory consolidation, stabilizing memories and, thus, leading to less time-dependent forgetting (Diekelmann & Born, 2010).

Consolidation processes

The memory process is typically divided into three sub-processes. After encoding of the learning material memories are supposed to undergo a period of consolidation, before they are to be retrieved after a certain period of time. The importance of consolidation processes for later successfull retrieval arises

from the fact that memories remain labile after encoding before they gradually become consolidated and transferred to a more stable state (Alberini, Milekic, & Tronel, 2006; Squire & Alvarez, 1995). First evidence for consolidation processes comes from Müller and Pilzecker (1900), who assumed that after the study of a list of syllables, stabilizing processes are activated and continue to be at work with decreasing intensity over a certain period of time. Further studies on the consolidation hypothesis have concentrated on inducing a serious intervention directly after the encoding of the learning material. They show that different psychological, pharmacological, or electrophysiological manipulations after aquisition of the learning material can reliably impair or benefit memories (e.g. McGaugh, 2000; Wixted, 2004). Critically, and in support of the consolidation hypothesis, such interventions are only effective in a sensible period of time immediately after encoding, as consolidation processes are supposed to be more or less completed after a certain time period (McGaugh & Gold, 1976). Newer studies even support the idea of a consolidation process that consists of different steps that gradually stabilize memories over a longer period of time interspersed with intermittend reactivation. In support of this idea, some studies have emphasized an important role of reactivation by a reminder or active retrieval for memories. Such reactivation is supposed to lead memories back into labile state which enables subsequent updating and further stabilization of memories. This has been referred to as reconsolidation (see e.g. Hupbach, Gomez, Hardt, & Nadel, 2007; Nader & Einarsson, 2010; Nader & Hardt, 2009; Sara, 2000).

The sleep effect - empirical and theoretical background

In a classic study Jenkins and Dallenbach (1924) investigated effects of sleep after encoding on memory. Using senseless syllables they compared memory after retention intervals of varying length. Participants spent these intervals either awake or asleep. Results indicated, for the first time, evidence for sleep-associated memory consolidation. Indeed, number of correctly recalled syllables was reliably higher after sleep versus wake delays. This pattern of results has later on been

replicated numerous times (Barrett & Ekstrand, 1972; Benson & Feinberg, 1977; Gais et al., 2006) using different learning material like word lists (Ficca, Lombardo, Rossi, & Salzarulo, 2000; Lahl, Wispel, Willigens, & Pietrowsky, 2008), paired associates (Barrett & Ekstrand, 1972; Plihal & Born, 1997), or spatial information (Talamini, Nieuwenhuis, Takashima, & Jensen, 2008). Thus, it is now beyond dispute that sleep can beneficially influence memory (see Diekelmann et al., 2009, but see Vertes, 2004). According to Jenkins and Dallenbach (1924) these beneficial effects are caused by sleep sheltering memories from detrimental effects of interfering information. During wake delays retroactive interference builds up through the ongoing aquisition of new learning material. As sleep is an "offline" state where memories are protected from such additional information, memories are preserved and less prone to time-dependent forgetting. Hence, Jenkins and Dallenbach (1924) viewed sleep as a rather passive state were memories are locked from outside influences, therefore being better recalled. However, research of the past decades has revealed evidence that calls this passive-protection hypothesis into question (see e.g. Ellenbogen, Payne, & Stickgold, 2006). The next paragraphs will give an overview over recent research supporting an active role of sleep as a beneficial factor for the consolidation of memories.

Not only declarative memory contents have been found to profit from periods of sleep. Several studies have even found evidence for a beneficial effect of sleep on procedural memories such as for motor sequences (Cohen, Pascual-Leone, Press, & Robertson, 2005; Korman et al., 2007), in serial reaction-time tasks (Fischer, Drosopoulos, Tsen, & Born, 2006), but even performance in visual-discrimination tasks could reliably profit from sleep (Gais, Plihal, Wagner, & Born, 2000; Stickgold, Whidbee, Schirmer, Patel, & Hobson, 2000). Plihal and Born (1997) investigated how different sleep stages influence different types of memories. Sleep is characterized by two core sleep stages that reoccur in a cyclic manner during a night of sleep: REM (rapid-eye-movement) sleep and SWS (slow-wave sleep). SWS is characterised by slow, high-amplitude waves in the EEG, whereas REM sleep results in fast, low-amplitude oscillatory activity. In human beings, periods of SWS are more pronounced during the first half of the sleep period and than decrease in intensity and duration, while periods of REM sleep are first less distinct but become

predominant during the second half of the sleep period. Using the early-late sleep comparison, Plihal and Born (1997) investigated how these different sleep stages influence memory consolidation. Participants in their study were asked to either study paired associates (declarative-memory task) or to engage in a mirror-tracing task (procedural-memory task). Subsequently, in the early-sleep condition, they either stayed awake or slept during the first half of the night and slept during the second half of the night. In contrast, in the late-sleep condition, they slept during the first half of the night but stayed awake or slept during the second half of the night. Results revealed that SWS-rich early sleep mainly benefited declarative memory contents while REM sleep rich late sleep mainly served performance in the procedural task. This differentiation has resulted in the dual-process theory which states that SWS benefits preferentially declarative memories while REM sleep benefits rather procedural memories (see also Daurat, Terrier, Foret, & Tiberge, 2007; Drosopoulos, Wagner, & Born, 2005).

The dual-process theory supports the idea of an active process during sleep, benefiting memory consolidation. In contrast to assumptions made by Jenkins and Dallenbach (1924), findings by Plihal and Born (1997) cannot be explained with sleep acting as a passive shelter for memories. If memories were passively protected by sleep, only the duration of the period of protection would matter but not which sleep stages are predominantly present during this time interval. Thus, evidence from the early-late sleep comparison clearly suggests there to be an active role of sleep in consolidation processes. Still, some research has called the strict dichotomy of the dual-process theory into question, reporting evidence for performance in non-declarative tasks benefiting from SWS (Gais et al., 2000; Huber, Ghilardi, Massimini, & Tononi, 2004) and performance in declarative tasks relying on REM sleep (Tilley & Empson, 1978; Empson & Clarke, 1970). Other research has focused more on the importance of the cyclic succession of both REM sleep and SWS. The sequential hypothesis suggests that if the whole cyclic sleep sequence is intact, memories can profit from sleep (e.g. Ficca et al., 2000; Ficca & Salzarulo, 2004; Giuditta et al., 1995). Thus there are different accounts on the relationship between sleep stages and consolidation processes. Overall, recent research has focused rather on in-depth analyses of electrophysiological correlates

of sleep-associated memory consolidation than on the distinction between different sleep stages. Such correlates include general slow-wave activity (<1Hz) and increased numbers and density of sleep spindles (e.g. Eschenko, Ramadan, Mölle, Born, & Sara, 2008; Schabus et al., 2004; Siapas & Wilson, 1998). Thus, in sum, evidence supports the core assumptions of the dual-process theory but newer research shows that the electrophysiological correlates underlying sleep-associated memory consolidation are more or less pronounced during different sleep stages; i.e. future research should be directed to such correlates rather than a strictly separated analysis of different sleep stages (see Diekelmann et al., 2009).

Boundary conditions and characteristics of the sleep effect

Over the years many studies have investigated boundary conditions of the sleep effect. A key question in this research is whether sleep can be beneficial for certain types memories and less so for other types, i.e. if sleep selectively enhances certain memories compared to others. In a study by Wilhelm et al. (2011) participants were asked to learn paired associates and subsequently one group was informed about the future relevance of this study material while the other group did not receive such information. Results showed that sleep-related benefits occured mainly if item material was of future relevance (for similar results, see Van Dongen, Thielen, Takashima, Barth, & Fernandez, 2012). Other studies have investigated the influence of sleep on emotional versus neutral learning material. Typically, emotional material leads to an increased sleep effect compared to neutral material (e.g. Payne et al., 2008; Wagner, Hallschmid, Rasch, & Born, 2006). Another interesting research question is whether memory strength might act as a modulating factor of the sleep effect. In a study by Drosopoulos et al. (2007) memory strength of paired associates was manipulated before participants either slept or stayed awake for a period of time. Using the anticipation-plus-study method (repeated testing followed by retrieval), item material was either strengthened to a lerning criterion of 60% (weak) correctly recalled items or a lerning criterion of 90% (strong) correctly recalled items during

the learning phase. Results after wake respectively sleep delay showed that mainly weak paired associates could profit from sleep-associted memory consolidation. Thus, there seem to be some modulating factors of the sleep effect, as sleep selectively enhances certain types of memories. This again speaks in favor of sleep actively supporting consolidation processes as a simple passive shelter would not lead to selective benefits for certain memories.

In a recent study by Abel and Bäuml (2012) retrieval-induced forgetting (RIF; Anderson et al., 1994) was investigated after 12-h wake versus 12-h sleep delay. Results showed that there was a significant sleep effect for unpracticed control items (NRP) while no such effect was observable for items that had been subject to retrieval practice (RP+) or unpracticed items related to them (RP-). Thus, though one cannot draw firm conclusions from this single experiment, this indicates that sleep effects can be reduced after retrieval practice. However, it remains unclear why even unpracticed items of practiced categories (RP-) did not show sleep effects. Overall, this offers hints about a possible relationship between the testing effect and sleep. The results of Drosopoulos et al. (2007) give insights about how differences in strength levels of the initial learning materials might influence the size of the sleep effect. Going beyond these findings by Abel and Bäuml (2012) even include information about possible differences in sleep effects between restudied items and items subjected to retrieval practice. So this poses the question whether memory strength itself can moderate the sleep effect or if it rather is type of practice that influences effects of sleep-associated memory consolidation.

One might ask if the delay between learning and sleep or the duration of sleep plays a role for the beneficial mnemonic effects. Gais et al. (2006) found that sleep can reliably reduce time-dependent forgetting if it is applied within 3 hours after learning and less if it follows only after more than 10 hours (see also Benson & Feinberg, 1977). Thus, in contrast to timing of typical learning occasions, e.g. in schools, sleep-associated memory consolidation seems to depend on sleep following learning in a certain time window. Regarding the duration of sleep, several studies have investigated whether even shorter periods of sleep can lead to a significant sleep effect. In these studies participants were typically asked to take a nap during the day after studying certain learning material. Results show that even shorter

naps can lead to sleep-associated memory consolidation (e.g. Tucker & Fishbein, 2008; Tucker et al., 2006). And even an ultra-short nap of 6 minutes has been found to be sufficient (Lahl et al., 2008), though the effect is more pronounced after longer naps. So sleep periods can even be quite short to result in mnemonic benefits.

Similarly another interesting question is how long sleep effects can persist over time. Most studies on the sleep effect have asked participants to retrieve the learning material right after the sleep manipulation. However, only few have investigated how long the delay between learning and final retrieval may be for benefits of sleep to persist. Results of such studies show that the sleep effect is found even if this delay is exceeded to 24 hours (Talamini et al., 2008) or even 48 hours (Gais et al., 2006). In fact, recall after a night of sleep and a subsequent 12-h wake delay has proved to be enhanced compared to after just a 12-h wake delay (Tucker, Tang, Uzoh, & Stickgold, 2011). This again shows that benficial processes are active during sleep and that memories can effectively be protected from time-dependent forgetting, even under conditions where duration of wake intervals is controlled. The most astonishing, and yet to be replicated, finding comes from a study by Wagner et al. (2006) who found sleep effects for emotional text material to persist after a period of four years. Overall, empirical evidence shows that sleep effects can last for a long time which underscores the importance and impact of sleep in every-day learning.

Another line of research has investigated the relationship between sleep and retroactive-interference effects. One might argue that sleep-associated memory consolidation should stabilize memories to such a degree that they become less susceptible to detrimental effects of interfering lerning material. In a study by Ellenbogen, Hulbert, et al. (2006) participants were asked to initially learn paired associates and then either stay awake or sleep during a 12-h interval. While in the interference condition participants were than asked to study additional interfering paired-associate material, the final test was administered without additional learning after the delay in the no-interference condition. Results suggest that sleep protects memory from retroactive interference as interference effects were significantly higher after wake than after sleep delay (for similar results, see

Ellenbogen, Hulbert, Jiang, & Stickgold, 2009). These studies clearly indicate that sleep might reduce susceptibility to retroactive interference and speak, therefore, in favor of active consolidation processes leading to sleep-related mnemonic benefits. In fact, if sleep would only provide a passive shelter from interference accumulating during wake delays, such results would be implausible, as sleep should not protect memories from interference occuring after the sleep interval (Ellenbogen, Payne, & Stickgold, 2006). However, there is even contrasting evidence, denying that sleep protects memories from retroactive interference (e.g. Deliens et al., 2013).

Reactivation of memories during sleep

So it is now beyond dispute that sleep can benefit consolidation processes in both declarative and procedural memory systems. Moreover, in contrast to prior assumptions, sleep does not only passively shelter memories from retroactive interference that builds up during wake periods. In fact, there is plenty of evidence supporting the idea of active beneficial processes that foster sleep-associated memory consolidation (see above). Recent research even goes beyond this prior work as it suggests that memories are reactivated during sleep, leading to a stabilization of memories. In fact, studies on rodents convincingly support the idea of memories being reactivated during periods of sleep following initial learning. Wilson and McNaughton (1994) showed that similar brain cells were simultaneously activated during a spatial-learning task and subsequent sleep. Later studies have even investigated the issue in human beings. In a fMRI study by Peigneux et al. (2004), participants were asked to learn routes in a virtual town. Results show that the same hippocampal brain areas that were active during the initial spatial learning task were activated during subsequent SWS. Interestingly, Peigneux et al. (2004) could link this reactivation to later recall performance in a final test, as hippocampal activity during SWS was positively correlated to the improvement of performance in route retrieval on the next day. Thus, studied information seems to be reactivated during periods of sleep and such reactivation processes can be associated with later recall performance.

Only few studies have experimentally manipulated such reactivation during

sleep to investigate its beneficial influence on memory recall. In a study by Rasch et al. (2007) participants learned a visuospatial object location task. Simultaneously an odor (the smell of a rose) was administered to establish a distinct cue linked to the learning material. During subsequent SWS either this odor cue or an odorless vehicle was applied repeatedly (in an alternating 30s off / 30s on cycle) to reactivate learned object locations. Recall test after the sleep period clearly showed that memory was enhanced when odor cues had been applied compared to when just odorless vehicles had been used. Additionally results showed that reactivation processes were directly linked to SWS. However, no effect of reactivation was found when odor cues were applied during REM sleep or in a wake period prior to sleep. Moreover, no effects were evident when odor was not present during learning but only during SWS, underscoring the importance of binding of the learning material to the reactivation cue. In a similar study by Rudoy, Voss, Westerberg, and Paller (2009) auditory cues were used instead to reactivate initially learned material after having paired each learned item with an individual cue. Results show that memories can be selectively reactivated during sleep. If the individual cue, that had been linked to an item, was applied during subsequent sleep, recall after sleep was significantly better than if other unrelated cues were applied. Overall, these results speak in favor of memories being reactivated during sleep, which results in enhanced consolidation processes and better recall performance.

Recent research has focused on the phenomenon of memory consolidation and its relationship to neuronal activity during and after periods of sleep. According to a two-stage model, the consolidation process is supposed to consist of two subtypes. Synaptic consolidation refers to strengthening at synaptic levels in localized neuronal circuits which can take place both in wake and sleep conditions and is initiated shortly after memory aquisition. In contrast, system consolidation is supposed to preferentially take place "offline" during sleep and might last for much longer periods. This process might be fostered by repeated reactivations of newly encoded memory representations. Empirical evidence (see above) points out that reactivations occur mainly during SWS and might mediate the redistribution of the temporarily stored representations to long-term storage where they become integrated into preexisting long-term memory networks. Through slow oscillations

and in conjunction with so-called sharp-wave ripples and thalamo-cortical spindles, repeated reactivation of hippocampal memory representations might lead to enduring changes in cortical areas. Thus, reactivation might induce an integration of memories into long-therm storage and a reorganization of memory representations. This results in stabilized memories and less time-dependent forgetting (see e.g. Rasch & Born, 2013; Born & Wilhelm, 2012). In an fMRI study by Gais et al. (2007) evidence for system consolidation was brought forward. While shortly after memory encoding neuronal activity was most pronounced in hippocampal areas, after several months the center of cerebral activity had shifted to the medial prefrontal cortex, if participants had slept after initial memory aquisition. No such shift was observed if participants had been deprived from sleep for 24 hours after the learning phase (for similar results, see Takashima et al., 2006). Thus, sleep seems to be a necessary state initiating reactivation of memories leading to active system consolidation, i.e. the integration of newer memories into existing memory networks at a cortical cerebral level.

Summary

Beneficial effects of sleep on memory representation have been investigated systematically for a long period of time. Ever since the seminal study by Jenkins and Dallenbach (1924) it has been found in a variety of experimental designs using different types of memories (Gais et al., 2000; Plihal & Born, 1997; Stickgold et al., 2000; for an overview, see Diekelmann et al., 2009; Rasch & Born, 2013). Refered to as passively protecting memories from interference, sleep has over time been found to lead to an active process benefitting the consolidation of memories (e.g. Plihal & Born, 1997). Recent studies suggest even that memories might be reactivated during sleep, leading to better recall after sleep versus wake delay (Peigneux et al., 2004; Rasch et al., 2007). Moreover studies on the neuronal processes active during sleep can shed light on how such processes might lead to updating of existing memory networks with new mnemonic information, underscoring the importance of sleep periods for a functioning memory system.

1.4 Goals of The Present Work

Both testing and sleep have been repeatedly found to be effective tools in promoting long-term retention of mnemonic information (Diekelmann et al., 2009; Roediger & Karpicke, 2006a). Besides being effective in reducing time-dependent forgetting, both retrieval-practice (e.g. Halamish & Bjork, 2011) and sleep (e.g. Ellenbogen, Payne, & Stickgold, 2006; but see Deliens et al., 2013) have been found to protect memories against effects of retroactive interference. Thus far, testing effects have been investigated over delays with varying duration, several studies suggesting that they increase with length of the retention interval (Roediger & Karpicke, 2006b; Toppino & Cohen, 2009; Wheeler et al., 2003). As longer delays typically include at least one sleep interval one might assume that sleep-associated memory consolidation plays an active role in enhancing testing effects. Alternatively, testing effects might be reduced or not affected at all by sleep versus wake intervals. However, if delay intervals include sleep or wake periods has never been systematically controlled for in prior investigations on the testing effect. Taking recent studies on the testing effect and on sleep-associated memory consolidation into consideration, the present work aims to disentangle their mnemonic effects and to investigate potential interactions. Results might contribute to better understanding of theoretical backgrounds and boundary conditions of the testing effect and the sleep effect (see below). Additionally, they might provide knowledge about how to best use two of the most effective tools for memory enhancement, maximizing long-therm retention of aquired information. As a result, they might promote for the practical applience of knowledge about these beneficial mnemonic effects in e.g. educational contexts (see above).

To investigate influences of sleep on the testing effect the present work bases on prior studies on the testing effect and sleep-associated memory consolidation. Using different kinds of learning material (categorized item material, paired associates), testing effects after shorter delays (20 minutes) were contrasted with testing effects after longer delays (12 hours). Longer delays included either wake or sleep intervals. Additionally, memory retention was investigated both in the presence and absence of retroactive interference during final test. In a series of 4

experiments the present work gradually analyzed how these factors affect memory recall. Based on a study by (Abel & Bäuml, 2012), the first two experiments used categorized item material and asked participants, after initial study, to either restudy or practice retrieval of the items before a sleep or wake delay was employed. In contrast, in a second section, Experiments 3 and 4 used semantically unrelated paired associates as learning material to relate to classical work on both the testing effect and the sleep effect and to explore how variations in initial success rates during retrieval practice influence final memory retention. While Experiments 1 and 3 focused on a possible interaction between testing effect and sleep, Experiments 2 and 4 were used to investigate effects of retroactive interference, introducing additional interfering item material after the sleep or wake delay

Regarding the testing effect, predicted influences of sleep depend on the theoretical framework (see above). According to the distribution-based bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011) some predictions can be made (see Figure 2). Following presumptions made by this model sleep delays should result in a reduction of the testing effect as it would predict sleep to differentially influence final memory recall of items after restudy versus retrieval-practice. Items that were successfully retrieved during the retrieval-practice phase are, according to the model, strenghened to a very high degree. Even sleep should lead to a strengthening of items, resulting in a shift to the right on the memory-strength axis. As successfully retrieved items are already high above recall threshold after wake delay, additional strengthening through sleep should not have any pronounced effect on later memory recall. In contrast, restudied items fall below recall threshold by a large margin over a longer delay. Thus, these items might profit from sleep-associated memory consolidation by a greater margin as sleep might help more of these items to cross recall threshold. Hence, compared to wake delay, sleep might diminish the gap in recall performance between restudied items and items subjected to retrieval practice; i.e. sleep versus wake delay might result in reduced testing effects (see Figure 2, lower panel). However, it is important to bear in mind that such predictions might only be made for a delay of a certain duration (approximately 12 hours). Extending delays to several days or even weeks

should lead to different predictions according to the model. After longer delays more of the restudied items initially strengthened by sleep would fall below recall threshold, while those that had been subject to retrieval practice still might be above threshold by a larger margin. Overall, this might lead to a reestablished testing effect as delay is extended even more.

According to the elaborative retrieval hypothesis (Carpenter & Delosh, 2006; Pyc & Rawson, 2009) testing effects are caused by processes triggered by effortful retrieval during the retrieval-practice phase. In contrast, restudy is supposed to lead to only passive and rather effortless reprocessing of initially learned information. Though the elaborative retrieval hypothesis does not directly make any predictions about sleep's influence on the testing effect there are still some conclusions that can be drawn from assumptions of this theoretical framework. There has been established a clear link between retrieval effort and the activation of additional semantic information (Carpenter, 2009; Pyc & Rawson, 2010, 2012) which is one explanation for benefits of retrieval practice at the final test. Sleep, on the other hand, has also been found to facilitate the activation of semantic networks around certain initially studied material. For instance, influences of sleep on false memories were investigated using the Deese-Roediger-McDermott paradigm (Roediger & McDermott, 1995). Results indicate that sleep does not only benefit recall for initially learned material but also for semantically related items (e.g. Darsaud et al., 2011). Moreover, going beyond similar profits of sleep for actually learned and semantically related information, some studies reported even greater benefits for the related but initially unstudied items (McKeon, Pace-Schott, & Spencer, 2012; Payne et al., 2009). In addition to these findings sleep has been found to prime associative semantic networks in the remote-associates task (Cai, Mednick, Harrison, Kanday, & Mednick, 2009). Thus, according to the elaborative retrieval hypothesis there are no reasons to expect reduced testing effects after sleep. In fact, this hypothesis would possibly even predict greater testing effects, as both sleep and retrieval practice have been found to activate semantic networks. Thus, overall, this is in contrast with predictions made by the bifurcation model.

A third perspective is offered by the episodic context account (Karpicke et al., 2014). Like the elaborative retrieval hypothesis this theoretical framework offers

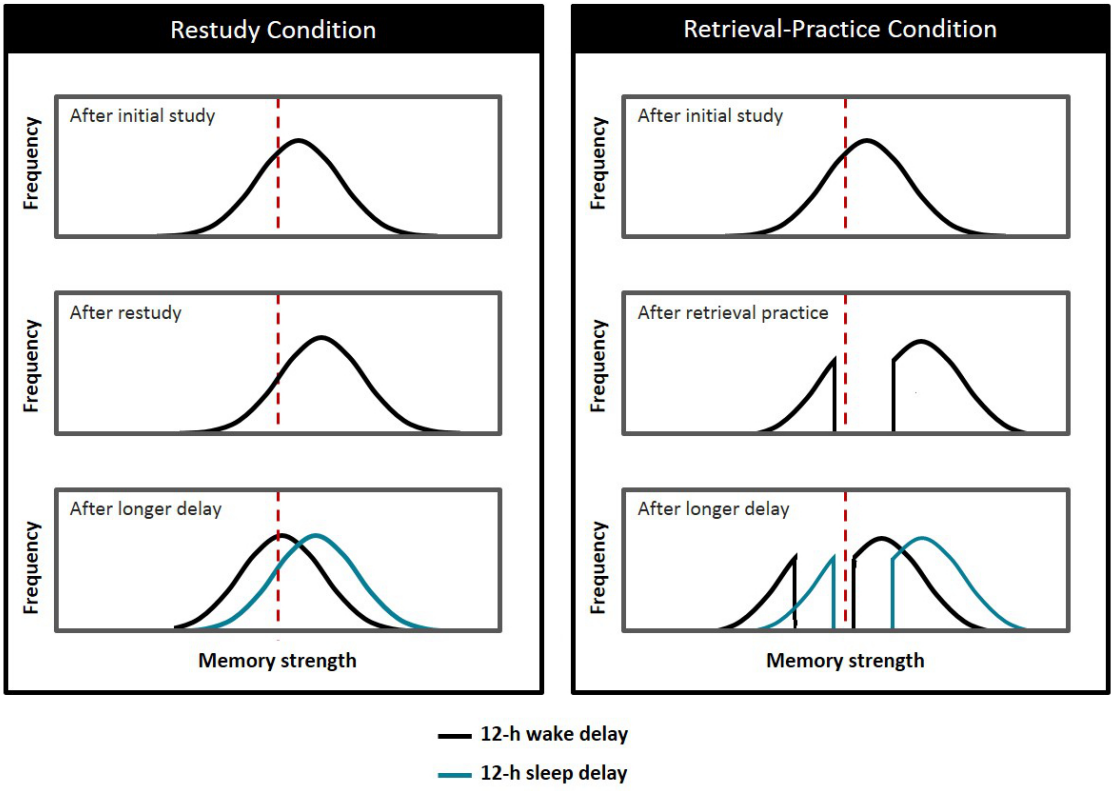

Figure 2: Memory strength distributions of two hypothetical sets of items, based on the bifurcation model (Kornell et al., 2011). The left column shows the restudy condition while the right column shows the retrieval-practice condition. In the top pair of panels memory strength after one initial study trial is illustrated; at this point the two distributions are identical. The second pair of panels shows how distributions are shifted after restudy or retrieval practice. While all restudied items gain memory strength, retrieval-practiced items become bifurcated. Successfully retrieved items are strengthened to a higher degree than restudied items, whereas nonretrieved items remain at the original strength level (see even Figure 1). The bottom pair of panels illustrates how distributions are shifted after wake delay (black curves) and sleep delay (blue curves), respectively; all items show the same amount of delay-induced forgetting and the same amount of sleep-induced strengthening. Red vertical dotted lines indicate recall threshold at the final test. Restudied items profit from sleep, because more of them cross the recall threshold after sleep. In contrast, items subjected to retrieval practice do not profit from sleep, because their strength level is still above threshold after wake delay.

an explanation for processes that lie behind the formation of the testing effect. Basically, following this account, retrieval practice and restudy are supposed to lead to a reinstatement of the original learning context and to an updating of the memory representation with information from the new temporal context during the practice phase. According to this view, both retrieval and restudy may induce context retrieval, but without intentional retrieval instructions, restudy might not rely as much on context retrieval as retrieval practice. This difference in context retrieval might induce a difference in the creation of unique context cues, and thus enhance long-term retention after retrieval practice more than after restudy. In fact, such unique context cues might help to reduce the size of the search set during final recall leading to benefits of retrieval practice over restudy. There is no direct evidence predicting influences of sleep on the testing effect deriving from this account. However, according to a study by Cairney et al. (2011), context effects might be primarily evident after wake delay (e.g. Godden & Baddeley, 1975) but less so after sleep delay. Basically, following this finding, one could assume that memory recall after sleep depends less on contextual cues. Additionally, studies on the sleep effect suggest that memories are being reactivated during sleep, leading to a better integration of such memories in existing mnemonic networks (Rasch et al., 2007). Overall, according to the episodic context account, the key advantage of retrieval practice over restudy is related to more adaptive cues by context reinstatement and integration of new contextual information during the practice phase. Thus, less reliance on such cues and a reactivation of memories during sleep might scale down the gap between retrieval practice and restudy during final test, i.e. sleep versus wake delays might reduce the testing effect. So, in contrast to the elaborative retrieval hypothesis, the episodic context account does not predict greater testing effects after sleep delay. In fact, testing effects are either not affected or decreased, according to this account.

Besides aiming to gather more information on the influence of sleep delay on the testing effect, an additional goal of the present work is to investigate the influence of retroactive interference on both the testing effect and sleep-associated memory consolidation. Previous work has identified the presence of retroactive interference as a moderating factor of beneficial effects of retrieval practice. In fact, testing

has been repeatedly found to insulate memories against retroactive interference compared to restudy, making memories less prone to the typical detrimental effects of interference induction (Abel & Bäuml, 2014; Potts & Shanks, 2012). This interesting beneficial feature of retrieval practice fits well into the bifurcation model, as mainly restudied items are supposed to suffer from interference effects but not items subjected to retrieval practice that are high above recall threshold (Halamish & Bjork, 2011). Thus, there is clear empirical and theoretical evidence in favor of the idea that testing can protect memories from the diminishing effects of retroactive interference. Regarding sleep's influence on interference effects, evidence is mixed. Inducing retroactive interference by asking participants to study additional interfering material after a sleep delay, Ellenbogen, Hulbert, et al. (2006) found sleep compared to wake intervals to protect initially studied material from interference effects. In a follow-up study they obtained similar results with a slightly modulated experimental design (Ellenbogen et al., 2009). However, contrasting evidence comes from a study by Deliens et al. (2013), who did not find sleep to protect memories from retroactive interference. In fact, results of this study even indicate the opposite: i.e. increased interference effects after a night of sleep compared to after a sleep-deprivation interval. Thus, whether sleep reduces memories' susceptibility to retroactive interference is an issue that needs to be further investigated. The present work aimed to tackle this question and is the first to investigate the influence of retrieval practice and sleep on interference effects in one study.

According to different theoretical frameworks of the testing effect there are diverging predictions regarding the outcome of the present work (see above). However, as mentioned above, the bifurcation model offers a strength-related explanation of the testing effect. In contrast, the elaborative retrieval account and the episodic context account are process-based accounts that try to explain cognitive processes resulting in the testing effect. Thus, the bifurcation model does not contradict predictions made by the other two accounts. Regarding influences of sleep, the bifurcation model would predict sleep to reduce the size of the testing effect, being beneficial mainly after restudy but less so after retrieval practice. The elaborative retrieval account and the episodic context account can only make

indirect predictions about if and how sleep migth influence the testing effect (see above). Consequently, the focus of this work will lie on predictions made by the bifurcation model. Regarding interference effects, on the basis of previous empirical evidence, retrieval practice should reduce susceptibility to retroactive interference. Whether even sleep has such an effect on memory is still unclear and has yet to be unraveled. Overall, the present work is the first to systematically investigate possible relationships between sleep and the testing effect. Results might offer insights about theoretical implications of the testing effect and the sleep effect promoting potential practical appliance of this knowledge in various contexts, i.e. educational settings.

Chapter 2

Sleep and the Testing Effect - Categorized Item Material

2.1 Experiment 1

The goal of Experiment 1 was to examine the effects of sleep and wake delay on item material that was either subject to restudy or to retrieval practice during the learning phase. For this purpose, subjects studied a semantically categorized item list in one initial study cycle. This type of item material was used in accordance with a study by Abel and Bäuml (2012) who found differential effects of sleep on restudied items and items subjected to retrieval practice. In the restudy condition, participants were asked to restudy the items of half of the categories once and the other half twice. In parallel, after the initial study cycle, subjects in the retrieval-practice condition were asked to actively retrieve the items of half of the categories once, the other half twice. After a 12-h delay of either diurnal wakefulness or nocturnal sleep subjects were asked to recall all items in a final memory test.

Based on previous findings on the testing effect (Hogan & Kintsch, 1971; Roediger & Karpicke, 2006b), after the wake delay, items that had been subjected to retrieval practice were expected to be recalled better than restudied items, both after one and after two practice cycles. The question of key interest in this experiment though was whether recall after sleep delay might yield a different pattern of results. Based on the the bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011), one would expect sleep to reduce the testing effect as items that were successfully retrieved during retrieval practice (as opposed to nonretrieved items) would already be strengthened to a very high degree. While restudied items would profit from sleep-associated memory consolidation, successfully retrieved items would already fall above recall threshold by a large margin in the absence of sleep, thus effects of sleep-induced strengthening on recall performance should be small or negligible for these items. Consistently, the advantage of retrieval practice over restudy at final recall would be reduced after sleep delay (see above).

Process-based accounts (elaborative retrieval account and episodic context account) make more indirect and vague predictions about how sleep might influence the testing effect. Based on the elaborative retrieval hypothesis (Carpenter, 2009;

Pyc & Rawson, 2010), retrieval practice would be expected to result in more elaborative processing than restudy. Indeed, according to this hypothesis, the attempt of retrieving a certain item might be accompanied by the activation of semantically related information (Carpenter, 2009; Pyc & Rawson, 2010, 2012). Together with studies proposing an active role of sleep in the buildup of semantic associative networks (Cai et al., 2009; McKeon et al., 2012; Payne et al., 2009) one could expect sleep to maintain or even increase the testing effect. Finally, based on the episodic context account (Karpicke et al., 2014), one might expect that sleep either leads to maintained or, alternatively, reduced testing effects. In fact, according to this account, the benefits of retrieval practice over restudy might be more evident the more recall during final test relies on contextual cues. Consistently, sleep might lead to less reliance on contextual cues (Cairney et al., 2011) and memories might be reactivated during sleep (Rasch et al., 2007), resulting in reduced testing effects.

Method

Participants

A sample of 216 healthy students from Regensburg University took part in the experiment voluntarily or in return for financial reimbursement ($M = 22.4$ years; range 18-30 years; 55 male). At the beginning of the experiment, all participants were asked to complete a screening questionnaire to make sure they did not suffer from neurological or psychiatric disorders, in particular sleep disorders. Subjects taking psychoactive medication or drugs affecting the central nervous system were excluded prior to randomization (see Ellenbogen, Hulbert, et al., 2006). Subjects in the 12-h sleep condition were instructed to try to sleep regularly during the night between sessions, while subjects in the 12-h wake condition were instructed not to take naps during the day. Every subject was asked to refrain from drinking alkoholic beverages between sessions. All participants were native German speakers and were distributed equally across conditions ($n = 36$ in each of the six conditions). Comparisons between conditions regarding subjects' age, habitual sleep duration, subjective ratings of sleep quality, IQ (estimated via

speed of cognitive processing; Oswald & Roth, 1987), and ratings on the Epworth Sleepiness Scale (Johns, 1991) did not reveal any differences (all $ps > .10$).

Material

Using items from different word norms (Scheithe & Bäuml, 1995; Van Overschelde, Rawson, & Dunlosky, 2004) an item list was constructed consisting of 24 concrete German nouns from four different semantic categories (six items per category). Whithin each category, all items had unique initial letters to establish distinct item cues. The items of two the categories were repeated once during retrieval practice or restudy, whereas the items of the remaining two categories were repeated twice; categories were distributed across practice levels in a balanced manner, i.e. all categories were used equally often for the low and the high practice level to prevent item material effects.

Design

The experiment had a $2 \times 2 \times 2$ mixed-factorial design with the between-subjects factors of TYPE OF PRACTICE (restudy, retrieval practice) and DELAY (12-h wake, 12-h sleep), and the within-subjects factor of PRACTICE LEVEL (low, high). After the initial study cycle, half of the subjects were asked to restudy the item list (restudy condition), while the other half was asked to engage in retrieval practice (retrieval-practice condition). There were two practice levels with differing numbers of practice cycles. For one half of the initially studied categories, subjects were given one restudy or retrieval-practice cycle (low practice level); for the other half, they were given two restudy or retrieval- practice cycles (high practice level).

Additionally, across delay conditions, one half of the subjects stayed awake during the delay between learning phase and test, while for the other half this interval was filled with sleep. In the 12-h wake condition, participants studied and practiced the items at 9 a.m., and final test was conducted at 9 p.m., after 12 hours of diurnal wakefulness; in contrast, in the 12-h sleep condition, participants studied and practiced the items at 9 p.m., and took the final test at 9 a.m., after

one night of nocturnal sleep (see Figure 3A for an illustration of the experimental design; for similar designs, see Abel & Bäuml, 2013a; Payne et al., 2008; Scullin & McDaniel, 2010). Because study and test sessions took place at different times of day across delay conditions, an additional short-delay condition was included to control for potential circadian effects. Half of the subjects in this condition participated at 9 a.m., the other half at 9 p.m., with only a short delay of 12 min between learning phase and test.

Procedure

Study and Practice Phase. At the beginning of the learnig phase, items were presented successively and together with their corresponding category labels in a random order, at a presentation rate of 3 sec per item. After one initial study cycle for all 24 items, there were additional practice cycles that comprised either restudy or retrieval-practice trials, depending on practice condition. In the restudy condition, the intact items and their category labels were reexposed at a 3 sec rate. Items of two of the four categories were restudied once (SS = study-study), whereas items from the other two categories were restudied twice (SSS = study-study-study). Semantic categories were distributed over the two practice levels in a balanced manner. In the retrieval condition, subjects were provided with the items' category labels and word stems for up to 5 sec each, and were asked to recall the corresponding items. Mean response time (which equals overall processing time) across all retrieval-practice trials was $M = 2.1$ sec ($SD = .27$). Retrieval of two categories' items was practiced once (ST = study-test), whereas retrieval of the other two categories' items was practiced twice (STT = study-test-test; see Figure 3B for an illustration of the two types of practice). In both the restudy and retrieval conditions, order of items was random with the restriction that items of the same category were never restudied or retrieved consecutively.

The learning phase was followed by a distractor phase of 12 min, during which participants engaged in several unrelated cognitive tasks. After this phase, subjects in the short-delay control conditions completed the final recall test. In contrast,

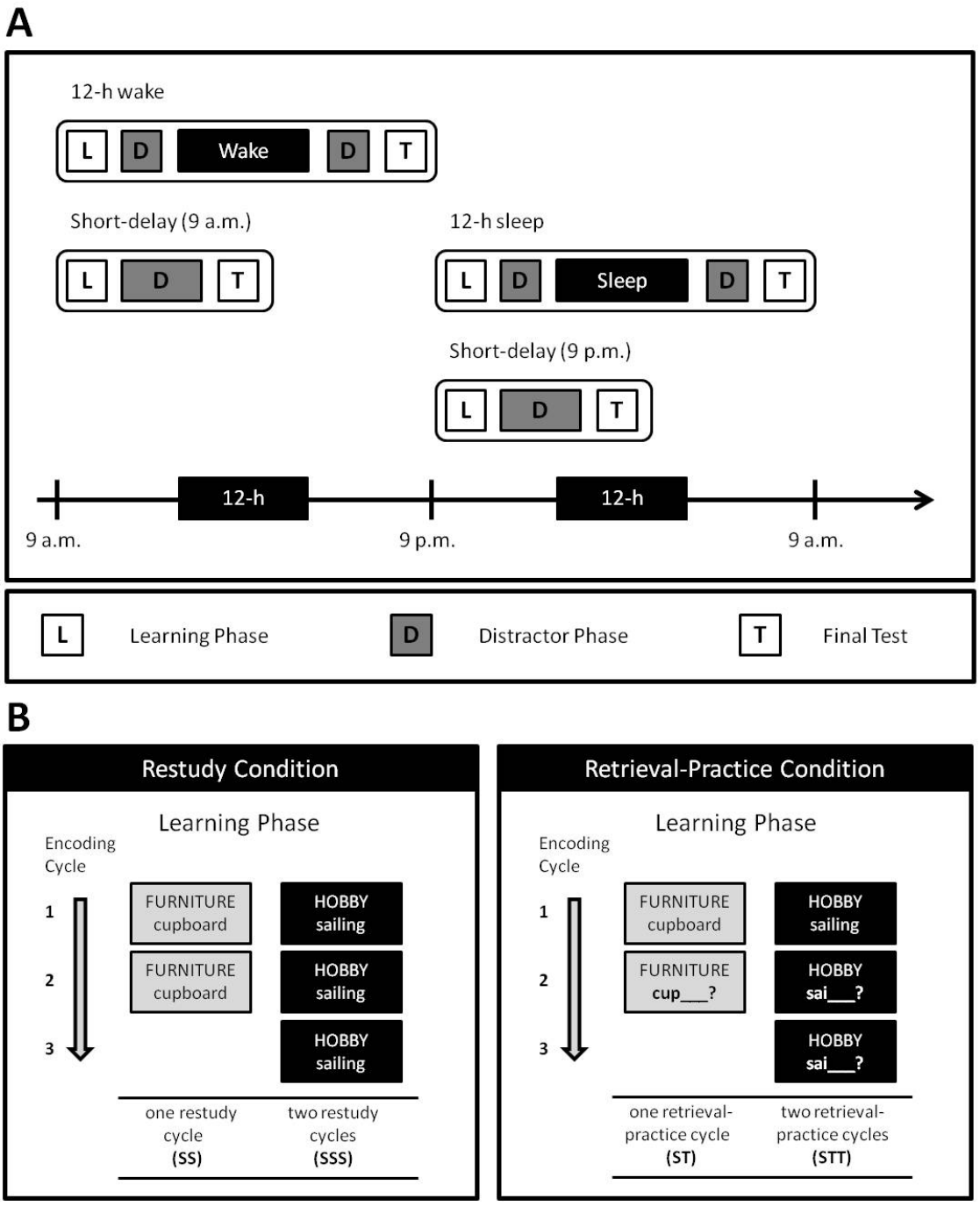

Figure 3: **(A)** Illustration of delay conditions employed in *Experiment 1*: In the 12-h wake conditon, the learning of the material took place at 9 a.m., before participants returned to the laboratory for the final test after a 12-h delay of diurnal wakefulness. In the 12-h sleep condition, the learning of the material took place at 9 p.m., and memory was tested after a 12-h delay comprising nocturnal sleep. The short-delay condition took place at 9 a.m., respectively 9 p.m. Here the final test was already administered after a short delay of 12 minutes. **(B)** Illustration of the two types of practice employed in *Experiment 1*: In the restudy condition, after the initial study cycle, there were one or two restudy cycles. In contrast, in the retrieval-practice condition initial study was followed by one or two retrieval-practice cycles.

subjects in the 12-h delay conditions were dismissed from the first session after having completed half of the distractor phase (6 min). After a delay of 12 hours that was either spent awake during the day or filled with normal nighttime sleep, subjects in the 12-h delay conditions returned to the laboratory and completed the second half of the distractor phase (6 min) before completing the same final recall test. Concerning compliance with instructions, subjects in the 12-h sleep condition reported to have slept regularly during the experimental night ($M = 7.9$ hrs; $SD = .83$). Time spent asleep during the experimental night did not differ from habitual sleep time of participants in this condition ($p > .10$). Subjects in the 12-h wake condition reported not to have taken any naps during the day. None of the subjects reported alcohol intake between sessions.

Test Phase. During final test, participants were presented with the category labels and initial letters of all 24 studied items for 7 sec each, and were asked to recall the appropriate item. Items from the same category were tested consecutively. Order of categories and order of items within categories was random.

Results

Success Rates during Retrieval-Practice Cycles

A 2 x 2 x 2 ANOVA with the factors of time of day (9 a.m., 9 p.m.), retrieval practice (ST, STT), and delay (short delay, 12-h delay) showed that retrieval success was higher after two than after one retrieval-practice cycle (95.3% vs. 89.8%), $F(1, 104) = 27.36$, $MSE = 56.61$, $p < .001$, $\eta^2 = .21$. There were no other main effects and no interactions, $ps > .10$.

Final Test (12-h Delay Conditions)

Figure 4 shows recall performance after the 12-h delay. A 2 x 2 x 2 ANOVA with the factors of type of practice (restudy, retrieval practice), delay (12-h wake, 12-h sleep), and practice level (low, high) revealed significant main effects of type of practice, $F(1, 140) = 12.13$, $MSE = 274.67$, $p = .001$,

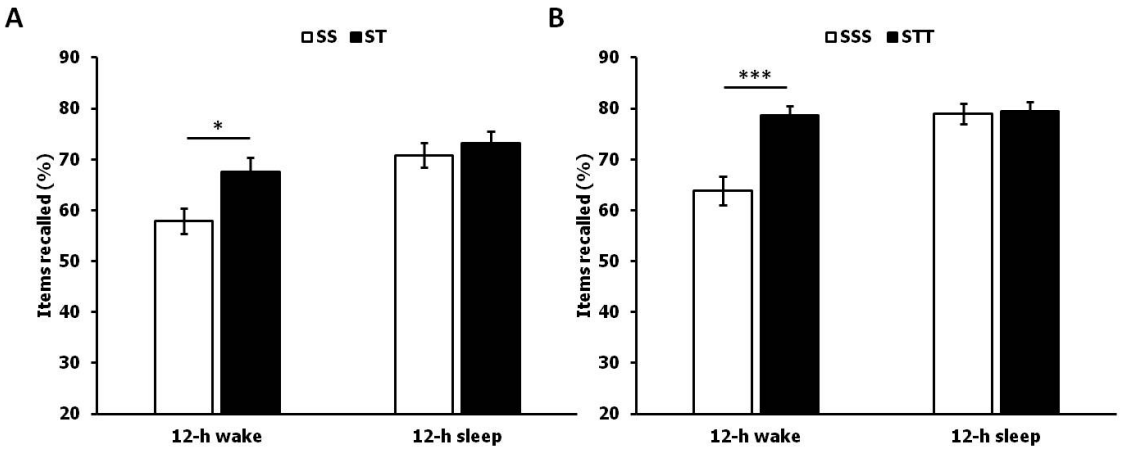

Figure 4: Mean recall performance on the final test of *Experiment 1* as a function of delay (12-h wake, 12-h sleep) and type of practice (restudy, retrieval practice), separately for the low (A) and the high (B) practice level. Condition labels indicate study (S) and retrieval practice (T) cycles. Error bars represent standard errors; $^{*}p < .05$; $^{***}p < .001$.

$\eta^2 = .08$, DELAY, $F(1, 140) = 19.22$, $MSE = 274.67$, $p < .001$, $\eta^2 = .12$, and PRACTICE LEVEL, $F(1, 140) = 34.79$, $MSE = 128.85$, $p < .001$, $\eta^2 = .20$. The main effect of TYPE OF PRACTICE reflects that there was overall higher recall in the retrieval-practice condition than the restudy condition (74.7% vs. 67.9%), whereas the main effect of DELAY shows that recall rates were overall higher in the 12-h sleep condition than in the 12-h wake condition (75.6% vs. 67.0%); the main effect of PRACTICE LEVEL indicates better recall after the high than the low practice level (75.2% vs. 67.3%). More important, while all other interactions were nonsignificant (all $ps > .05$), there was a reliable two-way interaction between TYPE OF PRACTICE and DELAY, $F(1, 140) = 7.75$, $MSE = 275.67$, $p = .006$, $\eta^2 = .05$, indicating that the difference in memory performance after retrieval practice in comparison to restudy was modulated by delay condition.

Consistently, planned comparisons showed a significant testing effect after the 12-h wake delay, both for the lower practice level (57.9% vs. 67.6%), $t(70) = 2.53$, $p = .013$, $d = .60$, and the higher practice level (63.9% vs. 78.7%), $t(70) = 4.44$, $p < .001$, $d = 1.05$, whereas no reliable TE arose after the 12-h sleep delay, both for the lower practice level (70.8% vs. 73.1%), $t(70) = .67$, $p = .505$, $d = .16$, and

the higher practice level (78.9% vs. 79.4%), $t(70) = .87$, $p = .872$, $d = .04$.

Further planned comparisons revealed that there was an effect of sleep-associated memory consolidation for items that had been restudied, irrespective of whether they were restudied once (SS), $t(70) = 3.66$, $p < .001$, $d = .86$, or twice (SSS), $t(70) = 4.35$, $p < .001$, $d = 1.03$. However, there was no effect of sleep-associated memory consolidation for items that had been subject to retrieval practice once (ST), $t(70) = 1.50$, $p = .138$, $d = .35$, or twice (STT), $t(70) = .27$, $p = .789$, $d = .06$.

Circadian Control (Short-Delay Condition)

Table 1 shows mean recall levels after the short delay. A 2 x 2 x 2 ANOVA with the factors of type of practice (restudy, retrieval practice), time of day (9 a.m., 9 p.m.), and practice level (low, high) revealed a significant main effect of practice level, $F(1, 68) = 8.67$, $MSE = 121.26$, $p = .004$, $\eta^2 = .11$, with the items of the higher practice level being recalled better than the items of the lower practice level (81.4% vs. 76.0%). No other effects emerged, indicating that there was no testing effect after short retention interval and recall was unaffected by circadian effects, all $ps > .10$.

Table 1: Mean recall performance in the short-delay condition of *Experiment 1* as a function of time of day (9 a.m., 9 p.m.), type of practice (restudy, retrieval practice), and practice level (low, high). Condition labels indicate study (S) and retrieval-practice (T) cycles. Standard errors are displayed in parentheses.

	Restudy		Retrieval Practice	
	SS	**SSS**	**ST**	**STT**
9 a.m.	75.4 (3.5)	79.9 (3.4)	77.7 (2.7)	82.8 (3.3)
9 p.m.	73.9 (3.8)	78.1 (2.7)	77.3 (4.0)	83.3 (2.7)
Combined	74.5 (2.5)	79.8 (2.2)	77.5 (2.4)	83.1 (2.0)

Discussion

The results replicate prior testing effect studies (Carrier & Pashler, 1992; Hogan & Kintsch, 1971; Roediger & Karpicke, 2006b; Wheeler & Roediger, 1992). After the 12-h wake delay, there was a significant benefit for items that had been subject to retrieval practice over items that had been restudied during the learning phase. This benefit held regardless of number of practice cycles, i.e. there was a similar benefit of retrieval practice after one and after two practice cycles. In contrast, after the 12-h sleep delay the testing effect was reduced by a large margin. Thus, there was only a very small numerical, but far from statistically relevant, benefit for items that had been subject to retrieval practice over items that had been restudied during the learning phase. Again, this pattern was evident regardless of level of practice, as a testing effect did neither arise after one, nor after two practice cycles. This indicates that the type of delay between the learning phase and the test phase (12-h wake versus 12-h sleep) did considerably influence the presence or absence of a testing effect, with significant testing effects only being measurable after the wake delay and not after the sleep delay.

Theoretically, the present results are in line with the distribution-based bifurcation model of the testing effect (Halamish & Bjork, 2011; Kornell et al., 2011). According to this model, items that are successfully retrieved during the retrieval-practice phase are strengthened by a greater margin than items that are restudied. In contrast, items that are not successfully retrieved, approximately remain at the strength level they already had after the initial study cycle. Due to this bifurcated distribution, successfully retrieved items might still be well above recall threshold at final test after the 12-h wake delay. Thus, sleep-associated memory consolidation could strengthen these items further but without any measurable effect. Restudied items, on the other hand, might pass the recall threshold by a greater proportion after sleep delay than after wake delay, resulting in a reliable sleep effect in the final test (see Figure 2). As this sleep-induced benefit is only evident after restudy and not after retrieval practice, the present results indicate that testing effects are reduced after sleep. Similarly, regarding process models, the results might even be in line with the episodic context account of the

testing effect (Karpicke et al., 2014). As testing effects are reduced after sleep but not after wake delay, one could argue that sleep versus wake delay results in less reliance on contextual cues (Cairney et al., 2011) leading to a reduction of benefits of retrieval practice over restudy. Additionally, bearing in mind that sleep might reactivate previously learned information (Rasch et al., 2007), this reactivation might also contribute to a reduced gap between restudy and retrieval practice at final test. In contrast, the finding of reduced testing effects after sleep seems to be less consistent with the elaborative retrieval hypothesis of the testing effect (Carpenter, 2009; Jacoby, 1978; Pyc & Rawson, 2009). This account would expect sleep to foster processes leading not only to a maintained but possibly even to an increased testing effect.

The present experiment goes beyond prior work, as it demonstrates that sleep can reduce or even eliminate the testing effect. Critical to the results of Experiment 1 is that it might be argued that they arose due to a ceiling effect. Indeed, as recall levels after retrieval practice are overall very high, there might not be much room for further effects of sleep-associated memory consolidation. Experiment 2 was designed to examine this issue by reducing overall recall levels through the induction of post-delay retroactive interference.

2.2 Experiment 2

Based on the findings of Abel and Bäuml (2012), one can assume that effects of sleep-associated memory consolidation are reduced or even eliminated after retrieval practice. Experiment 1 extends this empirical work by directly comparing a restudy with a retrieval-practice condition. The results suggest that only restudied items can profit reasonably from sleep, thus there is a reduction of the testing effect after a 12-h sleep delay compared to a 12-h wake delay.

However, to draw any firm conclusions from this experiment might be premature. Retrieval practice is a very powerful tool that can significantly reduce time-dependent forgetting. Apparently, for subjects in Experiment 1, this worked very well as recall rates in the retrieval-practice condition are quite high even after a 12-h delay. A consequence of these near-ceiling rates might have been that effects of sleep were not measurable, particularly in the high practice condition. This might be a simple explanation for the observed absence of sleep-associated memory effects after retrieval practice. Therefore, aim of Experiment 2 was to replicate the results of Experiment 1 and to examine the issue under more difficult conditions. Higher difficulty at the final test might reduce overall recall rates, thus permitting a more sustained conclusion about the relationship between sleep and retrieval practice.

According to the bifurcation model, increased final-test difficulty would raise the recall threshold. This would elicit a larger testing effect, as particularly restudied items would fall below recall threshold leading to an increased gap between retrieval practice and restudy (see Figure 5). A way to increase difficulty at final test is to induce retroactice interference through study of related item material. Indeed, Halamish and Bjork (2011) used such a manipulation showing that it can result in a significant increment of the testing effect (for related results, see Abel & Bäuml, 2014; Potts & Shanks, 2012).

Experiment 2 was designed in a similar way as Experiment 1. The main difference between the two experiments was the induction of retroactive interference at test. After the 12-h delay interval an additional item list was

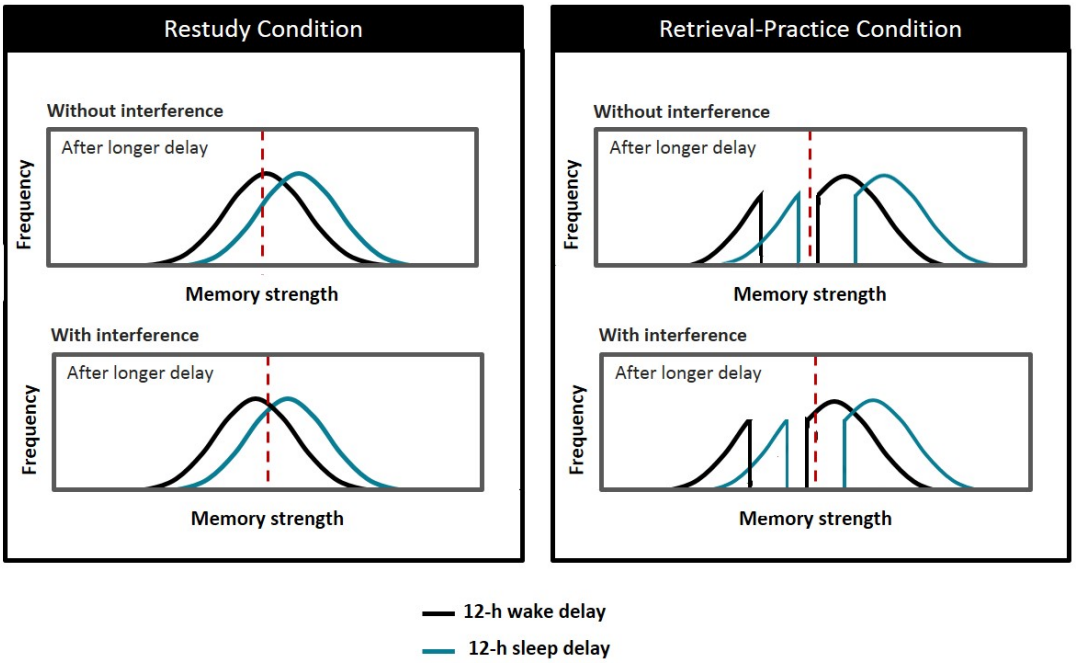

Figure 5: Memory strength distributions of two hypothetical sets of items after wake delay (black curves) and sleep delay (blue curves), based on the bifurcation model (Kornell et al., 2011). The left column shows the restudy condition while the right column shows the retrieval-practice condition. The top pair of panels indicates memory strength in a final recall test in the absence of retroactive interference. In contrast, the lower pair of panels shows memory strength in the presence of retroactive interference, curves being shifted to the left as memory strength supposedly is decreased. A comparison of amount of items above recall threshold (red dotted lines) indicates that retroactive interference increases testing effects at final test.

presented for study. Items of this second interfering list were chosen from the same semantic categories as items from the original study list. Study of the additional item list should induce retroactive interference as items with the same category cue should compete for retrieval during the final test. The induction of interference is a way of enhancing recall difficulty and, thus, should prevent possible ceiling effects. Like in Experiment 1, we expected to find a reliable testing effect after the wake delay. According to Halamish and Bjork (2011), this effect might even be increased relative to Experiment 1, as it is supposed to rise with difficulty of the final test. More interesting though, if results in the retrieval condition of Experiment 1 were mainly due to a ceiling effect, then the testing effect in this experiment might arise both after wake and sleep delay. Size of the effect should then be similar between both delays or even greater after sleep delay. In contrast, if the results of Experiment 1 were not due to a ceiling effect, then the testing effect may again be reduced, or even be eliminated, after sleep.

Method

Participants

144 healthy students from Regensburg University participated in the experiment (M = 22.9 years; 18-30 years; 44 male). Criteria for subject selection were the same as in Experiment 1. Subjects in the 12-h sleep condition were instructed to try to sleep regularly during the night between sessions, while subjects in the 12-h wake condition were instructed not to take naps during the day. Every subject was asked to refrain from drinking alkoholic beverages between sessions. Again all participants were native German speakers and were distributed equally across conditions ($n = 36$ in each of the four condiditions). Comparisons between conditions regarding subjects' age, habitual sleep duration, subjective ratings of sleep quality, and ratings on the Epworth Sleepiness Scale (Johns, 1991) did not reveal any differences (all $ps > .10$). IQ of participants was not estimated in this experiment, as the connect-the-numbers-task was replaced by the presentation of the second item list (nontarget list).

Material

Two separate item lists served as learning material in the experiment. The target list consisted of the same 24 nouns that were already used in Experiment 1. Additionally, to induce retroactive interference, a nontarget list with another 24 concrete German nouns was constructed. For this purpose, items were selected from the same four semantic categories that were already employed in the target list (six items per category Scheithe & Bäuml, 1995; Van Overschelde et al., 2004). Whithin each category, all target and nontarget items had unique initial letters.

Design

The experiment had a 2 × 2 × 2 mixed-factorial design with the between-subjects factors of type of practice (restudy, retrieval practice) and delay (12-h wake, 12-h sleep), and the within-subjects factor of practice level (low, high). Apart from minor exceptions, Experiment 2 was designed in an identical way as Experiment 1. To induce retroactive interference, participants studied the nontarget list at the beginning of the second session. Another change to the design of Experiment 1 was the omission of a short-delay condition (see Figure 6 for an illustration of the experimental design of Experiment 2).

Procedure

The first session of the experiment (Study and practice phase) was conducted in exactly the same way as in Experiment 1. After participants were asked to study all items of the target list in one initial study cycle, they either studied the items on one or two restudy cycles (restudy condition) or on one or two retrieval-practice cycles. Two of the categories were used for the low practice level, while the remaining two categories were used for the high practice level. Again, semantic categories were distributed over practice levels in a balanced manner. In the retrieval condition, subjects were again provided with the items' category labels and word stems for up to 5 sec each, and were asked to recall the corresponding items. This time mean response time (equalling overall processing time) across

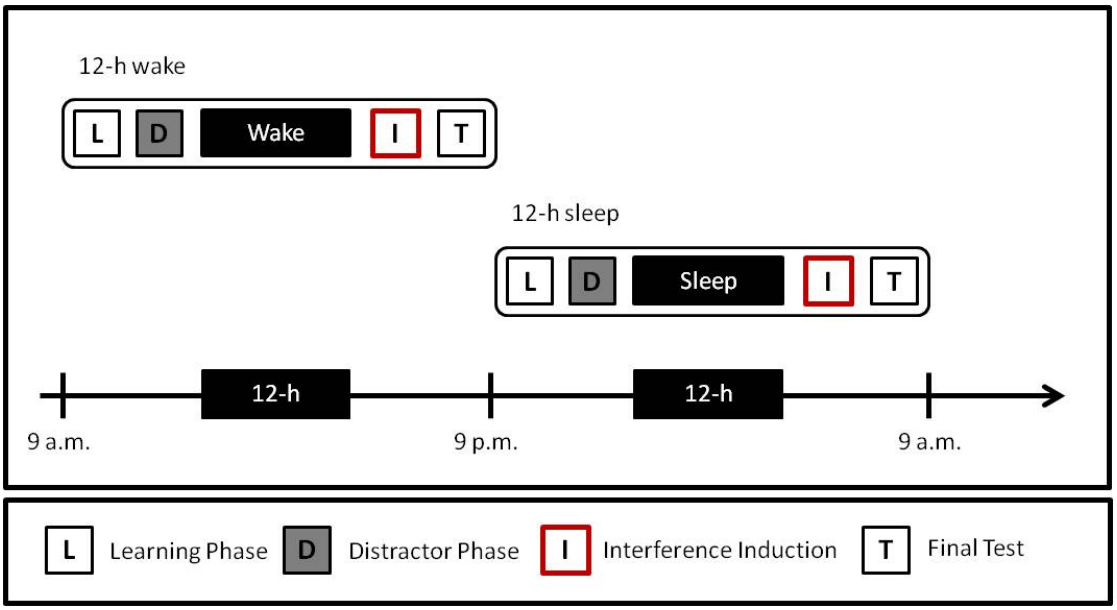

Figure 6: Illustration of conditions employed in *Experiment 2*: In the 12-h wake conditon, the learning of the material took place at 9 a.m., before participants returned to the laboratory for the final test after a 12-h delay of diurnal wakefulness. In the 12-h sleep condition, the learning of the material took place at 9 p.m., and memory was tested after a 12-h delay comprising nocturnal sleep. In contrast to *Experiment 1* a second item list (nontarget list) was studied after the 12-h delay to induce retroactive interference.

all retrieval-practice trials was $M = 2.2$ sec ($SD = .44$). After the 12-h delay, instead of employing the full second half of the distractor phase, the nontarget list was presented. The list was presented three times at a rate of 3 sec per item; items were presented in a random order. This phase was followed by a 30-sec backward-counting distractor task. Afterwards, the final test was conducted in an analogous manner to Experiment 1. The items of the target list were tested first and the items of the nontarget list second. Concerning compliance with instructions, subjects in the 12-h sleep condition reported to have slept regularly during the experimental night ($M = 7.8$ hrs; $SD = .93$). Time spent asleep during the experimental night did not differ from habitual sleep time of participants in this condition ($p > .10$). Subjects in the 12-h wake condition reported not to have taken any naps during the day. None of the subjects reported alcohol intake between sessions.

Results

Success Rates during Retrieval-Practice Cycles

A 2 x 2 ANOVA with the factors of time of day (9 a.m., 9 p.m.) and retrieval practice (ST, STT) showed that retrieval success was higher after two than after one retrieval-practice cycle (96.5% vs. 90.0%), $F(1, 70) = 24.21$, $MSE = 62.83$, $p < .001$, $\eta^2 = .26$. There was no main effect of time of day and no interaction, $ps > .10$.

Final Test (12-h Delay Conditions)

Figure 7 shows recall performance after the 12-h delay. A 2 x 2 x 2 ANOVA with the factors of type of practice (restudy, retrieval practice), delay (12-h wake, 12-h sleep), and practice level (low, high) revealed significant main effects of type of practice, $F(1, 140) = 53.00$, $MSE = 410.44$, $p < .001$, $\eta^2 = .28$, delay, $F(1, 140) = 9.26$, $MSE = 410.44$, $p = .003$, $\eta^2 = .06$, and practice level, $F(1, 140) = 33.60$, $MSE = 171.53$, $p < .001$, $\eta^2 = .19$. The main effect of type of practice again reflects overall higher recall in the retrieval condition than in the restudy condition (68.4% vs. 51.0%), whereas the main effect of delay shows that recall rates were overall higher in the 12-h sleep condition than in the 12-h wake condition (63.4% vs. 56.1%); the main effect of practice level indicates that items were recalled better with the higher than the lower practice level (64.2% vs. 55.3%). More important, while all other interactions were nonsignificant (all $ps > .05$), there was once more a reliable two-way interaction between type of practice and delay, $F(1, 140) = 4.39$, $MSE = 410.44$, $p = .038$, $\eta^2 = .03$, indicating that the beneficial effect of retrieval practice in comparison to restudy was again modulated by delay condition. Consistently, planned comparisons revealed significant testing effects after the 12-h wake delay for both the lower practice level (40.3% vs. 62.4%), $t(70) = 5.57$, $p < .001$, $d = 1.31$, and the higher practice level (49.5% vs. 72.1%), $t(70) = 6.01$, $p < .001$, $d = 1.42$, but numerically reduced testing effects after the 12-h sleep delay, for both the lower practice level (51.6% vs. 66.7%), $t(70) = 3.49$, $p = .001$, $d = .82$,

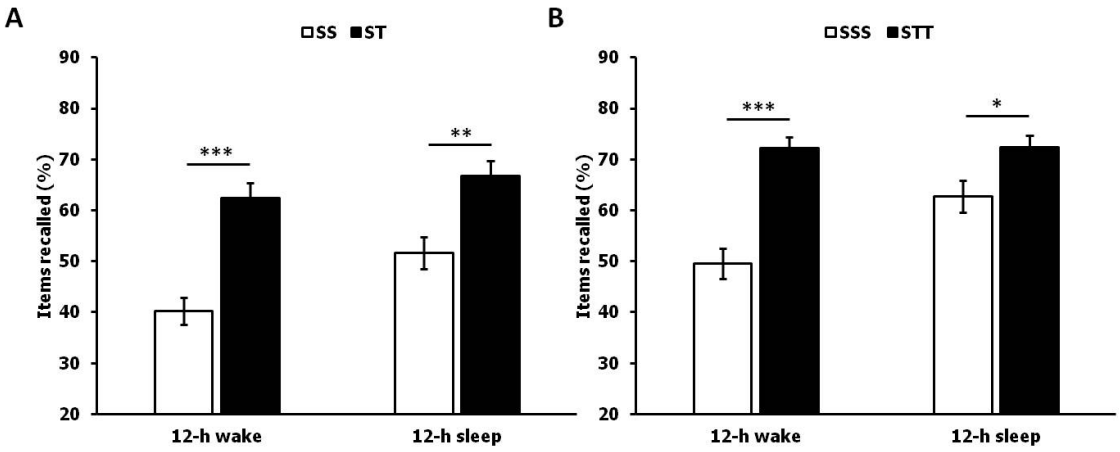

Figure 7: Mean recall performance on the final test of *Experiment 2* as a function of delay (12-h wake, 12-h sleep) and type of practice (restudy, retrieval practice), separately for the low (A) and the high (B) practice level. Condition labels indicate study (S) and retrieval practice (T) cycles. Error bars represent standard errors; $*p < .05$; $**p < .01$; $***p < .001$.

and the higher practice level (62.7% vs. 72.4%), $t(70) = 2.43$, $p = .018$, $d = .57$.

Consistently, there was an effect of sleep-associated memory consolidation for items that had been restudied once (SS), $t(70) = 2.73$, $p = .008$, $d = .64$, and for items that had been restudied twice (SSS), $t(70) = 3.03$, $p = .003$, $d = .71$. In contrast, there was no effect of sleep-associated memory consolidation for items that had been practiced once (ST), $t(70) = 1.02$, $p = .311$, $d = .24$, and no effect for items that had been practiced twice (STT), $t(70) = .08$, $p = .934$, $d = .02$.

Recall for the nontarget list was analyzed in a 2 x 2 ANOVA with the factors of type of practice (restudy, retrieval practice) and delay (12-h wake, 12-h sleep). The analysis revealed no significant effects, all $Fs < 1.0$, indicating that there were no reliable differences between practice conditions (58.6% vs. 57.9%) and no reliable differences between delay conditions (58.2% vs. 58.2%).

Additional Analyses

To analyze the effect of retroactive interference on memory retention as a function of experimental condition, data of Experiments 1 and 2 was directly compared.

This was done by means of a 2 x 2 x 2 x 2 ANOVA with the factors of type of practice (restudy, retrieval practice), delay (12-h wake, 12-h sleep), practice level (low, high), and interference (interference, no interference). This analysis again revealed significant main effects of type of practice, $F(1, 280) = 61.47$, $MSE = 342.55$, $p < .001$, $\eta^2 = .18$, delay, $F(1, 280) = 26.33$, $MSE = 342.55$, $p < .001$, $\eta^2 = .09$, and practice level, $F(1, 280) = 67.95$, $MSE = 150.19$, $p < .001$, $\eta^2 = .20$, as well as a significant interaction between type of practice and delay, $F(1, 280) = 11.46$, $MSE = 342.55$, $p = .001$, $\eta^2 = .04$. In addition, it revealed a significant main effect of interference, $F(1, 280) = 56.11$, $MSE = 352.55$, $p < .001$, $\eta^2 = .17$, indicating lower recall rates in Experiment 2 than in Experiment 1 (59.7% vs. 71.3%). Most important, there was a significant interaction between type of practice and interference, $F(1, 280) = 11.76$, $MSE = 342.55$, $p = .001$, $\eta^2 = .04$, indicating that the testing effect was influenced by interference. Indeed, planned comparisons revealed that the testing effect for the lower practice level was greater in the presence of interference (45.9% vs. 64.6%), $t(142) = 6.20$, $p < .001$, $d = 1.03$, than in its absence (64.3% vs. 70.3%), $t(142) = 2.24$, $p = .027$, $d = .37$, and the testing effect for the higher practice level was also greater in the presence of interference (56.1% vs. 72.3%), $t(142) = 5.70$, $p < .001$, $d = .95$, than in its absence (71.4% vs. 79.0%), $t(142) = 3.28$, $p = .001$, $d = .55$. No further effects were significant ($ps > .10$).

Discussion

The central methodological difference between Experiment 1 and Experiment 2 was the induction of retroactive interference through the introduction of a second study list (nontarget list). This list was semantically related to the target list and, in comparison to the results of Experiment 1 was supposed to impede recall during the final test. As results of Experiment 1 were close to ceiling (particularly in the high practice condition), higher difficulty of the final test was supposed to reduce the overall recall level, enabling a better interpretation of the data. Indeed, the present results reveal that overall recall rates were significantly reduced in Experiment 2 compared to Experiment 1 (59.7% vs. 71.3%) indicating a higher

final test difficulty after the induction of interference. Even under such conditions of increased test difficulty, results of Experiment 1 were largely replicated. Testing effects after sleep delay were reliably reduced compared to testing effects after the wake delay, regardless of practice level. This time though, testing effects arose both after wake and after sleep delay. As in Experiment 1 sleep effects were only present after restudy but absent after retrieval practice facilitating the reduction of the testing effect after sleep delay. Thus, the present results can be seen as a replication and extension of Experiment 1, excluding a ceiling effect as the source of the observations.

By inducing retroactive interference, Experiment 2 employed a more difficult final recall test than Experiment 1. According to the bifurcation model (Halamish & Bjork, 2011), this should raise the recall threshold, resulting in overall increased testing effects (see Figure 5). Comparisons between results of Experiment 1 and Experiment 2 show that testing effects are higher in the presence of retroactive interference than in its absence. This replicates findings of prior studies (Abel & Bäuml, 2014; Halamish & Bjork, 2011; Potts & Shanks, 2012). In contrast to results on the testing effect, comparisons between Experiment 1 and Experiment 2 indicate no significant influence of sleep on effects of retroactive interference. This contradicts earlier studies indicating that sleep can protect memories from the detrimental influence of retroactive interference (Ellenbogen, Hulbert, et al., 2006; Ellenbogen et al., 2009). As mentioned before, evidence regarding this issue is mixed though, as Deliens et al. (2013) found effects of retroactive interference to be even more pronounced after sleep versus wake delay. All previous studies that have investigated this relationship used different item material (paired associates) than was used in the present experiment. This might explain the results indicating no influence of sleep on effects of retroactive interference. Experiment 4 will take up the issue again being designed in a similar way as Experiment 2 but using paired associates instead of categorized items as learning material.

2.3 Summary

The results of Experiments 1 and 2 replicate studies on the testing effect (e.g. Carrier & Pashler, 1992; Hogan & Kintsch, 1971; Karpicke & Roediger, 2008). Consistent, over both experiments there was a significant testing effect after the 12-h wake delay. An effect that was observable regardless of level of practice and the presence or absence of retroactice interference at test. On average, items that had been subject to retrieval practice were recalled considerably better than restudied items (70.2% vs. 52.9%). Interestingly this effect was not present yet after the short delay of Experiment 1 but only arose after the 12-h delay, a phenomenon that has been observed in prior studies (Roediger & Karpicke, 2006b; Toppino & Cohen, 2009). These findings emphasize the considerable importance of the length of the retention interval for the occurence of a testing effect (test-delay interaction).

In addition to replicating previous work, aim with the present experiments was to examine a possible influence of sleep on the testing effect. Numerous studies have investigated sleep's influence on memory representation and evidence on sleep-associated memory consolidation is compelling (see Diekelmann & Born, 2010; Diekelmann et al., 2009). Additionally, recent studies have pointed out that beneficial effects of sleep can be selective for certain types of memories (see e.g. Payne et al., 2008; Wilhelm et al., 2011). Thus, memorial consequences of wake vs. sleep delay are very differential and depend on the type of memory. The results of Experiments 1 and 2 consistently indicate that type of delay (wake vs. sleep) is a moderating factor of the testing effect. In contrast to results after wake delay, testing effects were reduced or even eliminated when retention interval was filled with nocturnal sleep. In both experiments this extenuated testing effect was presumably caused by differences in sleep-induced benefits after restudy and retrieval practice. While, on average over both experiments, items benefitted reliably from sleep delay after restudy (sleep effect: 13,1%), there was almost no such benefit after retrieval practice (sleep effect: 2,7%). Thus, memorial benefits of sleep versus wake were related to the type of practice during the learning phase.

The results of Experiment 1 and 2 fit nicely into the distribution-based

bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011). As is illustrated in Figure 2, the model assumes that items that were successfully retrieved during the retrieval-practice phase (as opposed to the nonretrieved items) are strenghtened to a higher degree than restudied items. As a consequence, successfully retrieved items might still be above recall threshold after wake delay and, hence, reliable benefits of sleep delay should not be observable in the final test. In contrast, as the memory strength distribution of restudied items is not bifurcated, the model would predict more of these items to be above recall threshold after sleep than after wake delay, resulting in sleep-associated benefits at the final test. This would mean that, consistent with the present results, the testing effect may be present after wake delay but reduced or even absent after sleep delay. Another assumption of the bifurcation model is that the testing effect depends on final test difficulty (Halamish & Bjork, 2011). Recall threshold should rise with higher test difficulty, resulting in greater testing effects as particularly restudied items fall below recall threshold (see Figure 5). Comparisons between Experiments 1 and 2 confirm this prediction and replicate other studies that investigated the issue (Abel & Bäuml, 2014; Potts & Shanks, 2012).

Another account that seems to be in line with the present results, is the encoding context account (Karpicke et al., 2014). This account, in contrast to the bifurcation model, makes basic assumptions about the nature of processes active during retrieval practice. Retrieval practice is supposed to be more powerful in enhancing long-term memory than restudy as it leads to a reactivation of the learning context, complementing it with new context information. Thus, it reduces the size of the search set at final recall. In contrast to the bifurcation model, the account does not make any direct assumptions about a possible relationship between sleep and the testing effect. However, evidence on how sleep interacts with contextual memory effects, shows that sleep but not wake delays can reduce the reliance on contextual cues during final recall (Cairney et al., 2011). Additionally, mnemonic information is supposed to be reactivated during sleep, enhancing the integration of memories into existing memory networks (Rasch et al., 2007). Thus, benefits of retrieval practice might be less evident after sleep than after wake delay. As the episodic context account is not contradictory to assumptions made by

the bifurcation model, both might co-exist and offer explanations for the present results.

The present results appear to be less well consistent with another process model of the testing effect. According to the elaborative retrieval hypothesis (Carpenter, 2009; Pyc & Rawson, 2010), the attempt of retrieving memorial information might enhance elaborative processing. Thus, there might be a co-activation of semantically related information that becomes linked to the target information. This is supposed to foster the testing effect as mainly difficult (retrieval-practice) tasks and less easier tasks or restudy cycles should result in such increased elaborative processing. With this theoretical framework in mind, one would expect sleep to maintain or even increase the testing effect, as sleep has repeatedly been found to raise memory for semantic information (Cai et al., 2009; McKeon et al., 2012; Payne et al., 2009).The present results cannot confirm these predictions. In contrast, the finding of a reduced or eliminated testing effect after sleep can hardly be explained with the elaborative retrieval hypothesis. The present experiments though might not be able to test the elaborative retrieval hypothesis properly as it might be argued that initial success rates during retrieval practice were to high. So it might be that retrieval practice did not induce sufficient elaboration. However, testing effects being clearly evident after wake delay elaboration induced by retrieval practice should have at least been adequate to produce such effects. This will be further investigated in Experiments 3 and 4 of the present work.

On the basis of findings on the relationship between retrieval-induced forgetting and sleep (Abel & Bäuml, 2012) aim with Experiments 1 and 2 was to investigate the influence of sleep on the testing effect. The results replicate the finding of Abel and Bäuml (2012), showing that benefits of sleep are absent after retrieval practice. Using the retrieval-practice paradigm (Anderson et al., 1994) it could show that there was a beneficial effect of sleep delay for control items that had only been presented intactly to subjects without asking them to actively engage in retrieval practice of the items. In contrast, and in accordance with the present results, items subjected to retrieval practice during the learning phase did not show such benefits in recall performance after sleep delay. Such results seem to be in conflict with prior evidence on effects of sleep-associated memory consolidation after repeated

study-test cycles (e.g. Drosopoulos et al., 2007; Ellenbogen, Hulbert, et al., 2006). Indeed, several studies found sleep effects using such a learning method that involves a mixture of retrieval-practice and restudy (feedback) trials. Thus, these studies did not discriminate between sleep effects for restudy versus retrieval practice. Basically, they only offer information about how items with varying strength level might be differentially affected by sleep, being silent about if and how sleep interacts with different types of practice. Thus, the critical methodological differences between the present experiment and these prior studies, which might account for the disparity in results, might be the use of pure retrieval-practice trials without intermixed restudy trials (see also Abel & Bäuml, 2012) and possibly even the application of categorized item material instead of paired associates.

In sum, the first two experiments replicate several findings on the testing effect and even go beyond them in some aspects. As the first two experiments were built on the basis of categorized item material which is commonly used in experiments on retrieval-induced forgetting (e.g. Anderson et al., 1994) the question remains if the present results are replicable with other item material. Studies on the influence of sleep on declarative memory contents have commonly used paired associates as item material (e.g. Fogel, Smith, & Cote, 2007; Gais et al., 2006; Plihal & Born, 1997; Tucker et al., 2006). Likewise, many studies on the testing effect have previously used word pairs as learning material (e.g. Allen et al., 1969; Carpenter, 2009; Carrier & Pashler, 1992; Karpicke & Roediger, 2008). Experiments 3 and 4 were designed to investigate the relationship between sleep and the testing effect with semantically unrelated word pairs which resembles typical item material used in studies on the sleep effect as well as the testing effect.

Chapter 3

Sleep and the Testing Effect - Paired Associates

3.1 Experiment 3

The results of Experiments 1 and 2 indicate that wake and sleep delay can affect the testing effect differentially. Testing effects were consistently reduced or even eliminated after 12-h delays filled with nocturnal sleep. This was found regardless of level of practice and the presence or absence of retroactive interference. In both experiments categorized item lists were used as learning material to replicate and further explore findings by Abel and Bäuml (2012) on the relationship between retrieval-induced forgetting and sleep. In contrast, Experiment 3 used semantically unrelated word pairs as learning material to replicate the findings of the first two experiments under conditions that resemble previous studies on both the sleep effect and the testing effect (see Diekelmann et al., 2009; Roediger & Butler, 2011).

Paired associates have been common item material in studies on the testing effect ever since a seminal study by Allen et al. (1969). Subjects in this study were asked to learn a list of paired associates either 5 or 10 times and then take no, one, or five tests on the items. During final test they were provided with the left-hand member of each word pair and asked to recall the second item. While differences in recall performance only varied modestly with number of restudy trials, it was significantly improved with increasing numbers of retrieval-practice trials. Similarly, in a study by Jacoby (1978) subjects were during retrieval practice asked to come up with the second item of initially studied word pairs when given the left-hand member and a fragmented form of the second item. Final recall performance after retrieval practice was once more better than after restudy, i.e. there was a significant testing effect (for related results see, e.g. Carrier & Pashler, 1992; McDaniel & Masson, 1985). Thus both categorized item material (see above) and paired associates have been used in studies on the testing effect showing that benefits of retrieval practice generalize to many different item materials (for an overview see Roediger & Karpicke, 2006a).

According to the elaborative retrieval hypothesis (Carpenter, 2009; Pyc & Rawson, 2010), the memorial advantage of retrieval practice over restudy is fostered by elaborative processing. In fact, several studies show that testing effects

increase with difficulty of the retrieval-practice task (Carpenter & Delosh, 2006; Pyc & Rawson, 2009). As mentioned above, the results of Experiments 1 and 2 were not in accordance with the elaborative retrieval hypothesis as it would predict testing effects to be preserved or even increased after sleep. Yet, as the retrieval-practice task was quite simple in these experiments, presumably resulting in low retrieval effort, one cannot draw firm conclusions. Experiment 3 further explores this issue by increasing difficulty of the retrieval-practice task through the use of more challenging learning material (paired associates). With difficulty of retrieval, according to the elaborative retrieval hypothesis, the testing effect should increase in size (e.g. Carpenter, 2009). Similarly, according to the episodic context account (Karpicke et al., 2014), increased difficulty of the retrieval-practice task should enhance testing effects. This account suggests that more difficult retrieval-practice requires more context reinstatement increasing benefits of retrieval practice over restudy. It is unclear thus far whether the results of the first two experiments can be replicated under such changed conditions i.e. if they generalize to different and more challenging learning material. It might e.g. be that testing effects are increased resulting in reliable differences between restudy and retrieval practice not only after wake but even after sleep delay.

Experiment 3 was very similar to Experiment 1 but used different learning material. Participants were asked to learn a list of unrelated paired associates instead of categorized items. In accordance with Experiment 1, there were restudy and retrieval practice conditions. Subjects in the restudy condition were asked to restudy one half of the initially studied paired associates once, and the other half twice; analogical, in the retrieval-practice condition, subjects were asked to practice retrieval of half of the paired associates once, and that of the other half twice. Again, the final test of the item material was conducted after a 12-h delay that included either regular sleep or wakefulness. Subjects were presented the stimulus words of the paired associates and were asked to recall the appropriate response words. The results of the experiment will help clarifying whether the reduction of the testing effect after sleep, as it was found in Experiments 1 and 2, is restricted to categorized item material, or generalizes to paired associates.

Method

Participants

A sample of 216 healthy students from Regensburg University took part in the experiment voluntarily and in return for financial reimbursement ($M = 22.0$ years; range 18-30 years; 31 male). As in Experiments 1 and 2, all participants were asked to complete a screening questionnaire at the beginning of the experiment. Criteria for subject selection were the same as in the first two experiments. Subjects in the 12-h sleep condition were instructed to try to sleep regularly during the night between sessions, while subjects in the 12-h wake condition were instructed not to take naps during the day. Every subject was asked to refrain from drinking alkoholic beverages between sessions. Participants were native German speakers and were distributed equally across conditions ($n = 36$ in each of the six conditions). Once more, comparisons between conditions regarding subjects' age, habitual sleep duration, subjective ratings of sleep quality, IQ (estimated via speed of cognitive processing; Oswald & Roth, 1987), and ratings on the Epworth Sleepiness Scale (Johns, 1991) did not reveal any differences (all $ps > .05$).

Material

32 unrelated neutral, one- and two-syllable, words were drawn from different semantic categories (Van Overschelde et al., 2004). 16 of these items were randomly chosen as stimulus words. The remaining 16 items were used for a second list of response items. A list of paired associates was created by randomly pairing the stimulus list with the response list. After initial study of the whole list, 8 of the paired associates were repeated once during retrieval practice or restudy, whereas the remaining paired associates were repeated twice; sets of paired associates were distributed across practice levels in a balanced manner.

Design.

The experiment had the same 2 × 2 × 2 mixed-factorial design as Experiment 1. The factors of TYPE OF PRACTICE (restudy, retrieval practice) and DELAY (12-h wake, 12-h sleep) were again manipulated between subjects, the factor of PRACTICE LEVEL (low, high) was again manipulated within subjects. Analogically to Experiment 1 after one initial study cycle, half of the subjects were asked to restudy the paired associates (restudy condition), while the other half was asked to engage in retrieval practice (retrieval-practice condition). There were again two practice levels with differing numbers of practice cycles. For one half of the initially studied paired associates, subjects were given one restudy or retrieval-practice cycle (low practice level); for the other half, they were given two restudy or retrieval- practice cycles (high practice level; see Figure 8). Additionally, like in Experiment 1, there were two delay conditions. Participants in the 12-h wake condition studied and practiced the word pairs at 9 a.m., and final test was conducted at 9 p.m., after 12 hours of wakefulness; in contrast, in the 12-h sleep condition, participants studied and practiced the word pairs at 9 p.m., and took the final test at 9 a.m., after one night of nocturnal sleep. To control for potential circadian effects, a short-delay condition was included. Half of the subjects in this condition participated at 9 a.m., the other half at 9 p.m.

Procedure.

Study and Practice Phase. Experiment 3 was conducted in a similar fashion as Experiment 1. During study, paired associates were presented successively in a random order, at a presentation rate of 5 sec each. After one initial study cycle for all 16 word pairs, there were one or two additional cycles comprising either restudy or retrieval-practice trials, depending on practice condition. In restudy conditions, subjects were again presented with the paired associates for 5 sec each. Half of the paired associates were restudied once (SS), whereas the other half of paired associates was restudied twice (SSS). In retrieval conditions, subjects were provided with the stimulus words and word stems of the response items for up to 7 sec each, and were asked to recall the corresponding items (mean time

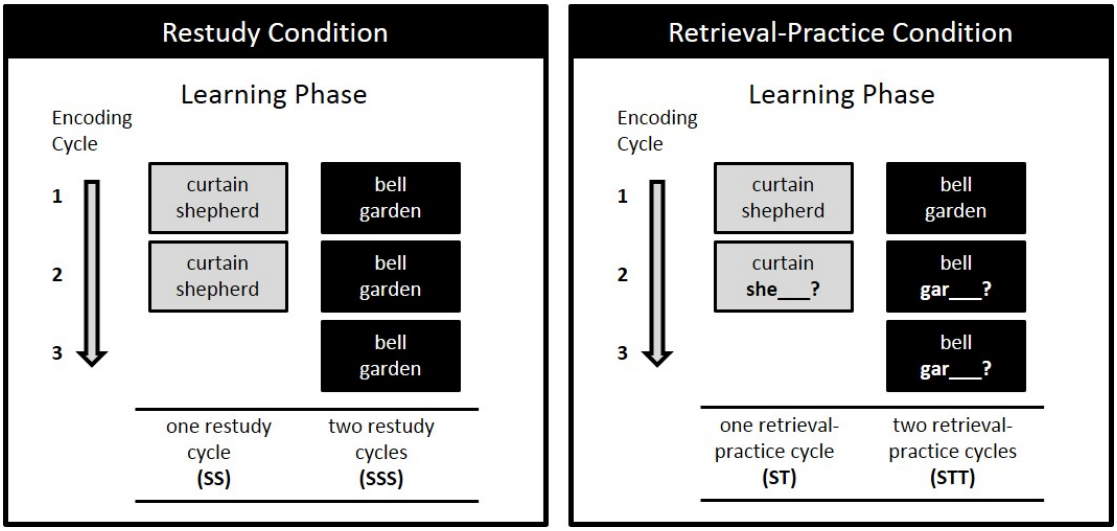

Figure 8: Illustration of the two types of practice employed in *Experiment 3*: In the restudy condition, after the initial study cycle, paired associates were learned in one or two restudy cycles. In contrast, in the retrieval-practice condition initial study of paired associates was followed by one or two retrieval-practice cycles.

until response: $M = 2.3$ sec; $SD = .51$). Retrieval practice was conducted in a covert fashion to allow for the simultaneous testing of two subjects. Instead of answering verbally during retrieval practice, participants were asked to press a corresponding key, thereby indicating whether they successfully retrieved the response item; covert retrieval has been shown to lead to similar testing effects as overt retrieval does (Putnam & Roediger, 2013; M. A. Smith & Roediger, 2013). Retrieval of the response item of half of the paired associates was practiced once (ST), whereas retrieval of the other half was practiced twice (STT). In both restudy and retrieval conditions, order of the paired associates was random.

Participants spent the interval between encoding phase and test phase in parallel to Experiment 1. Consistent with Experiments 1 and 2, subjects in the 12-h sleep conditions reported to have slept regularly during the night ($M = 7.4$ hrs; $SD = 1.07$). Time spent asleep during the experimental night did not differ from habitual sleep time of participants in this condition ($p > .05$). Subjects in the 12-h wake condition reported not to have taken any naps during the day. None of the subjects reported alcohol intake between sessions.

Concerning compliance with instructions, subjects in the 12-h sleep condition reported to have slept regularly during the experimental night (M = 7.9 hrs; SD = .83). Time spent asleep during the experimental night did not differ from habitual sleep time of participants in this condition ($p > .10$). Subjects in the 12-h wake condition reported not to have taken any naps during the day. None of the subjects reported alcohol intake between sessions.

Test Phase. At test, participants were presented with the stimulus words and initial letters of the response items of all 16 studied paired associates for 7 sec each, and were asked to recall the appropriate response item. Order of paired associates was random.

Results

Success Rates during Retrieval-Practice Cycles

A 2 x 2 x 2 ANOVA with the factors of TIME OF DAY (9 a.m., 9 p.m.), RETRIEVAL PRACTICE (ST, STT), and DELAY (short delay, 12-h delay) revealed that retrieval success was higher after two than after one retrieval-practice cycle (86.5% vs. 80.8%), $F(1, 104) = 11.13$, $MSE = 137.28$, $p = .001$, $\eta^2 = .10$, as indicated by subjects' key presses. There were no other main effects and no interactions, $ps > .10$.

Final Test (12-h Delay Conditions)

Figure 9 shows recall performance after the 12-h delay. A 2 x 2 x 2 ANOVA with the factors of TYPE OF PRACTICE (restudy, retrieval practice), DELAY (12-h wake, 12-h sleep), and PRACTICE LEVEL (low, high) revealed significant main effects of DELAY, $F(1, 140) = 9.26$, $MSE = 673.69$, $p = .003$, $\eta^2 = .06$, and PRACTICE LEVEL, $F(1, 140) = 38.89$, $MSE = 235.77$, $p < .001$, $\eta^2 = .22$, but no main effect of TYPE OF PRACTICE, $F(1, 140) = .89$, $MSE = 673.69$, $p = .348$, $\eta^2 = .01$. The main effect of DELAY reflects that recall rates were overall higher in the 12-h sleep condition than in the 12-h wake condition (67.3% vs. 58.0%), whereas the

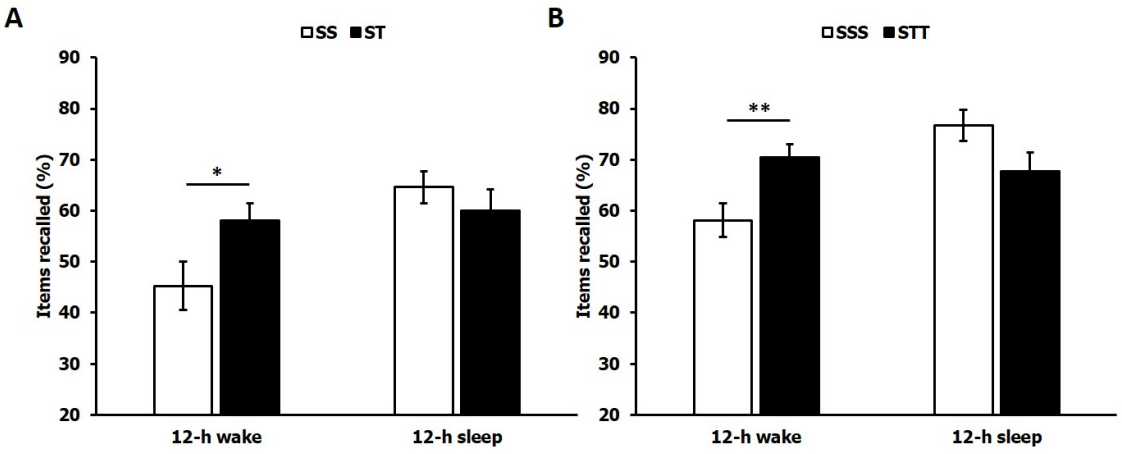

Figure 9: Mean recall performance on the final test of *Experiment 3* as a function of delay (12-h wake, 12-h sleep) and type of practice (restudy, retrieval practice), separately for the low (A) and the high (B) practice level. Condition labels indicate study (S) and retrieval practice (T) cycles. Error bars represent standard errors; *$p < .05$; **$p < .01$.

main effect of PRACTICE LEVEL indicates better recall after the high than the low practice level (68.3% vs. 57.0%). More important, all other interactions being nonsignificant (all $ps > .10$), there was a reliable two-way interaction between TYPE OF PRACTICE and DELAY, $F(1, 140) = 9.96$, $MSE = 673.69$, $p = .002$, $\eta^2 = .07$, indicating that the difference in memory performance after retrieval practice in comparison to restudy was modulated by delay condition. Consistently, planned comparisons showed significant testing effects after the 12-h wake delay, for both the lower practice level (45.3% vs. 58.0%), $t(70) = 2.25$, $p = .028$, $d = .53$, and the higher practice level (58.1% vs. 70.5%), $t(70) = 2.94$, $p = .005$, $d = .69$, whereas no testing effects arose after sleep - but they were even numerically reversed -, both for the lower practice level (64.6% vs. 60.1%), $t(70) = .85$, $p = .399$, $d = .20$, and the higher practice level (76.7% vs. 67.7%), $t(70) = 1.89$, $p = .064$, $d = .44$.

Planned comparisons revealed that there was a sleep effect for items that had been restudied once (SS), $t(70) = 3.33$, $p = .001$, $d = .79$, or twice (SSS), $t(70) = 4.09$, $p < .001$, $d = .96$. However, there was no sleep effect for items that had been subject to retrieval practice once (ST), $t(70) = .40$, $p < .688$, $d = .09$, or twice (STT), $t(70) = .62$, $p < .536$, $d = .15$.

Table 2: Mean recall performance in the short-delay condition of *Experiment 3* as a function of time of day (9 a.m., 9 p.m.), type of practice (restudy, retrieval practice), and practice level (low, high). Condition labels indicate study (S) and retrieval-practice (T) cycles. Standard errors are displayed in parentheses.

	Restudy		Retrieval Practice	
	SS	**SSS**	**ST**	**STT**
9 a.m.	81.9 (5.5)	88.9 (3.8)	71.5 (6.8)	82.6 (3.5)
9 p.m.	82.6 (5.2)	86.8 (4.9)	72.9 (5.6)	82.6 (4.9)
Combined	82.3 (3.7)	87.8 (3.0)	72.2 (4.3)	82.6 (3.0)

Circadian Control (Short-Delay Condition)

Table 2 shows mean recall levels after the short delay. A 2 x 2 x 2 ANOVA with the factors of type of practice (restudy, retrieval practice), time of day (9 a.m., 9 p.m.), and practice level (low, high) revealed a significant main effect of practice level, $F(1, 68) = 20.49$, $MSE = 111.88$, $p < .001$, $\eta^2 = .23$, with items of the higher practice level being recalled better than items of the lower practice level (85.2% vs. 77.3%). No other effects emerged, indicating that there was no testing effect after short retention interval and recall was unaffected by circadian effects, all $ps > .10$.

Additional Analyses

Experiments 1 and 3 were similar in most aspects but differ in the learning material employed. In Experiment 3 paired associates were used to induce a more effortful retrieval-practice task than with categorized item material that was used in Experiment 1. Indeed, a comparison between mean success rates during retrieval-practice suggests that the task was more difficult in Experiment 3 (Experiment 1: 92.6%; Experiment 3: 83.7%), $t(214) = 7.32$, $p < .001, d = 1.00$.

This difference did even result in a numerically larger test-delay interaction in Experiment 3. Indeed, when categorized item material was used (Experiment 1), mean recall reduction from short delay to 12-h wake delay - averaged over both practice levels - was 16.9% after restudy and 7.1% after retrieval practice. Whereas, when paired associates were used as item material (Experiment 3), forgetting over the delay was 33.7% after restudy but only 12.7% after retrieval practice. The results of a 2 x 2 x 2 ANOVA with the factors of TYPE OF PRACTICE (restudy, retrieval practice), DELAY (9 a.m. short delay, 12-h wake), and EXPERIMENT (Experiment 1, Experiment 3) confirmed the numerical indication of a test-delay interaction, $F(1, 208) = 11.56$, $MSE = 241.10$, $p = .001$, $\eta^2 = .05$, showing more forgetting after restudy than after retrieval practice cycles; the numerically larger interaction in Experiment 3 than Experiment 1 did not reach significance, $F(1, 208) = 1.57$, $MSE = 241.10$, $p = .212$, $\eta^2 = .01$. Finally, EXPERIMENT affected delay-induced forgetting, with a larger amount of forgetting in Experiment 3 than Experiment 1, $F(1, 208) = 6.36$, $MSE = 241.10$, $p = .012$, $\eta^2 = .03$.

Discussion

Experiment 3 was implemented to investigate the relationship between sleep-associated memory consolidation and the testing effect. Aim was a replication of Experiment 1 using paired associates instead of categorized items, extending these findings to other learning material. Paired associates have most commonly been used in earlier studies on the sleep effect and the testing effect (see Diekelmann et al., 2009; Roediger & Butler, 2011). As the difficulty of the retrieval-practice task in Experiment 1 was quite low, with high success rates during retrieval practice (mean: 92.6%), goal of Experiment 3 was even to raise retrieval difficulty. As intended, the learning of paired associates proved to be more challenging resulting in lower success rates during retrieval practice in this experiment (mean: 83.7%), indicating that more effort was needed to retrieve the items.

Like in Experiment 1, the present results replicate prior testing effect studies

(e.g. Roediger & Karpicke, 2006b; Wheeler & Roediger, 1992). After 12-h wake delay there were reliable benefits of retrieval practice over restudy, regardless of number of practice cycles. In contrast, and in line with results of Experiment 1, there was no such benefit after 12-h sleep delay. Moreover, the present results show even a numerical benefit of restudy over retrieval practice, thus an inverted testing effect after 12-h sleep delay. Like in Experiment 1, this pattern of results was related to differences in sleep-associated memory consolidation depending on type of practice. Again, sleep effects were present after restudy but absent after retrieval practice. Thus, in addition to replicating Experiment 1, results of this experiment are again in line with findings of Abel and Bäuml (2012). Overall, the present results show the generalizability of these prior findings to other learning material, i.e. paired associates and can once more be accounted for by the bifurcation model (Halamish & Bjork, 2011).

According to the elaborative retrieval hypothesis (Carpenter, 2009; Pyc & Rawson, 2010) more retrieval effort should increase testing effects. In contrast, a comparison of the present results with the results of Experiment 1 did not reveal any differences in size of the testing effect. Moreover, like in the results of Experiment 1 and 2, the present finding of a reduced testing effect after sleep is not in accordance with the elaborative retrieval hypothesis that would predict at least preserved testing effects. Thus the relationship between the testing effect and sleep cannot be easily reconciled with the elaborative retrieval hypothesis. Similarly, regarding the episodic context account (Karpicke et al., 2014), greater difficulty during retrieval practice should demand greater context reinstatement resulting in increased testing effects compared to Experiment 1. Thus, the fact that testing effects are similar in size in the present experiment is not in line with predictions of the episodic context account.

3.2 Experiment 4

The first three experiments show that testing effects can be reduced after sleep delay as apparently only restudied item material profits from sleep-associated memory consolidation. Experiment 2 investigated the issue in the presence of retroactive interference at final test. Results show that testing effects can be increased in the presence but not in the absence of retroactive interference, which is in line with previous research on the issue (Abel & Bäuml, 2014; Halamish & Bjork, 2011; Potts & Shanks, 2012). Experiment 4 was implemented to investigate effects of retroactive interference on the testing effect. Thus, it was very similar to Experiment 2 but item material was changed from categorized items to paired associates. This was done to investigate the issue using item material typically used in studies on the testing effect possibly replicating previous results (see Experiment 3). Results of Experiments 1 and 2 did not suggest that sleep effects are increased by retroactive interference. Previous evidence on the issue is mixed (Ellenbogen, Hulbert, et al., 2006; Deliens et al., 2013). All previous studies on the issue used paired associates as item material. Thus, Experiment 4 might even offer further information regarding this issue.

As mentioned above, the results of Experiment 2 were not in accordance with the elaborative retrieval hypothesis predicting testing effects to be preserved or even increased after sleep. Yet, the retrieval-practice task used might have been too simple to induce effortful retrieval (success rates during retrieval practice: 93.25%), so drawing firm conclusion from these results might be premature. Experiment 4 will further analyze this issue by increasing difficulty of the retrieval-practice task through the use of more challenging learning material (paired associates). With difficulty of retrieval, according to the elaborative retrieval hypothesis, the testing effect should increase in size (e.g. Carpenter, 2009). Similarly, according to the episodic context account (Karpicke et al., 2014), increased difficulty of the retrieval-practice task should enhance testing effects (see above). However, in accordance with the bifurcation model (Halamish & Bjork, 2011) (see above), results of Experiment 3 suggest that increased difficulty during retrieval practice might not lead to increased testing effects. Experiment 4 investigates the issue

further by striving to replicate previous results.

Experiment 4 was designed in a similar way as Experiment 2, but item material was changed to paired associates (see Experiment 3). The learning phase was similar to Experiment 3 and retroactive interference was induced after the 12-h delay. Participants were asked to learn additional word pairs that always consisted of the stimulus words of the original learning list, while response items were changed. This was done to induce interference at the final test and increase recall difficulty. Goal of Experiment 4 was to replicate results of Experiment 2, i.e. increased testing effects under conditions of retroactive interference using different item material. As the retrieval practice task used in Experiment 2 might be to simple to induce effortful retrieval, paired associates were used to increase retrieval difficulty. According to the elaborative retrieval hypothesis and the episodic context account this should increase testing effects possibly altering the pattern of results observed in Experiment 2. However, results of Experiment 3 indicate that this shift in retrieval difficulty might not lead to differences in the size of the testing effect, which is in accordance with the bifurcation model. In conjunction with the results of Experiment 2, using different types of learning material, results of the current experiment will give information about the role of retroactive interference for the testing effect.

Method

Participants

144 healthy students from Regensburg University participated in the experiment (M = 21.9 years; 18-30 years; 20 male). Criteria for subject selection and were the same as in the previous experiments. Subjects in the 12-h sleep condition were instructed to try to sleep regularly during the night between sessions, while subjects in the 12-h wake condition were instructed not to take naps during the day. Every subject was asked to refrain from drinking alkoholic beverages between sessions. Once more, all participants were native German speakers and were distributed equally across conditions (n = 36 in each of the four condiditions).

Comparisons between conditions regarding subjects' age, habitual sleep duration, subjective ratings of sleep quality, and ratings on the Epworth Sleepiness Scale (Johns, 1991) did not reveal any differences (all $ps > .05$). As in Experiment 2, the connect-the-numbers-task was replaced by the presentation of the second item list (nontarget list).

Material

The same list of 16 paired associates that had already been used in Experiment 3 was applied as learning material (target list). 8 paired associates were used for low practice level while the other 8 paired associates were used for the high practice level. Sets of paired associates were distributed over practice levels in a counterbalanced manner. Additionally, to induce retroactive interference, a nontarget list with another unrelated and neutral 16 paired associates was constructed. These paired associates consisted of the 16 stimulus items of the original target list that were each randomly paired with one of another concrete German one- or two-syllable word (response items) from different semantic categories (Van Overschelde et al., 2004). Response items of the target list and those of the nontarget list had unique initial letters.

Design

The experiment had a $2 \times 2 \times 2$ mixed-factorial design with the between-subjects factors of TYPE OF PRACTICE (restudy, retrieval practice) and DELAY (12-h wake, 12-h sleep), and the within-subjects factor of PRACTICE LEVEL (low, high). Like in Experiment 2, to induce retroactive interference, participants studied the nontarget list at the beginning of the second session and there was no short-delay condition (see Figure 6 for an illustration of the experimental design of Experiments 2 and 4).

Procedure

The first session of the experiment (Study and practice phase) was conducted in exactly the same way as in Experiment 3. During study, paired associates were presented successively in a random order, at a presentation rate of 5 sec each. After one initial study cycle, participants either studied the paired associates on one or two restudy cycles (restudy condition) or on one or two retrieval-practice cycles. 8 of the paired associates were used for the low practice level, while the remaining paired associates were used for the high practice level. As in Experiment 3, in restudy conditions, subjects were presented with the paired associates for 5 sec each. Half of the paired associates were restudied once (SS), whereas the other half of paired associates was restudied twice (SSS). In retrieval conditions, subjects were provided with the stimulus words and word stems of the response items for up to 7 sec each, and were asked to recall the corresponding items (mean time until response: $M = 2.1$ sec; $SD = .68$). To allow for the simultaneous testing of two subjects, retrieval practice was again conducted in a covert fashion (see Experiment 3). Like previously, retrieval of the response item of half of the paired associates was practiced once (ST), whereas retrieval of the other half was practiced twice (STT). In both restudy and retrieval conditions, order of the paired associates was random. Like in Experiment 2, after the 12-h delay the nontarget list was presented. The list was presented three times at a rate of 5 sec per item; items were presented in a random order. This phase was followed by a 30-sec backward-counting distractor task. After this, the final test was conducted in an analogous manner to Experiment 3. The paired associates of the target list were tested first and the items of the nontarget list second. Concerning compliance with instructions, subjects in the 12-h sleep condition reported to have slept regularly during the experimental night ($M = 7.7$ hrs; $SD = 1.11$). Time spent asleep during the experimental night did not differ from habitual sleep time of participants in this condition ($p > .10$). Subjects in the 12-h wake condition reported not to have taken any naps during the day. None of the subjects reported alcohol intake between sessions.

Test Phase. At test, participants were first presented with the stimulus words and initial letters of the response items of all 16 paired associates of the target list for 7 sec each, and were asked to recall the appropriate response item. Afterwards they were asked to recall all of the response items of the nontarget list in the same way. Order of paired associates was always random.

Results

Success Rates during Retrieval-Practice Cycles

A 2 x 2 ANOVA with the factors of TIME OF DAY (9 a.m., 9 p.m.) and RETRIEVAL PRACTICE (ST, STT) showed that retrieval success was higher after two than after one retrieval-practice cycle (88.7% vs. 83.7%), $F(1, 70) = 12.21$, $MSE = 74.75$, $p = .001$, $\eta^2 = .15$. There was no main effect of TIME OF DAY and no interaction, $ps > .10$.

Final Test (12-h Delay Conditions)

Figure 10 shows recall performance after the 12-h delay. A 2 x 2 x 2 ANOVA with the factors of TYPE OF PRACTICE (restudy, retrieval practice), DELAY (12-h wake, 12-h sleep), and PRACTICE LEVEL (low, high) revealed that there were significant main effects of TYPE OF PRACTICE, $F(1, 140) = 18.19$, $MSE = 527.76$, $p < .001$, $\eta^2 = .12$, DELAY, $F(1, 140) = 9.28$, $MSE = 527.76$, $p = .003$, $\eta^2 = .06$, and PRACTICE LEVEL, $F(1, 140) = 51.54$, $MSE = 197.59$, $p < .001$, $\eta^2 = .27$. The main effect of TYPE OF PRACTICE once more reflects overall higher recall in the retrieval condition than in the restudy condition (57.9% vs. 46.4%), whereas the main effect of DELAY shows that recall rates were overall higher in the 12-h sleep condition than in the 12-h wake condition (56.3% vs. 48.0%); the main effect of PRACTICE LEVEL indicates that items were recalled better at the higher than the lower practice level (58.1% vs. 46.2%). More important, all other interactions being nonsignificant (all $ps > .05$), there was once more a reliable two-way interaction between TYPE OF PRACTICE and DELAY, $F(1, 140) = 14.07$, $MSE = 527.76$,

$p < .001$, $\eta^2 = .09$, indicating that the beneficial effect of retrieval practice in comparison to restudy was again modulated by delay condition. Consistently, planned comparisons revealed significant testing effects after the 12-h wake delay for both the lower practice level (30.9% vs. 53.8%), $t(70) = 5.83$, $p < .001$, $d = 1.37$, and the higher practice level (43.4% vs. 63.9%), $t(70) = 5.09$, $p < .001$, $d = 1.20$, but no reliable testing effect after the 12-h sleep delay, for the lower practice level (47.6% vs. 52.4%), $t(70) = .94$, $p = .352$, $d = .23$, and a slight numerical benefit for restudy for the higher practice level (63.5% vs. 61.5%), $t(70) = .44$, $p = .658$, $d = .10$.

Consistently, there was an effect of sleep-associated memory consolidation for items that had been restudied once (SS), $t(70) = 3.20$, $p = .002$, $d = .75$, and for items that had been restudied twice (SSS), $t(70) = 3.93$, $p < .001$, $d = .93$. In contrast, there was no effect of sleep-associated memory consolidation for items that had been practiced once (ST), $t(70) = .36$, $p = .723$, $d = .08$, and no effect for items that had been practiced twice (STT), $t(70) = .70$, $p = .484$, $d = .16$.

Recall for the nontarget list was analyzed in a 2 x 2 ANOVA with the factors of type of practice (restudy, retrieval practice) and delay (12-h wake, 12-h sleep). The analysis revealed no significant effects, all $Fs < 1.0$, indicating that there were no reliable differences between practice conditions (70.7% vs. 73.6%) and no reliable differences between delay conditions (71.1% vs. 73.2%).

Additional Analyses

To analyze the effect of retroactive interference on memory retention as a function of experimental condition, data of Experiments 3 and 4 was directly compared. This was done by means of a 2 x 2 x 2 x 2 ANOVA with the factors of type of practice (restudy, retrieval practice), delay (12-h wake, 12-h sleep), practice level (low, high), and interference (interference, no interference). This analysis again revealed significant main effects of type of practice, $F(1, 280) = 12.47$, $MSE = 600.73$, $p < .001$, $\eta^2 = .04$, delay, $F(1, 280) = 18.46$, $MSE = 600.73$, $p < .001$, $\eta^2 = .06$, and practice level, $F(1, 280) = 89.25$, $MSE = 216.68$, $p < .001$, $\eta^2 = .24$, as well as a significant interaction between

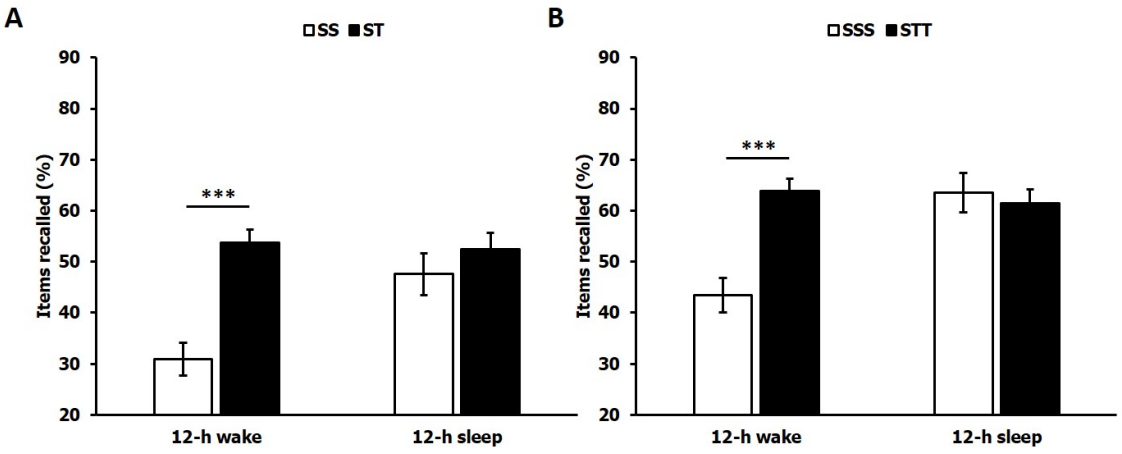

Figure 10: Mean recall performance on the final test of *Experiment 4* as a function of delay (12-h wake, 12-h sleep) and type of practice (restudy, retrieval practice), separately for the low (A) and the high (B) practice level. Condition labels indicate study (S) and retrieval practice (T) cycles. Error bars represent standard errors; $^{***}p < .001$.

type of practice and delay, $F(1, 280) = 23.52$, $MSE = 600.73$, $p < .001$, $\eta^2 = .08$. In addition, it revealed a significant main effect of interference, $F(1, 280) = 26.40$, $MSE = 600.73$, $p < .001$, $\eta^2 = .09$, indicating lower recall rates in Experiment 4 than in Experiment 3 (52.1% vs. 62.6%). Most important, there was a significant interaction between type of practice and interference, $F(1, 280) = 4.50$, $MSE = 600.73$, $p = .035$, $\eta^2 = .02$, indicating that the testing effect was influenced by interference. Indeed, planned comparisons over both delay conditions revealed that there was a significant testing effect for the lower practice level in the presence of interference (39.2% vs. 53.1%), $t(142) = 4.11$, $p < .001$, $d = .68$, but not in its absence (54.9% vs. 59.0%), $t(142) = 1.02$, $p = .311$, $d = .17$, and a testing effect for the higher practice level in the presence of interference (53.5% vs. 62.7%), $t(142) = 2.79$, $p < .001$, $d = .47$, but not in its absence (67.4% vs. 69.1%), $t(142) = .47$, $p = .620$, $d = .08$. No further effects were significant ($ps > .10$).

Comprehensive Analyses of Interference Effects

Retroactive interference and testing effects. In both Experiments 2 and 4 retroactive interference was induced after the 12-h delay to explore testing effects in the presence and absence of such interference. To further investigate these effects over different types of item material, results of all reported experiments after 12-h wake delay were compared (see Figure 11). This was done by means of a 2 x 2 x 2 ANOVA with the factors of TYPE OF PRACTICE (restudy, retrieval practice), PRACTICE LEVEL (low, high), INTERFERENCE (interference, no interference). Like before, this analysis revealed significant main effects of TYPE OF PRACTICE, $F(1, 284) = 94.20$, $MSE = 453.04$, $p < .001$, $\eta^2 = .25$, and PRACTICE LEVEL, $F(1, 284) = 81.81$, $MSE = 194.22$, $p < .001$, $\eta^2 = .22$. Additionally there was a main effect of INTERFERENCE, $F(1, 284) = 34.61$, $MSE = 453.04$, $p < .001$, $\eta^2 = .11$, indicating lower overall recall rates in Experiments 2 and 4 than in Experiments 1 and 3 (52.1% vs. 62.5%). Most important, there was a significant interaction between TYPE OF PRACTICE and INTERFERENCE, $F(1, 284) = 7.41$, $MSE = 453.04$, $p = .007$, $\eta^2 = .03$, indicating that the testing effect was reliably influenced by interference. Consistently, planned comparisons revealed that there was a testing effect for the lower practice level in the presence of interference (35.6% vs. 58.1%), $t(142) = 7.83$, $p < .001$, $d = 1.30$, and a reduced testing effect in its absence (51.6% vs. 62.8%), $t(142) = 3.19$, $p = .002$, $d = .53$, and a testing effect for the higher practice level in the presence of interference (46.5% vs. 68.0%), $t(142) = 7.69$, $p < .001$, $d = 1.28$, and a reduced testing effect in its absence (61.0% vs. 74.6%), $t(142) = 4.97$, $p < .001$, $d = .83$. No further effects were significant ($ps > .10$).

Retroactive interference and sleep effects. Analyses of recall rates during the final test show that there was no greater sleep effect in the presence of retroactive interference compared to in its absence, neither in Experiments 1 and 2 nor in Experiments 3 and 4. To further investigate this issue, only results after restudy but not retrieval practice of all reported experiments were analysed in a 2 x 2 x 2 ANOVA with the factors of DELAY (12-h wake, 12-h sleep), PRACTICE LEVEL (low, high), INTERFERENCE (interference, no interference). This analysis again revealed

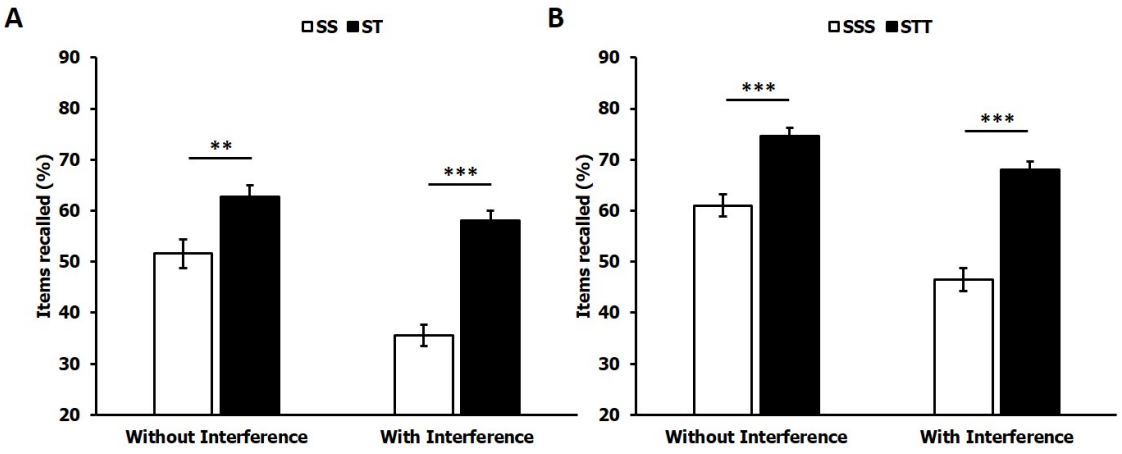

Figure 11: Mean recall performance on the final test after 12-h wake delay of *Experiments 1-4* as a function of interference (without interference, with interference) and type of practice (restudy, retrieval practice), separately for the low (A) and the high (B) practice level. Condition labels indicate study (S) and retrieval practice (T) cycles. Error bars represent standard errors; $^{**}p < .01$; $^{***}p < .001$.

significant main effects of DELAY, $F(1, 284) = 64.30$, $MSE = 566.73$, $p < .001$, $\eta^2 = .19$, PRACTICE LEVEL, $F(1, 284) = 85.31$, $MSE = 204.19$, $p < .001$, $\eta^2 = .23$, and INTERFERENCE, $F(1, 284) = 63.71$, $MSE = 566.73$, $p < .001$, $\eta^2 = .18$. No further effects were significant ($ps > .10$).

Discussion

Experiment 4 was conducted in order to investigate the relationship between sleep and the testing effect in the presence of retroactive interference, striving to replicate results of Experiment 2. However, in contrast to Experiment 2 paired associates were used as learning material, conforming to previous studies on the testing effect. As the difficulty of the retrieval-practice task in Experiment 2 was quite low, with high success rates during retrieval practice (mean: 93.25%), goal of Experiment 4 was to raise retrieval difficulty. As intended, the learning of paired associates proved to be more challenging resulting in lower success rates during retrieval practice in this experiment (mean: 86.2%), indicating that more effort was needed to retrieve the items.

Like in the first three experiments, the present results show that testing effects can be reliably reduced after 12-h sleep delay compared to 12-h wake delay. Again, this was observable independent from practice level (with even a slight advantage for restudy on the high practice level). Mean benefit of retrieval practice over restudy over both practice levels was 21.7% after 12-h wake delay and was reduced to 1.4% after 12-h sleep delay. This pattern of results replicates data from the first experiments and has its origin in the fact that only restudied items could profit reliably from sleep, while there was no sleep-associated benefit for items subjected to retrieval practice. So results of Experiment 4, like results of Experiment 3, show that findings from the first two experiments can be generalized to different item material, i.e. paired associates.

Goal of the present experiment was to investigate the relationship of testing effects and retroactive interference. In fact, testing effects have been found to rise in the presence of retroactive interference compared to its absence (Abel & Bäuml, 2014; Halamish & Bjork, 2011; Potts & Shanks, 2012). According to the bifurcation model (Halamish & Bjork, 2011) recall threshold is lifted by higher final-test difficulty so that mainly restudied items fall below this boundary and can no longer be retrieved, increasing the gap between items subjected to restudy versus retrieval practice. Thus, retrieval practice can stabilize memories, reducing susceptibility to such interference (see Figure 5). Results of Experiment 2 are in line with this empirical work, as testing effects were increased in the presence of interference. Likewise, a comparison between data of Experiments 3 and 4 shows that, over both delay conditions and practice levels, there was a significant testing effect in the presence of interference (11.6%) but a reduced and nonsignificant testing effect in its absence (2.9%).

Regarding the relationship of sleep and retroactive interference, prior studies found sleep effects in the presence of retroactive interference but reduced sleep effects in its absence (Ellenbogen, Hulbert, et al., 2006; Ellenbogen et al., 2009). Thus, susceptibility to retroactive interference was reliably reduced after nocturnal sleep. Contrasting evidence from Deliens et al. (2013), on the other hand, suggests that effects of retroactive interference might be even more pronounced after sleep versus wake delay. The present results fall in between these two contradictory

findings, as sleep effects remain at approximately the same level in the presence and absence of interference. This is in line with results of Experiment 2 showing no difference in effects in the presence of retroactive interference. Thus, type of study material cannot account for these prior results.

3.3 Summary

While Experiments 1 and 2 used categorized item material to investigate the relationship between sleep-associated memory consolidation and the testing effect, paired associates were implemented as learning material in Experiments 3 and 4. The first two experiments aimed to further investigate the finding of a reduced sleep benefit after retrieval practice (Abel & Bäuml, 2012) in connection with the testing effect. Overall they show that testing effects can be reduced after 12-h sleep delay, as only restudied items profit reliably from sleep. This observation goes beyond prior work and is interesting in the framework of research on the sleep effect as well as the testing effect. A possible flaw of these findings might be that, in previous studies on these effects, usually paired associates were used as learning material (e.g. Allen et al., 1969; Gais et al., 2006; Karpicke & Roediger, 2008; Tucker et al., 2006). Therefore, and to replicate data from the first two experiments with differing learning material, paired associates replaced categorized items in Experiments 3 and 4. The results show the generalizability of the previous observations to paired associates. Indeed, similar to the first two experiments, Experiments 3 and 4 found reliable testing effects over both practice levels after 12-h wake delay (mean: 17.1%) but numerically inverted testing effects (benefit of restudy over retrieval practice) after 12-h sleep delay (mean: -2.8%).

The results of Experiments 3 and 4 are again in line with the bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011). Like in the first two experiments, the reduction of testing effects after 12 h of nocturnal sleep were mainly caused by the fact that only restudied items showed reliable sleep-related benefits. If, after initial study, participants were asked to try to actively retrieve the items, there was hardly any benefit of sleep versus wake delay. Consistently, the gap between items subjected to restudy versus retrieval practice (i.e. the testing effect) was considerably reduced after nocturnal sleep. According to the bifurcation model, this can be explained with the bifurcated distribution of item strength after retrieval practice (Kornell et al., 2011). Experiments 3 and 4, like the first two experiments, are well in line with this theoretical framework and show that it can provide a solid groundwork for the comprehension of the relationship

between sleep and retrieval processes.

In accordance with the desirable difficulty framework (Bjork, 1994, 1999), the elaborative-retrieval hypothesis (Carpenter, 2009; Pyc & Rawson, 2010) suggests that benefits of retrieval practice are directly connected to the retrieval difficulty during such practice. Testing effects are, in this theoretical framework, induced by increased elaboration during active retrieval processes compared to simple restudy of items. Due to a promotion of semantic networks around the to-be-retrieved items, memory for items subjected to retrieval practice is fostered at final test. Experiments 3 and 4 aimed to further analyze the role of elaborative retrieval for the testing effect by increasing difficulty during retrieval practice with reduced success rates as a consequence. With the relatively simple retrieval-practice task implemented in the first two experiments, elaboration should have been very low not contributing much to the benefits of retrieval practice. Similarly, according to the episodic context account (Karpicke et al., 2014), increased difficulty during retrieval practice should foster efforts to reinstate the initial learning context, resulting in increased testing effects. However, increased difficulty during retrieval practice in Experiments 3 and 4 did not considerably change the pattern of results observed in the first two experiments. Thus, the present findings cannot be explained by means of the elaborative-retrieval hypothesis or the episodic context account. Possible reasons and suggestions for future research are discussed below.

Chapter 4

General Discussion

4.1 The Influence of Sleep on the Testing Effect

In a series of 4 experiments, the present work investigated the influence of sleep delay on the testing effect. Based on a prior study by Abel and Bäuml (2012) Experiments 1 and 2 employed categorized item material, while Experiments 3 and 4 used paired associates as learning material to resemble material used in classical studies on the testing effect. Over all 4 experiments there was a reliable testing effect after 12-h wake delay (mean testing effect after 12-h wake delay across all 4 experiments: 17.2%), replicating prior work (Hogan & Kintsch, 1971; Roediger & Karpicke, 2006b). In line with previous research on the testing effect, such effects were observable both after one and after two practice trials. Besides being an effective tool in reducing time-dependent forgetting, retrieval practice has previously been found to protect memories from the detrimental effects of retroactive interference (e.g. Abel & Bäuml, 2014; Potts & Shanks, 2012). Interference effects after the 12-wake delay were analyzed over all reported experiments. Results reveal reliable testing effects in the presence of interference (mean testing effect after wake delay in the presence of interference and over both practice levels: 22.0%) but reduced testing effects in its absence (mean testing effect after wake delay in the absence of interference and over both practice levels: 12.5%). Thus, besides replicating studies showing that retrieval practice can effectively reduce time-dependent forgetting, the present pattern of results replicates even prior findings on the testing effect and retroactive interference. Additionally, while the testing effect has reliably been reported to be present after longer delays, there have been studies showing that it can be reduced or even reversed (i.e. better recall after restudy) after shorter delays (test-delay interaction; see Roediger & Karpicke, 2006b; Toppino & Cohen, 2009; Wheeler et al., 2003). The present results replicate this finding, as testing effects are present after 12-h wake delay, but absent after shorter delays of 20 minutes (Experiments 1 and 3).

Moreover, the present work goes beyond prior work as it explored effects of retrieval practice versus restudy after sleep delay. While prior studies have

investigated testing effects with delays of varying duration (see above), the influence of sleep on the testing effect has not been explored so far. In contrast, the typical testing-effect studies have usually employed retention intervalls filled both with wake and sleep periods. Research on sleep-associated memory consolidation suggests that there are considerable differences between mnemonic influences of wake versus sleep delay (Diekelmann et al., 2009; Diekelmann & Born, 2010; Rasch & Born, 2013). The results after restudy of the item material replicate prior studies on the sleep effect in all 4 experiments, as items were consistently recalled better after sleep than after wake delay. In fact, time-dependent forgetting was constantly reduced after sleep independent of levels of practice (mean sleep effect after restudy over all 4 experiments and both practice levels: 15.9%). However, when applying retrieval practice during the learning phase such reliable sleep-related benefits where no longer observable at the final test. Indeed, results indicate that this was consistently the case over all 4 experiments and all practice levels (mean sleep effect after retrieval practice over all 4 experiments and both practice levels: 0.8%). This mirrors findings by Abel and Bäuml (2012) who, investigating the influence of sleep on retrieval-induced forgetting, reported no effects of sleep-associated memory consolidation for items that had been subject to retrieval practice (RP+). Moreover, the present work extends these findings by providing a restudy control group and, thus, enables investigations on sleep's influence on the testing effect. Accordingly, the present findings suggest that sleep compared to wake delay can reliably modify the size of the testing effect. Interestingly, and consistent over all reported experiments, testing effects were reduced after a night of nocturnal sleep compared to wake delay. In fact, testing effects were even eliminated after 12-h sleep delay in most experiments (Experiments 1, 3, and 4).

4.2 Theoretical Implications regarding the Testing Effect

Goal of the present work was to find out more about sleep's influence on the testing effect. Results provide interesting insights regarding theoretical frameworks

that are commonly used to explain benefits of retrieval practice over restudy. As has been mentioned before, the distribution-based bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011) might be capable of providing an explanation for the present pattern of results. According to the model, at the final test, recall performance is determined in an all-or-none fashion, as recall threshold is supposed to separate successful and unsuccessful retrieval depending on items' memory strength. Items that are successfully retrieved during the practice phase are supposed to be strengthened by a greater margin than restudied items. This results in more items being still above threshold at the final test after wake delay than if restudy is applied - the typical testing effect. However, sleep-associated memory consolidation is supposed to strengthen memories, benefiting mainly restudied items enabling them to cross recall threshold. In contrast, items that were subject to retrieval practice and successfully retrieved are already strengthened to a level high above this recall threshold. Thus, these items might further gain in strength through sleep but this, according to the model, will not be measured by the test (see Figure 2). Hence, the results of the present experiments, showing testing effects after 12-h wake delay but reduced testing effects after 12-h sleep delay, are in line with the bifurction model.

Besides exploring sleep's influence on the testing effect the present work even aimed to investigate both the testing effect and the sleep effect in the presence and absence of retroactive interference. Previous studies show that retrieval practice but not restudy can reduce items' susceptibility to retroactive interference, resulting in enhanced testing effects in the presence but not in the absence of retroactive interference (Abel & Bäuml, 2014; Potts & Shanks, 2012). This underscores the power of testing, not only leading to less time-dependent forgetting but also stabilizing memories to be less vulnerable to interference. Halamish and Bjork (2011) showed that even this effect fits well into the bifurcation model. Indeed, retroactive interference at final test might, according to the model, mainly be detrimental for restudied items and less for items subjected to retrieval practice. After retrieval practice items are supposed to be high above recall threshold and might therefore not suffer as much from interference effects (see Figure 5). Consistent with this presumption, the present results show that, independent of

practice level, testing effects are influenced by retroactive interference. Indeed, while there were reliable testing effects in the presence of retroactive interference, testing effects were reduced in its absence. This was observable when using categorized item lists (Experiments 1 and 2) as well as when using paired associates as study material (Experiments 3 and 4). Thus, the results of the present work are in line with previous studies emphasizing the power of testing not only to reduce time-dependent forgetting but even to stabilize memories against detrimental effects of retroactive interference.

The bifurcation model offers even a reasonable explanation for the existence of the above mentioned test-delay interaction. At shorter delays, restudied items might still be above recall threshold by a large margin, possibly resulting in advantages of restudy over retrieval practice or at least preventing reliable testing effects. At longer delays, mainly items that were successfully retrieved during retrieval practice remain above threshold, while most of the restudied items cross below threshold, leading to a testing effect (see Halamish & Bjork, 2011). Thus, the present results after wake delay are in line with the bifurcation model as they reveal a clear test-delay interaction. In contrast, after 12-h sleep delay, this test-delay interaction is undermined by reduced benefits of retrieval practice over restudy. This is supposedly caused by sleep-related strenghening, leading to a greater margin of restudied items above recall threshold (see Figure 2). Thus, the present results replicate prior findings regarding the testing effect and retroactive interference and test- delay interaction which is in line with assumptions made by the bifurcation model. Moreover, they go beyond previous work showing that sleep might reduce the testing effect, even that being in accordance with the bifurcation model (see above).

In contrast, the present findings are less well in line with the elaborative retrieval hypothesis (Carpenter & Delosh, 2006; Pyc & Rawson, 2009). According to this hypothesis, testing effects are the result of processes caused by effortful retrieval during the retrieval-practice phase. In contrast, restudy is supposed to result in lower recall rates, because it is a rather passive and effortless way of learning. Retrieval effort has been linked to the activation of semantic networks, creating additional recall routes that can be useful at the final test (Carpenter,

2009; Pyc & Rawson, 2010, 2012). As sleep has been found to have a similar effect on memories (Cai et al., 2009; Darsaud et al., 2011; McKeon et al., 2012; Payne et al., 2009), according to this hypothesis, one might expect sleep to foster beneficial effects of retrieval practice. As a result, testing effects should be preserved or even enhanced after sleep delay. The present finding of a reduced testing effect after sleep cannot be easily explained by means of this theoretical framework. As retrieval effort is linked to the difficulty of the retrieval-practice task, with higher difficulty evoking greater effort, easy retrieval-practice tasks might not suffice to produce reliable testing effects. As the retrieval-practice task was quite easy in Experiments 1 and 2, a more difficult task was applied in Experiments 3 and 4. Still, success rates during retrieval-practice cycles were quite high (mean: 85.3%). However, there were reliable testing effects after wake delay in all of the 4 experiments, suggesting that elaboration must have been sufficient or, alternatively, that elaboration and additional semantic processing may not have caused testing effects in the present work.

The third account on the testing effect, the episodic context account (Karpicke et al., 2014), offers a context-based explanation for benefits of retrieval practice over restudy. According to this account, testing improves recall performance by reinstating the original encoding context of the learning phase, updating the memory representation with contextual information from the new temporal context during retrieval practice, finally resulting in a restricton of the search set during final test. There is no evidence on how sleep might affect the size of the testing effect according to this account, so one might assume that the influence is negligible, i.e. testing effects remain unaffected. However, findings of a study by Cairney et al. (2011) suggest that dependence on contextual cues during final recall might be reduced after sleep versus wake delay. Additionally sleep has been found to foster reactivation of learning material, promoting the integration of new information into existing mnemonic networks (Rasch et al., 2007). Thus, another prediction of this account might be that testing effects are reduced after sleep, as benefits of retrieval practice over restudy might be less evident. Thus, the results of Experiment 1 and 2 are not only in line with the bifurcation model (see above) but even the episodic context account.

However, regarding difficulty of the retrieval-practice task, increased difficulty should, according to this account, result in greater testing effects (similar to predictions by the elaborative retrieval hypothesis). Higher difficulty is supposed to foster processes of context-reinstatement during retrieval that should result in increased testing effects after wake delay. As retrieval difficulty was increased in Experiments 3 and 4, but testing effects remained on a similar level as in Experiments 1 and 2, this is not in accordance with the episodic context account. Thus, one can suppose that retrieval difficulty was not great enough even in Experiments 3 and 4 but, again, reliable testing effects after wake delay were observable in all 4 experiments of the present work. This suggests that retrieval difficulty should have been sufficient. Accordingly, the reduced testing effect after sleep delay is in line with assumptions of the episodic context account. In contrast, the absence of increased testing effects in the face of increased retrieval difficulty cannot easily be conciliated with this account.

4.3 Theoretical Implications regarding Sleep-Associated Memory Consolidation

Early studies on sleep-associated memory consolidation have associated the benefits of sleep with a shelter passively protecting memories from interference that usually builds up during wake intervals (Jenkins & Dallenbach, 1924). This has, over time, been called into question by several researchers who provided evidence for active processes fostering memories during sleep. Clearly speaking in favor of such active processes, sleep has been found to be selectively beneficial for some types of memories and less so for other types. For instance, sleep has been found to mainly benefit relevant (Van Dongen et al., 2012; Wilhelm et al., 2011) and emotional memories (Payne et al., 2008; Wagner et al., 2006). Indeed, if sleep would work only like a passive shelter, it should benefit all types of memories equally (see even Ellenbogen, Payne, & Stickgold, 2006). The present results are further support for active processes leading to sleep-related mnemonic benefits. All 4 experiments reported in the present work found sleep-associated memory

consolidation to reliably benefit memories after restudy but no such benefits after retrieval practice. Thus memories were selectively enhanced depending on the type of practice applied. However, bearing in mind the assumptions made by the bifurcation model, one could also argue that equal benefits after restudy and retrieval practice might likewise have resultet in the present pattern of results. All items are supposed to be recalled at final test, if only their memory strength is sufficient to cross recall threshold. However, strengthening beyond threshold will, according to the model, not lead to better recall performance. Thus, there might be sleep-related strengthening after retrieval practice, maybe even resembling the amount of strengthening after restudy. Still, this might only lead to better recall for restudied items at the final test, which fits the present results.

Another study, speaking in favor of selectice sleep effects, investigated the relationship between memory strength and sleep-associated memory consolidation (Drosopoulos et al., 2007). Results revealed that there were sleep effects for items with low memory strength (60% initially recalled), but no such sleep effects for items with high memory strength (90% initially recalled). To let participants study until they reached the desired learning criterion, the anticipation-plus-study method with retrieval-practice trials followed by feedback was used. The present study disentangled effects of restudy and retrieval practice to investigate how sleep influences item material that has been subjected to different types of practice. Additionally, different levels of practice were used to investigate the role of memory strength for sleep-related mnemonic benefits. In contrast to the findings by Drosopoulos et al. (2007), in all reported experiments, sleep effects were not influenced by level of practice (i.e. memory strength) but rather by type of strengthening (restudy versus retrieval practice). Thus, results are in accordance with another recent study by Abel and Bäuml (2012). Reason for this disparity might be rooted in methodological differences, i.e. differences during the learning phase. Prior evidence shows that retrieval practice with feedback might lead to mediated testing effects, and thereby to very different results than when feedback is not provided (Pyc & Rawson, 2012). Hence, sleep might lead to benefits after retrieval practice *with* feedback but not *without* feedback trials. This might even explain why typical studies on the sleep effect found reliable sleep-associated

benefits even after retrieval practice cycles. In fact, these studies (like Drosopoulos et al., 2007) usually applied the anticipation-plus-study method, not separately analyzing effects of restudy and retrieval practice.

Finally, aim of the present work was not only to investigate the influence of retroactive interference on the testing effect but even how it can affect the sleep effect. So far evidence regarding this issue is mixed. Ellenbogen, Hulbert, et al. (2006) showed that sleep can protect memories against the effects of such interference, i.e. there were reliable interference effects after wake delay but reduced effects after sleep delay (see even Ellenbogen et al., 2009). However, Deliens et al. (2013) report quite different results in a study investigating the issue. In fact, they found even an opposite effect with stronger interference effects after sleep delay than after wake delay. The results of the present work fall in between this contrasting evidence, with comparable interference effects after sleep and after wake delay. Indeed, analyzes of the results provide no indications that sleep might protect memories from detrimental effects of retroactive interference. This, might be partly explainable by methodological differences between the present work and the study by Ellenbogen, Hulbert, et al. (2006). Still, the present results indicate that sleep might not reduce memories' interference susceptibility under all circumstances.

4.4 Future research perspective

The present work was conducted to investigate the influence of sleep on the testing effect. In a series of 4 experiments, memories were analyzed after different types of practice (restudy vs. retrieval practice), different types of delay (wake delay vs. sleep delay), and in the presence or absence of retroactive interference. Study material varied over the experiments with categorized items as well as paired associates being used. Results show that findings of prior studies could be replicated as, overall, retrieval practice resulted in better final memory recall than restudy (testing effect) and memory performance was better after sleep delay than after wake delay (sleep effect). Interestingly, the present results go beyond

prior work showing that sleep can influence the testing effect, as predicted by the bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011). An interesting question that should be tackled by future research is how extended delay intervals might influence the observed pattern of results. Indeed, if assumptions of the bifurcation model are realistic, expanding delays to considerably longer duration than 12 hours should at some point result in testing effects, even if initial learning is followed by a night of sleep. Indeed, with longer delays memory strength should decrease considerably, so that even items successfully retrieved during the practice phase should fall below recall threshold. As a result, sleep effects would be measurable even after retrieval practice, and consistently, testing effects might be reestablished even if sleep follows the practice phase (see Figure 2, right bottom panel). Thus, longer delays might alter the present pattern of results being an effective way to further investigate the role of bifurcated distributions for the test-delay interaction and the relationship between sleep and the testing effect.

As mentioned above, sleep has been found to be selectively beneficial for some types of memories and less so for other types (Payne et al., 2008; Van Dongen et al., 2012; Wagner et al., 2006; Wilhelm et al., 2011). This has earlier been seen as one of the many indicators of active consolidation processes during sleep. Even the present set of results appears to speak in favor of such an approach, showing differences in sleep effects to rely on type of initial practice of study material. However, when analyzing the present results on the basis of the bifurcation model sleep effects after retrieval practice and restudy might even be similar in size. Indeed, gains in memory strength might not be distinguishable when they appear at a strength level high above recall threshold (see above). Thus, the findings of the present work remain silent about whether consolidation processes during sleep might differ as a function of type of practice. Future research might further investigate the role of selective sleep associated memory consolidation for the present results.

The present results remain even silent about cognitive processes resulting in beneficial mnemonic effects of retrieval practice. However, the overall observed pattern of results with reduced testing effects after sleep does not appear to be in line with the elaborative retrieval hypothesis (Carpenter & Delosh, 2006). As

success rates during retrieval practice were quite high, this might have been caused by low degree of elaborative processing. However, even increased difficulty during retrieval practice did not change this overall pattern of results. Future research should further investigate the role of elaboration during retrieval practice for the relationship between sleep and the testing effect. For instance, increasing retrieval difficulty during the practice phase to an even higher level than in Experiments 3 and 4 might be a possible way to further assess this in the present set of results. However, increasing difficulty to a very high degree might again result in reduced testing effects caused by very low success rates. Another process-related account, the episodic context account (Karpicke et al., 2014), offers a new and interesting explanation for possible processes resulting in mnemonic benefits of retrieval practice and might even offer an explanation for the observed pattern of results. However, the present work is not clearly designed to explore the nature of such processes and therefore cannot support or disprove either of these accounts. Future work should tackle this question possibly leading to better knowledge about how retrieval practice can foster memories.

To clearly distinguish between effects of restudy and retrieval practice, the present experiments did not provide feedback after retrieval-practice trials. Consistently, being in line with this approach, the bifurcation model describes a situation where only direct effects of retrieval practice are assessed. However, several studies on the testing effect have used feedback after every trial, thereby reliably increasing the size of the testing effect (Roediger & Butler, 2011). Moreover, if feedback is provided, mediated effects of retrieval practice might facilitate subsequent restudy, e.g. through a more effective encoding strategy (see Cull, 2000; Roediger & Karpicke, 2006a). Consistently, the mediator shift hypothesis (Pyc & Rawson, 2012) suggests that test–feedback practice is beneficial because retrieval failures - as opposed to successfull retrieval - during practice allow to evaluate the effectiveness of mediators enabling a shift from less effective to more effective mediators. Thus, if elaborative processes lead to more effective mediators between cue and target memory (Pyc & Rawson, 2010), such elaboration might be more important for mediated than for direct effects of testing. Consistently, sleep might have very distinguished influences on direct versus mediated testing effects.

While direct effects of testing are reduced, mediated effects may be maintained or even enhanced by sleep. Overall, future studies should investigate this issue by exploring sleep's influence on mediated effects of testing.

Thus, future studies might increase knowledge about cognitive processes active during retrieval practice. Additionally, they might further investigate possible parallels between sleep effect and testing effect. Research on the sleep effect shows that sleep actively stimulates processes benefitting memories (see above). Prior work on sleep-related memory consolidation shows that memories might even be reactivated during sleep, fostering later recall (Peigneux et al., 2004). If information is learned in connection with an odor or auditory cue, reapplication of such cues during sleep can foster sleep-associated memory consolidation (Rasch et al., 2007; Rudoy et al., 2009). If and how such reactivation processes and the resulting mnemonic benefit relate to processes during active retrieval of mnemonic contents during wake periods might be an interesting question of future work. One might, for instance, explore neuronal processes active during sleep and retrieval practice to find out more about possible similarities and differences. It could be interesting to investigate if and how possible reactivation processes during sleep and retrieval practice might interact with each other. Overall, results of such work might provide new information about the nature of mnemonic processes and stimulate further research about how they might be optimized.

Interestingly, while replicating earlier findings on the influence of retroactive interference on the testing effect, the present results do not show a relationship between sleep and retroactive interference. Thus, the results of the present work fall in between previous findings of reduced effects of retroactive interference (Ellenbogen, Hulbert, et al., 2006) and increased effects of retroactive interference after sleep (Deliens et al., 2013). Future research should even target this issue and try to detect boundary conditions of the power of sleep to protect against detrimental effects of retroactive interference. Potential results might offer valuable information, i.e. regarding the question of possible parallel or differential mnemonic processes active during sleep and retrieval practice.

4.5 Conclusions

Taking recent studies on the testing effect and on sleep-associated memory consolidation into consideration, the present work aimed to disentangle their mnemonic effects and to investigate potential interactions. Results do not only replicate prior work showing reliable testing effects and sleep effects, but go beyond prior work by showing that sleep might reduce the size of the testing effect. This finding can give valuable empirical information about theoretical frameworks regarding the testing effect underscoring the importance of future research on the issue to gain further knowledge about possible boundary conditions of each theoretical explanation. Additionally, regarding research on the sleep effect the results of the present work can also inspire future research on processes fostering active sleep-associated memory consolidation. The present results might even be interesting for practical appliance of retrieval practice and sleep as tools to improve memory performance in e.g. educational contexts. Overall, both the testing effect and the sleep effect are still underestimated in their impact on memory and the present results might promote for future research on how they interact to improve knowledge about adaptive and effective learning strategies.

References

Abbott, E. E. (1909). On the analysis of the factor of recall in the learning process. *Psychological Monographs: General and Applied*, *11*, 159–177.

Abel, M., & Bäuml, K.-H. T. (2012). Retrieval-induced forgetting, delay, and sleep. *Memory*, *20*, 420–428.

Abel, M., & Bäuml, K.-H. T. (2013a). Sleep can eliminate list-method directed forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 946–952.

Abel, M., & Bäuml, K.-H. T. (2014). The roles of delay and retroactive interference in retrieval-induced forgetting. *Memory & Cognition*, *42*, 141–150.

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

Alberini, C. M., Milekic, M. H., & Tronel, S. (2006). Mechanisms of memory stabilization and destabilization. *Cellular and Molecular Life Sciences*, *63*, 999–1008.

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

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, *49*, 415–445.

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

Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. *Memory*, *22*, 237–313.

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of educational research*, *61*, 213–238.

Barnes, J. M., & Underwood, B. J. (1959). "Fate" of first-list associations in transfer theory. *Journal of experimental psychology*, *58*, 97–105.

Barrett, T. R., & Ekstrand, B. R. (1972). Effect of sleep on memory. 3. Controlling for time-of-day effects. *Journal of Experimental Psychology*, *93*, 321–327.

Bäuml, K.-H. T., & Dobler, I. M. (2015). The two faces of selective memory retrieval: Recall specificity of the detrimental but not the beneficial effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 246–253.

Bäuml, K.-H. T., & Kliegl, O. (2013). The critical role of retrieval processes in release from proactive interference. *Journal of Memory and Language*, *68*(1), 39–53.

Bäuml, K.-H. T., & Samenieh, A. (2010). The two faces of memory retrieval. *Psychological Science*, *21*, 793–795.

Bäuml, K.-H. T., & Samenieh, A. (2012). Selective memory retrieval can impair and improve retrieval of other memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 488–494.

Bäuml, K.-H. T., & Schlichting, A. (2014). Memory retrieval as a self-propagating process. *Cognition*, *132*, 16–21.

Benson, K., & Feinberg, I. (1977). The beneficial effect of sleep in an extended jenkins and dallenbach paradigm. *Psychophysiology*, *14*, 375–384.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The loyola symposium* (pp. 123–149). Hillsdale.

Bjork, R. A. (1989). Retrieval inhibition as an adaptive mechanism in human memory. In H. Roediger III & F. Craik (Eds.), *Varieties of memory & consciousness: Essays in honour of endel tulving* (pp. 309–330). Hillsdale.

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

Born, J., & Wilhelm, I. (2012). System consolidation of memory during sleep. *Psychological research*, *76*, 192–203.

Bower, G. H. (1981). Mood and memory. *American psychologist*, *36*, 129–148.

Bäuml, K.-H., Pastötter, B., & Hanslmayr, S. (2010). Binding and inhibition in episodic memory — Cognitive, emotional, and neural processes. *Neuroscience & Biobehavioral Reviews*, *34*, 1047–1054.

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., Marsh, E. J., Goode, M. K., & Roediger, H. L., III. (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology*, *20*, 941–956.

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*, 514–527.

Butler, A. C., & Roediger, H. L., III. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*, 604–616.

Cai, D. J., Mednick, S. A., Harrison, E. M., Kanday, J. C., & Mednick, S. (2009). Rem, not incubation, improves creativity by priming associative networks. *Proceedings of the National Academy of Sciences*, *106*, 10130–10134.

Cairney, S. A., Durrant, S. J., Musgrove, H., & Lewis, P. A. (2011). Sleep and environmental context: interactive effects for memory. *Experimental brain research*, *214*, 83–92.

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

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, *14*, 474–478.

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

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

Chan, J. C. (2009). When does retrieval induce forgetting and when does it induce facilitation? implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language*, *61*, 153–170.

Chan, J. C., McDermott, K. B., & Roediger, H. L., III. (2006). Retrieval-induced facilitation: initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*, 553–571.

Cohen, D. A., Pascual-Leone, A., Press, D. Z., & Robertson, E. M. (2005). Off-line learning of motor skill memory: A double dissociation of goal and movement. *Proceedings of the National Academy of Sciences USA*, *102*, 18237–18241.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*, 407–428.

Cuban, L. (1993). *How teachers taught: Constancy and change in american classrooms, 1890-1990*. Teachers College Press.

Cuddy, L. J., & Jacoby, L. L. (1982). When forgetting helps memory: An analysis of repetition effects. *Journal of Verbal Learning and Verbal Behavior*, *21*, 451–467.

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

Darsaud, A., Dehon, H., Lahl, O., Sterpenich, V., Boly, M., Dang-Vu, T., ... others (2011). Does sleep promote false memories? *Journal of Cognitive Neuroscience*, *23*, 26–40.

Daurat, A., Terrier, P., Foret, J., & Tiberge, M. (2007). Slow wave sleep and recollection in recognition memory. *Consciousness and cognition*, *16*, 445–455.

Davies, G. (1986). Context effects in episodic memory: A review. *Cahiers de Psychologie*, *6*, 157–174.

Delaney, P. F., Verkoeijen, P. P., & 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.

Deliens, G., Schmitz, R., Caudron, I., Mary, A., Leproult, R., & Peigneux, P. (2013). Does recall after sleep-dependent memory consolidation reinstate sensitivity to retroactive interference? *PLoS ONE*, *8*, e68727.

Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience*, *11*, 114–126.

Diekelmann, S., Wilhelm, I., & Born, J. (2009). The whats and whens of sleep-dependent memory consolidation. *Sleep Medicine Reviews*, *13*, 309–321.

Drosopoulos, S., Schulze, C., Fischer, S., & Born, J. (2007). Sleep's function in the spontaneous recovery and consolidation of memories. *Journal of Experimental Psychology*, *136*, 169–183.

Drosopoulos, S., Wagner, U., & Born, J. (2005). Sleep enhances explicit recollection in recognition memory. *Learning & Memory*, *12*, 44–51.

Duchastel, P. C., & Nungester, R. J. (1982). Testing effects measured with alternate test forms. *The Journal of Educational Research*, 309–313.

Ellenbogen, J. M., Hulbert, J. C., Jiang, Y., & Stickgold, R. (2009). The sleeping brain's influence on verbal memory: boosting resistance to interference. *PLoS One*, *4*, e4117.

Ellenbogen, J. M., Hulbert, J. C., Stickgold, R., Dinges, D. F., & Thompson-Schill, S. L. (2006). Interfering with theories of sleep and memory: Sleep, declarative memory, and associative interference. *Current Biology*, *16*, 1290–1294.

Ellenbogen, J. M., Payne, J. D., & Stickgold, R. (2006). The role of sleep in declarative memory consolidation: passive, permissive, active or none? *Current Opinion in Neurobiology*, *16*, 716–722.

Empson, J. A. C., & Clarke, P. R. F. (1970). Rapid eye movements and remembering. *Nature*, *227*, 288–289.

Eschenko, O., Ramadan, W., Mölle, M., Born, J., & Sara, S. J. (2008). Sustained increase in hippocampal sharpwave ripple activity during slow-wave sleep after learning. *Learning & Memory*, *15*, 222–228.

Ficca, G., Lombardo, P., Rossi, L., & Salzarulo, P. (2000). Morning recall of verbal material depends on prior sleep organization. *Behavioural Brain Research*, *112*, 159–163.

Ficca, G., & Salzarulo, P. (2004). What in sleep is for memory. *Sleep medicine*, *5*, 225–230.

Fischer, S., Drosopoulos, S., Tsen, J., & Born, J. (2006). Implicit learning-explicit knowing: a role for sleep in memory system interaction. *Journal of Cognitive Neurosciences*, *18*, 311–319.

Fogel, S. M., Smith, C. T., & Cote, K. A. (2007). Dissociable learning-dependent changes in rem and non-rem sleep in declarative and procedural memory systems. *Behavioural brain research*, *180*, 48–61.

Foster, H. H. (1901). The necessity for a new standpoint in sleep theories. *The American Journal of Psychology*, 145–177.

Gais, S., Albouy, G., Boly, M., Dang-Vu, T. T., Darsaud, A., Desseilles, M., ... Peigneux, P. (2007). Sleep transforms the cerebral trace of declarative memories. *Proceedings of the National Academy of Sciences*, *104*, 18778–18783.

Gais, S., Lucas, B., & Born, J. (2006). Sleep after learning aids memory recall. *Learning & Memory*, *13*, 259–262.

Gais, S., Plihal, W., Wagner, U., & Born, J. (2000). Early sleep triggers memory for early visual discrimination skills. *Nature Neuroscience*, *3*, 1335–1339.

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

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *40*, 1–114.

Giuditta, A., Ambrosini, M. V., Montagnese, P., Mandile, P., Cotugno, M., Zucconi, G. G., & Vescia, S. (1995). The sequential hypothesis of the function of sleep. *Behavioural brain research*, *69*, 157–166.

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

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

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

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

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, & Cognition*, *37*, 801–812.

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

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 269–299.

Huber, R., Ghilardi, M. F., Massimini, M., & Tononi, G. (2004). Local sleep and learning. *Nature*, *430*, 78–81.

Hupbach, A., Gomez, R., Hardt, O., & Nadel, L. (2007). Reconsolidation of episodic memories: a subtle reminder triggers integration of new information. *Learning & Memory*, *14*, 47–53.

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

Jenkins, J. G., & Dallenbach, K. M. (1924). Obliviscence during sleep and waking. *The American Journal of Psychology*, *25*, 605–612.

Johns, M. W. (1991). A new method for measuring daytime sleepiness: The epworth sleepiness scale. *Sleep*, *14*, 540–545.

Kahana, M. J., & Greene, R. L. (1993). Effects of spacing on memory for homogeneous lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 159–162.

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

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

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

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. *The psychology of learning and motivation*, *61*, 237–284.

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

Killgore, W. D. (2010). Effects of sleep deprivation on cognition. *Progress in brain research*, *185*, 105–129.

Kliegl, O., & Bäuml, K.-H. T. (2016). Retrieval practice can insulate items against intralist interference: Evidence from the list-length effect, output interference, and retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Knutson, K. L., Spiegel, K., Penev, P., & Van Cauter, E. (2007). The metabolic consequences of sleep deprivation. *Sleep medicine reviews*, *11*, 163–178.

Kolers, P. A. (1973). Remembering operations. *Memory & Cognition*, *1*, 347–355.

Korman, M., Doyon, J., Doljansky, J., Carrier, J., Dagan, Y., & Karni, A. (2007). Daytime sleep condenses the time course of motor memory consolidation. *Nature Neuroscience*, *10*, 1206–1213.

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

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.

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

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, *1*, 279–308.

Lahl, O., Wispel, C., Willigens, B., & Pietrowsky, R. (2008). An ultra short episode of sleep is sufficient to promote declarative memory performance. *Journal of Sleep Research*, *17*, 3–10.

Landauer, T. K., & Eldridge, L. (1967). Effect of tests without feedback and presentation - Test interval in paired-associate learning. *Journal of Experimental Psychology*, *75*, 290–298.

Lange, T., Dimitrov, S., & Born, J. (2010). Effects of sleep and circadian rhythm on the human immune system. *Annals of the New York Academy of Sciences*, *1193*, 48–59.

Larsen, D. P., Butler, A. C., & Roediger, H. L., III. (2008). Test-enhanced learning in medical education. *Medical education*, *42*, 959–966.

Lehman, M., & Malmberg, K. J. (2013). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review*, *120*, 155–189.

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and

elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 1787–1794.

Mamelak, M. (1997). Neurodegeneration, sleep, and cerebral energy metabolism: a testable hypothesis. *Journal of geriatric psychiatry and neurology*, *10*, 29–32.

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, *14*, 194–199.

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., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, *16*, 192–201.

McDaniel, M. A., Kowitz, M. D., & Dunay, P. K. (1989). Altering memory through recall: The effects of cue-guided retrieval processing. *Memory & Cognition*, *17*, 423–434.

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

McGaugh, J. L. (2000). Memory - a century of consolidation. *Science*, *287*, 248–251.

McGaugh, J. L., & Gold, P. E. (1976). Modulation of memory by electrical stimulation of the brain. *Neural mechanisms of learning and memory*, 549–560.

McGeoch, J. A., & McDonald, W. T. (1931). Meaningful relation and retroactive inhibition. *The American Journal of Psychology*, 579–588.

McKeon, S., Pace-Schott, E. F., & Spencer, R. M. C. (2012). Interaction of sleep and emotional content on the production of false memories. *PLOS One*, *7*, e49353.

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.

Müller, G. E., & Pilzecker, A. (1900). *Experimentelle Beiträge zur Lehre vom Gedächtniss* (Vol. 1). JA Barth.

Nader, K., & Einarsson, E. Ö. (2010). Memory reconsolidation: an update. *Annals of the New York Academy of Sciences*, *1191*, 27–41.

Nader, K., & Hardt, O. (2009). A single standard for memory: the case for reconsolidation. *Nature Reviews Neuroscience*, *10*, 224–234.

Oswald, W. D., & Roth, E. (1987). *Der Zahlen-Verbindungs-Test (ZVT)* [connect-the-numbers test]. Göttingen: Hogrefe.

Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K.-H. T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 287–297.

Payne, J. D., Schacter, D. L., Propper, R. E., Huang, L.-W., Wamsley, E., Tucker, M. A., ... Stickgold, R. (2009). The role of sleep in false memory formation. *Neurobiology of Learning and Memory*, *92*, 327–334.

Payne, J. D., Stickgold, R., Swanberg, K., & Kensinger, E. A. (2008). Sleep preferentially enhances memory for emotional components of scenes. *Psychological Science*, *19*, 781–788.

Peigneux, P., Laureys, S., Fuchs, S., Collette, F., Perrin, F., Reggers, J., ... Luxen, A. (2004). Are spatial memories strengthened in the human hippocampus during slow wave sleep? *Neuron*, *44*, 535–545.

Plihal, W., & Born, J. (1997). Effects of early and late nocturnal sleep on declarative and procedural memory. *Journal of Cognitive Neuroscience*, *9*, 534–547.

Postman, L., & Keppel, G. (1977). Conditions of cumulative proactive inhibition. *Journal of Experimental Psychology: General*, *106*, 376–403.

Potts, R., & Shanks, D. R. (2012). Can testing immunize against interference? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 1780–1785.

Putnam, A. L., & Roediger, H. L., III. (2013). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition*, *41*, 36–48.

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.

Pyc, M. A., & Rawson, K. A. (2012). Why is test-restudy practice beneficial for memory? an evaluation of the mediator shift hypothesis. *Journal of Memory and Language*, *38*, 737–746.

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

Rasch, B., Büchel, C., Gais, S., & Born, J. (2007). Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science*, *315*, 1426–1429.

Rasch, B., & Born, J. (2013). About sleep's role in memory. *Physiological reviews*, *93*, 681–766.

Rechtschaffen, A., & Bergmann, B. M. (1995). Sleep deprivation in the rat by the disk-over-water method. *Behavioural brain research*, *69*, 55–63.

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

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

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

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

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*, 803–814.

Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation*, *55*, 1–36.

Rudoy, J. D., Voss, J. L., Westerberg, C. E., & Paller, K. A. (2009). Strengthening individual memories by reactivating them during sleep. *Science*, *326*, 1079–1079.

Sara, S. J. (2000). Retrieval and reconsolidation: toward a neurobiology of remembering. *Learning & Memory*, *7*, 73–84.

Schabus, M., Gruber, G., Parapatics, S., Sauter, C., Klosch, G., Anderer, P., ... Zeitlhofer, J. (2004). Sleep spindles and their significance for declarative memory consolidation. *Sleep*, *27*, 1479–1485.

Scheithe, K., & Bäuml, K.-H. (1995). Deutschsprachige Normen für Vertreter von 48 Kategorien [German-language norms for representatives of 48 categories]. *Sprache & Kognition*, *14*, 39–43.

Scullin, M. K., & McDaniel, M. A. (2010). Remembering to execute a goal: Sleep on it! *Psychological Science*, *21*, 1028–1035.

Siapas, A. G., & Wilson, M. A. (1998). Coordinated interactions between hippocampal ripples and cortical spindles during slow-wave sleep. *Neuron*, *21*, 1123–1128.

Siegel, J. M. (2005). Clues to the functions of mammalian sleep. *Nature*, *437*, 1264–1271.

Skaggs, E. B. (1925). Further studies in retroactive inhibition. *Psychology Monograph*, 1–60.

Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory and Language*, *26*, 589–607.

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

Smith, S. M. (1985). Background music and context-dependent memory. *The American Journal of Psychology*, 591–603.

Soraci, S. A., Carlin, M. T., Chechile, R. A., Franks, J. J., Wills, T., & Watanabe, T. (1999). Encoding variability and cuing in generative processing. *Journal of Memory and Language*, *41*, 541–559.

Soraci, S. A., Franks, J. J., Bransford, J. D., Chechile, R. A., Belli, R. F., Carr, M., & Carlin, M. (1994). Incongruous item generation effects: A multiple-cue perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 67–78.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*(9), 641–656.

Squire, L. R., & Alvarez, P. (1995). Retrograde amnesia and memory consolidation: a neurobiological perspective. *Current Opinion in Neurobiology*, *5*, 169–177.

Stickgold, R., Whidbee, D., Schirmer, B., Patel, V., & Hobson, J. A. (2000). Visual discrimination task improvement: A multi-step process occurring during sleep. *Journal of Cognitive Neuroscience*, *12*, 246–254.

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

Takashima, A., Petersson, K. M., Rutters, F., Tendolkar, I., Jensen, O., Zwarts, M. J., ... Fernandez, G. (2006). Declarative memory consolidation in humans: a prospective functional magnetic resonance imaging study. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 756–761.

Talamini, L. M., Nieuwenhuis, I. L., Takashima, A., & Jensen, O. (2008). Sleep directly following learning benefits consolidation of spatial associative memory. *Learning & Memory*, *15*, 233–237.

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.

Tilley, A. J., & Empson, J. A. C. (1978). Rem sleep and memory consolidation. *Biological psychology*, *6*, 293–300.

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

Tucker, M. A., & Fishbein, W. (2008). Enhancement of declarative memory performance following a daytime nap is contingent on strength of initial task acquisition. *Sleep*, *31*, 197–203.

Tucker, M. A., Hirota, Y., Wamsley, E. J., Lau, H., Chaklader, A., & Fishbein, W. (2006). A daytime nap containing solely non-rem sleep enhances declarative but not procedural memory. neurobiology of learning and memory. *Neurobiology of learning and memory*, *86*, 241–247.

Tucker, M. A., Tang, S. X., Uzoh, A., A.and Morgan, & Stickgold, R. (2011). To sleep, to strive, or both: how best to optimize memory. *PloS one*, *6*, e21737.

Underwood, B. J. (1948). Retroactive and proactive inhibition after five and forty-eight hours. *Journal of Experimental Psychology*, *38*, 29–38.

Underwood, B. J. (1957). Interference and forgetting. *Psychological review*, *64*, 49–60.

Van Dongen, E. V., Thielen, J.-W., Takashima, A., Barth, M., & Fernandez, G. (2012). Sleep supports selective retention of associative memories based on relevance for future utilization. *PLoS ONE*, *7*, e43426.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the norms. *Journal of Memory and Language*, *50*, 289–335.

Vertes, R. P. (2004). Memory consolidation in sleep; dream or reality. *Neuron*, *44*, 135–148.

Wagner, U., Hallschmid, M., Rasch, B., & Born, J. (2006). Brief sleep after learning keeps emotional memories alive for years. *Biological psychiatry*, *60*, 788–790.

Watkins, M. J. (1979). Engrams as cuegrams and forgetting as cue overload: A cueing approach to the structure of memory. *Memory organization and structure*, 347–372.

Watkins, M. J., & Watkins, O. C. (1976). Cue-overload theory and the method of interpolated attributes. *Bulletin of the Psychonomic Society*, *7*, 289–291.

Weinstein, Y., McDermott, K. B., & Szpunar, K. K. (2011). Testing protects against proactive interference in face–name learning. *Psychonomic bulletin & review*, *18*, 518–523.

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

Wheeler, M. A., & Roediger, H. L. (1992). Disparate effects of repeated testing: Reconciling ballard's [1913] and bartlett's [1932] results. *Psychological Science*, *3*, 240–245.

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

Wilhelm, I., Diekelmann, S., Molzow, I., Ayoub, A., Mölle, M., & Born, J. (2011). Sleep selectively enhances memory expected to be of future relevance. *Journal of Neuroscience*, *31*, 1563–1569.

Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, *265*, 676–679.

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, *55*, 235–269.

Wixted, J. T. (2005). A theory about why we forget what we once knew. *Current Directions in Psychological Science*, *14*, 6–9.

Wixted, J. T., & Rohrer, D. (1993). Proactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 1024–1039.

Wixted, J. T., & Rohrer, D. (1994). Analyzing the dynamics of free recall: An integrative review of the empirical literature. *Psychonomic Bulletin & Review*, *1*, 89–106.