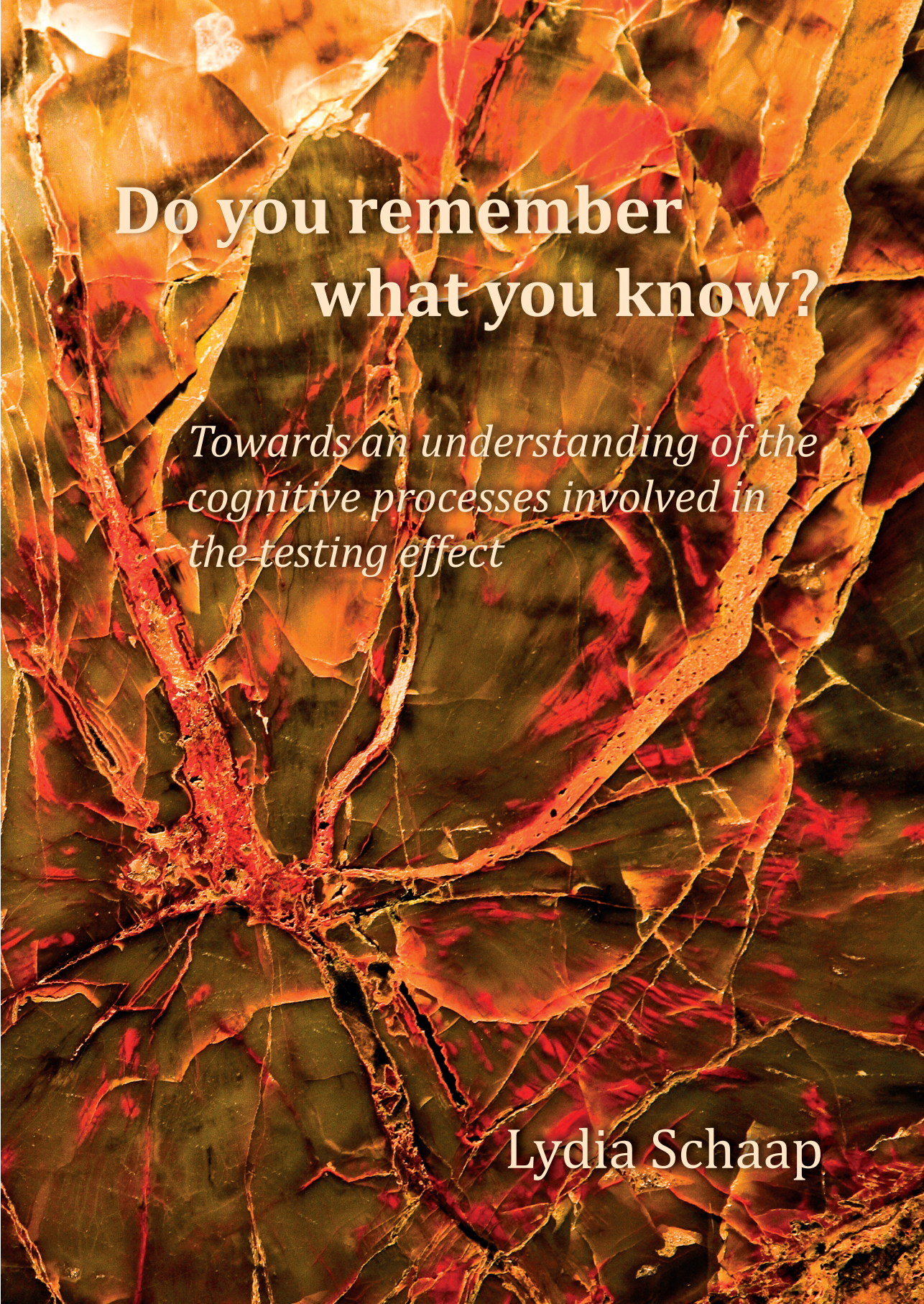

Do you remember what you know?

Towards an understanding of the cognitive processes involved in the testing effect

Lydia Schaap

Do you remember what you know?

Towards an understanding of the cognitive processes involved in the testing effect

Lydia Schaap

ISBN: 978-94-6169-383-9

Cover design: Frank van Erp
Layout: Optima Grafische Communicatie, Rotterdam

Printed by Optima Grafische Communicatie, Rotterdam

DO YOU REMEMBER WHAT YOU KNOW?

Towards an Understanding of the Cognitive Processes Involved in the Testing Effect

Proefschrift

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

De openbare verdediging zal plaatsvinden op
vrijdag 7 juni 2013 om 13.30 uur
door

Lydia Schaap
geboren te Leersum

PROMOTIECOMMISSIE

Promotor	Prof.dr. H.G. Schmidt
Overige leden	Prof.dr. R.M.J.P. Rikers Prof.dr. L. Kester Dr. K. Dijkstra
Co-promotor	Dr. P.P.J.L. Verkoeijen

CONTENTS

1 The testing effect in memory: An introduction

INTRODUCTION

There is an old saying that you cannot fatten a hog by weighing it, which means that the simple act of weighing a pig every day will not increase its weight. This saying is sometimes employed by opponents of the increase of the use of tests in educational practice, because simply testing students on their knowledge will not make them any smarter. Although this is probably true, using tests to assess students' knowledge level seems inevitable in educational practice and is not a bad thing per se. It can be used to indicate where a student stands against peers or a fixed standard after a learning phase, but it can also be used during a learning phase to guide student learning with help from feedback obtained by the results of a test.

One of the propositions belonging to this dissertation therefore is: You cán fatten a pig by weighing it! This proposition is not stated to claim that students could become smarter by testing them frequently, but that students can benefit from taking tests. In particular, one insight from cognitive psychology strongly suggests that testing students on their knowledge can strengthen their memory for that knowledge.

This insight is called the testing effect and is named after the empirical finding that testing students' memory after an initial learning phase will improve performance on a subsequent memory test. The effect holds even when compared to restudying the information and is most often found after a multiday retention interval (e.g., Roediger & Butler, 2011). The testing effect has not been widely adopted by educational practice, yet there are initiatives that, sometimes unintentionally, implemented an assessment method that could improve long-term retention of memory as a result of testing. One example from higher education is the so called progress test. The progress test is an assessment method that comprises administering tests that cover all topics of a (part of a) curriculum several times an academic year (e.g., Van der Vleuten, Verwijnen, & Wijnen, 1996). In other words, students are re-tested on the learning material of their curriculum periodically and need therefore to recall this information several times a year. Based on the research on the testing effect, progress testing will probably improve long-term retention of knowledge of the curriculum, albeit progress testing is often implemented for other reasons than the results of this research on the testing effect.

The present dissertation is concerned with how to improve long-term retention of knowledge, and focuses mainly on the testing effect, the cognitive processes involved in explaining this effect, and to a smaller extent on the possible applications of this effect in educational practice. In this first chapter, an overview of research on the testing effect as well as the different explanations of this effect will be described first, followed by a description of the possible applications of this effect in educational practice. Next some elementary

issues on human memory and information processing will be outlined. This introductory chapter will conclude with the research questions addressed in this dissertation, and an outline of the studies it contains.

The testing effect

The beneficial effects of taking a test while studying were already recognized centuries ago by Francis Bacon in 1620 and William James in 1890 (e.g., Roediger & Karpicke, 2006b), but they mainly emphasized the feedback function that tests can have and how tests can drive student learning (for example, by restudying those items which one failed to recall in a test). Another reason for why testing benefits learning, which is the main focus of this dissertation, is that the process of retrieval itself is beneficial for memory. At the beginning of the 20th century, research on the testing effect commenced. For example, Gates (1917) described the beneficial effects of "reciting" (i.e. retrieving information from memory) nonsense and sense material after studying this material as compared to re-reading the material. He studied this with 4th, 6th, and 8th graders and adults and generally found for both sense as well as nonsense material that reciting resulted in memory benefits when compared to rereading the material. Since then, many other researchers have investigated the testing effect.

One of the largest testing effect studies that has been conducted in educational practice is the study by Spitzer (1939). Over 3600 sixth grade children of schools in nine Iowa cities (USA) participated in this study. Children of the nine schools were arbitrarily divided into ten groups. Spitzer used two factual texts (A and B) of approximately 600 words as study materials. For both texts a twenty-five item multiple choice test was constructed to assess retention of the facts from the studied text. Group one to eight read text A and took test A, and then read text B. Group one and two took test B immediately after reading text B. Group three to eight took another, unrelated test immediately after reading text B, which had nothing to do with the content of text B: these groups were given test B at different retention intervals after studying text B (1 to 63 days). For some groups test B was repeated after several time intervals (1 to 63 days). Participants in group nine completed test B immediately after reading text B and also repeated test B after completing the first time that they took test B to see the effect of repeating the test.

Finally, the participants of group ten did not read text B, but did take test B, to assess previous knowledge on the subject matter of test B. An overview of the procedure can be seen in Table 1.

The results of this study by Spitzer (1939) showed that the test scores on test A were equal for all ten groups. This indicated that the groups were comparable in the amount of information they learned from reading text A and it was therefore presumed that they would also equally learn from reading text B. The scores at the different retention intervals

Table 1
Procedure of groups taking test B1, B2, and B3 in the Spitzer (1939) study

Time in days	0	1	7	14	21	28	63
Groups							
1	B1	B2	..	..	B3		
2	B1	..	B2	..	..	..	B3
3	..	B1	..	B2			
4	..	..	B1	..	B2		
5	..	..	..	B1	..	B2	
6	..	..	..	..	B1	..	B2
7	..	..	..	..	..	B1	
8	..	..	..	..	..	..	B1
9	B1; B2	..	..	..	..	..	..
10	B1*	..	..	..	..	..	..

**without reading text B*

were therefore compared as if the scores were obtained by comparable groups in different conditions.

The results showed that at day one the score on test B2 of group one was significantly higher than the score on test B1 of group three. Group three did not have an intervening test at day 0, while group one did. Similar findings were obtained for other B1-B2 comparisons. This indicated that taking an intervening test directly after studying text B, increased memory performance compared to a situation without testing after reading text B.

Other researchers also compared the impact of immediate intervening tests with the influence of no immediate test on final recall. For example, Darley and Murdock (1971) showed that an immediate free recall test after a learning phase, improved memory performance on a final test as compared to no free recall test after the initial learning phase. Also McDaniel, Kowitz, and Dunay (1989) compared the impact of three different immediate cued-recall tests with no cued-recall tests on performance on a final cued-recall test. They found beneficial effects for immediate recall as compared to no immediate recall. Although the results were impressive, critics pointed out that these results could be interpreted by the simple fact that testing ones memory was another restudy opportunity and was therefore beneficial to memory. This was later called the 'amount of processing hypothesis' (e.g., Carrier & Pashler, 1992).

Nowadays, testing studies often have a 'standard design' to rule out this 'amount of processing' confound. They typically start with an initial learning phase, which is either followed by restudy or by testing. These different conditions are often indicated with the abbreviation S for (re)study and T for test. For example STTT is used to indicate a study phase followed by three testing sessions or, or SSSS is used to indicate a study phase followed by

three restudy sessions. Differences in performance between conditions are then measured with a final retention test after a certain retention interval.

Tulving (1967) investigated the learning rates of three learning conditions: STST, STTT, and SSST and showed that the three conditions resulted in about the same form of the learning curve. This was a notable result because participants in the SSST condition could study the word list three times, while the participants in the STTT condition could study the word list only one time. Tulving (1967) showed with this study that testing was at least as beneficial as studying. He did not administer a final recall test after a multiday retention interval.

In the early 1970's Hogan and Kintsch (1971) were one of the first to show with a final test after a multiday retention interval, that retrieval can improve memory performance. They conducted two experiments. In experiment one, two conditions were compared: a restudy condition that ended with either a recall (SSST_recall) or a recognition test (SSST_recognition) and a test condition, consisting of one study trial followed by either three recall tests (STTT_recall) or three recognition tests (STTT_recognition). Performance of both groups was compared on a final recognition test and on a free recall test that were administered 48 hours after the learning phase. At the final free recall test, no testing effect was found. Both the restudy condition and the testing condition with free recall intervening tests had comparable scores on the final free recall test and the same applied to the final recognition test. However, when the restudy condition was compared to the testing condition with recognition as intervening test, a testing effect was found when the final test was a free recall test, but not when the final test was a recognition test, on which scores were comparable. Note that, the restudy condition in experiment one always ended with either a free recall or a recognition test. This might have influenced final test performance. Therefore Hogan and Kintsch (1971) conducted a second experiment, in which they compared a group of participants having four study trials (SSSS) with a group of participants having one study trial followed by three testing trials (STTT) on a final recall and recognition test or solely on a final recognition test, after 48 hours. Scores on the final recognition test (which was not preceded by a free recall test) were better for the SSSS condition, while a testing effect, that is better performance for the STTT condition, was found on the final free recall scores. To sum up, in the Hogan and Kintsch (1971) study, a testing effect was found when the intervening tests as well as the final test were a free recall test. No testing effect was found when the intervening tests and the final test were recognition tests.

During the years a series of replications and variations of testing effect studies have been conducted. For instance, Carrier and Pashler (1992) investigated the testing effect with paired associates and compared a pure restudy condition (SSSSS) with a combined study-test condition (STTST) in four experiments. Their results demonstrated that testing is more beneficial to memory than restudying the material with cued-recall

tests and after various retention intervals (varying from 2 minutes to 27 hours). Next to free recall and cued-recall, the testing effect has also been found with recognition tests (e.g., Runquist, 1983). In sum, the testing effect has been established with different types of tests under different learning conditions (for an overview see Roediger & Butler, 2011; Roediger & Karpicke, 2006b).

After the publication of a study by Roediger and Karpicke (2006a) the testing effect has been studied more intensively than ever. Many factors were varied to investigate the circumstances that would or would not result in a testing effect. For instance, one of the conditions that have been varied is the time lag between practice trials. Studies by for example Cull (2000) and Karpicke and Roediger (2007) showed that the presentation of test trials with a certain time lag (i.e. spaced) was more beneficial for memory performance than test trials that were presented directly after each other (i.e. massed). In addition, the study materials have also been varied to show that the testing effect is not only applicable to word lists. Research showed that next to simple facts (e.g., Carpenter, Pashler, Wixted, & Vul, 2008), word lists (e.g., Wheeler, Ewers, & Buonanno, 2003), and paired associates (e.g., Karpicke & Roediger, 2008) the benefits of testing can also be found with materials that are more likely to be used in educational practice. For instance, Roediger and Karpicke (2006a) investigated the effect with prose passages, and found that prior testing resulted in memory benefits as compared to restudying the prose passages at a final test after two days and after one week. Also, McDaniel and Fisher (1991) established a testing effect with educationally relevant material, namely general knowledge facts (comparable to facts from the *Trivial Pursuit* game). They compared a testing condition (with and without feedback) to a restudy condition and found that tested knowledge facts were better retained in memory than the restudied knowledge facts after a retention interval of 24-48 hours.

In addition to studies with more educationally relevant material, several studies have recently made a successful attempt to transfer the testing effect to the classroom (e.g., McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; McDaniel, Anderson, Derbish, & Morisette, 2007; Roediger, Agarwal, McDaniel, & McDermott, 2011). Because of these studies we now know for example, that quizzing (relative to non-quizzing) in 8th grade science classes was beneficial for students' scores on their summative unit examinations (McDaniel et al., 2011). Quizzing also improved performance on 6th grade social studies chapter and semester exams even when compared to a rereading control condition (Roediger et al., 2011). Finally, the testing effect was also established with symbol-word pairs (Coppens, Verkoeijen, & Rikers, 2011), videotaped lectures (Butler & Roediger, 2007), visuo-spatial materials such as maps (Carpenter & Pashler, 2007), and multimedia materials such as animations (Johnson & Mayer, 2009).

Although the large number of studies on the testing effect has made clear that the testing effect is a robust effect, researchers were also interested in explaining the underlying mechanisms of the effect. These mechanisms are still being debated nowadays.

Mechanisms underlying the testing effect

Two main views have been put forward in the literature as possible explaining theories of the testing effect: the Transfer Appropriate Processing (TAP) view (e.g., Morris, Bransford & Franks, 1977) and the Elaborative Processing (EP) view (e.g., Glover, 1989). The TAP view explains the testing effect by the match between the processes that take place during the initial test and the final test (e.g., Roediger & Karpicke, 2006b). According to this view, the cognitive processes required at an initial test and at a final test show much more overlap than the processes required for restudying and a final test. Because of these similarities in processes, the memory performance on a final test is expected to be better after an initial test than after restudying. To test this hypothesis, Carpenter and DeLosh (2006, experiment 1) varied the degree of similarity between initial tests and final tests. For example, either both the initial test and the final test were free recall tests (high degree of similarity) or the initial test was a recognition test and the final test was a free recall test (low degree of similarity). The results of this experiment did not support the TAP view, because more similar testing formats did not result in better performance on the final test than dissimilar testing formats.

Another, recently more popular theory for explaining the testing effect, is the elaborative processing (EP) view. This view focuses on the processes that occur during the initial test as compared to restudy. According to the EP view learners are more actively engaged in elaborative reprocessing the learning material during retrieval practice than during restudying this material (e.g., Anderson, 1983; Carpenter, 2009). Specifically, during an initial test, learners are assumed to be more actively engaged in elaborately reprocessing of the material than during restudy of the material. At an initial test, the act of retrieval, but not, or to a much lesser extent restudying, will result in the activation of information that is linked to the tested material, resulting in multiple retrieval routes to this material. Retrieving information will therefore result in more elaborative memory traces. Hence, memory performance will be better after testing than after restudying information (e.g., Carpenter, 2009).

Glover (1989) implicitly contrasted the TAP and EP view (experiment 4). Glover argued that the amount of elaborative processing needed, differs between recognition tests, cued-recall tests, and free recall tests. Recognition tests require less elaborative processing than cued-recall tests, which in turn require less elaboration than free recall tests. If the EP view is correct, the memory performance on any final test should be highest after an initial free

recall test, followed by a cued-recall test and the lowest performance would be expected after an initial recognition test. On the other hand, if the TAP view is correct, memory performance on the final test should be best if the overlap between initial test and final test is large (see, Carpenter & DeLosh, 2006). Glover (1989) combined these two predictions in one experiment. He varied the initial tests as well as the final tests (recognition, cued-recall and free recall), comparing combinations of the three types of initial tests with the three types of final tests. He found that taking a free recall initial test resulted in the best memory performance on the final test, compared to taking an initial cued-recall or recognition test, irrespective of the format of the final test. The second best performance was after an initial cued-recall test and the worst memory performance on a final test (irrespective of format) was after an initial recognition test. These results seem to clearly support the EP view, but according to Carpenter and DeLosh (2006) there were some points to improve in terms of experimental design (e.g., to equate time on task for all conditions).

Therefore, Carpenter and DeLosh (2006, experiment 2 and 3) also investigated the EP view and varied the amount of elaboration at the initial cued-recall tests in terms of the number of cues given during these tests. Participants had to study lists of words, followed by an initial cued-recall test consisting of different numbers of letters as cues for the to be recalled word. For example if they had studied the word doctor, the cued-recall test item could be d _ _ _ _ _, d o _ _ _ _, or d o c _ _ _. Fewer cues were assumed to require more elaboration to recall the studied word. The final test in this study was a free recall test after a distraction task of five minutes. Carpenter and DeLosh (2006) found support for the EP hypothesis with these experiments, because the target words that were tested with fewer cues at the initial test were better remembered at the final test.

Carpenter (2009) also found confirmation for the EP view of the testing effect. She conducted two experiments in which participants encoded weakly or strongly related word pairs. The encoding of weakly related word pairs was thought to require more elaboration than strongly related word pairs, because the link between the weakly related word pairs needs to be contrived by the participants themselves. After the encoding phase, participants had to either restudy the word pairs or were tested with a cued-recall test. Retrieving the target from a weakly related word pair was also considered more elaborative than retrieving a strongly related word pair, because the target and the cue are less obviously connected. Because of the more elaborative retrieval, weakly related word pairs were expected to be better remembered at a final test. The results of this study indeed showed that targets from weakly related word pairs were better retained than strongly related word pairs at a final test (five minutes retention interval) and therefore confirmed the EP view of the testing effect.

Another study investigating this view is from Pyc and Rawson (2010). They tested the so-called mediator effectiveness hypothesis. This hypothesis states that memory is enhanced

by testing because of the use of mediators during encoding. A mediator was defined as a word, phrase or concept that links the cue to the target (of for example a word pair). Participants studied Swahili-English word pairs (initial study) followed by three blocks of practice trials. A practice trial consisted of either restudying the word pairs or taking a cued-recall test followed by restudy. During the initial study and the restudy trials, participants had to think of a mediator that could facilitate their learning. After one week a final cued-recall test was administered. At this final test the target was cued in three different ways. Either only the original cue was given (C-group), or the cue plus the generated mediator was given (CM-group). Or, as a third way of cueing the target, only the cue was given with the instruction that participants had to recall the previously generated mediator first, before recalling the target word (CMR-group).

Pyc and Rawson (2010) predicted that mediators that were generated during a test plus restudy condition would be more likely to improve memory performance than mediators that were generated during a restudy only condition. The underlying rationale for this is that the mediator generated during testing will increase the strength of the link between cue and target. Because of the retrieval of the mediator during testing, the chance of future retrieval of the mediator is also increased and will therefore aid retrieval of the target from memory after the subject is being presented with the cue. In this study, a general testing effect was found at the final cued-recall test in all three cueing conditions (C, CM and CMR group). At the final cued-recall test in the C-group, memory performance was almost three times better in the test-restudy condition than in the restudy condition. In the CM-group a general testing effect was also found. However, when the final test scores of the C-group and the CM-group were compared, memory performance only increased for the restudy condition between these two groups, indicating that providing the mediators was only beneficial in the restudy condition and seemed to be superfluous in the test-restudy condition. In other words, providing mediators (CM-group) resulted in a smaller testing effect (smaller difference between the test-restudy condition and the restudy condition) than in the C-group. When the final test performance in the CMR-group was examined, the testing effect was of a similar size as in the C-group. However, recall of mediators was greater in the test-restudy condition than in the restudy condition (51% versus 34%). For the mediators that were actually retrieved, memory performance (indicated by the percentage correct recalled targets after correctly recalling the mediator) was best for the test-restudy condition, but the restudy condition also benefited from the mediator retrieval.

Pyc and Rawson (2010) concluded from these results that in the test-restudy condition more successful mediators were generated than in the plain restudy condition. If a mediator was retrieved at the final test, retrieving the accompanying target was more successful in the test-restudy condition than in the restudy condition. This was explained by the assumption

that successful retrieval of mediators during tests may strengthen the memory paths between the cue and the mediator, and between the mediator and the target. Additionally, they suggested that unsuccessful retrieval of mediators during learning may incite participants to think of another more successful mediator, which in turn increases memory performance in the long run. This assumption was supported by the finding that participants changed their mediators more often in the test-restudy condition than in the restudy condition (in 25% versus 19% of the trials). Pyc and Rawson (2010) concluded from their study that both mediator retrieval and decoding contributes to the beneficial effects of mediators in testing. Thinking of a mediator, being able to retrieve it and to decode it (retrieve the target with help from the mediator) is far more elaborative than restudying word pairs.

Yet another explanation of the testing effect is the retrieval effort hypothesis (Pyc & Rawson, 2009). This hypothesis states that the amount of retrieval effort at an initial test is positively related to memory performance at a final retention test and this fits within the desirable difficulties framework (e.g., Bjork, 1994; Pyc & Rawson, 2009). In their first experiment, Pyc and Rawson (2009) confirmed this hypothesis. Participants had to study Swahili-English word pairs, followed by practice comprising cued-recall tests and restudy trials. If a target was correctly recalled at a cued-recall test, it was dropped from further study. If a target was incorrectly recalled, the word pair had to be studied again, until it was correctly recalled. Furthermore, the number of times a target should be correctly recalled (i.e., the criterion level) varied between 1, 3, 5, 6, 7, 8, or 10 times. As soon as participants reached the assigned criterion level for an item, the item was dropped from further practice. This manipulation was assumed to reflect a variation of retrieval difficulty, because successive retrieval of a target becomes easier when it has been retrieved before. Moreover, this facilitation increases with the number of successful retrievals. So, the last retrieval of nine successful earlier attempts should require less effort than the last retrieval of five successful earlier attempts. Next to that, the inter stimulus interval (ISI) was manipulated, with longer intervals between study and test also (presumably) reflecting more retrieval effort. A final cued-recall test was administered either after 25 minutes or after one week. The results of this experiment indicated that retention on both final tests (after 25 minutes and one week) was enhanced most by effortful and successful retrieval. This means that final test performance was better after a long ISI than after a short ISI, at least when the target could be retrieved. Next to that, the positive influence of testing decreased as a function of the number of times a target was correctly retrieved. Pyc and Rawson (2009) therefore showed that the harder successful retrieval is, the stronger the testing effect is.

Thus far it seems that the elaborative processing hypothesis of the testing effect is quite plausible and more or less the same can be said for the retrieval effort hypothesis, because it is likely to presume that effortful retrieval will also include elaborative processing (e.g.,

Roediger & Butler, 2011). However, recent results from some studies are opposing this view (e.g., Karpicke & Smith, 2012). One of the predictions that arise from the EP view is that if elaboration is the explanatory factor in the testing effect, an elaborative restudy condition should result in a comparable memory performance as testing does. Karpicke and Smith (2012) tested this prediction by conducting two testing experiments in which an elaborative restudy condition was added to a standard testing design. Participants had to study and test themselves on word pairs until they could correctly recall the target with the cue. After a target was recalled correctly for the first time, three possible conditions could follow. Either nothing happened and the item was dropped from further practice, or the item was either elaborately restudied, or retested. Participants used either an imagery-based keyword method or a verbal elaboration method to elaborate on the learning material during a restudy trial. A final cued-recall test was administered after a one week retention interval. From this study it appeared that testing was still more beneficial than restudying, even when restudying was elaborate. Karpicke and Smith (2012) therefore concluded that the EP view cannot explain the testing effect.

To sum up, the testing effect has been studied extensively and seems to be very promising for educational practice. However, despite several studies in this direction, the applications of the testing effect in educational practice have not been fully investigated. Moreover, despite the large number of studies on this topic, the cognitive mechanisms underlying the testing effect are still not completely clear. The present dissertation aims to contribute to the research on these two issues. To understand the cognitive mechanisms underlying the testing effect, it is important to describe some elementary information on human memory first, before continuing to the research questions and outline of the studies of this dissertation.

HUMAN MEMORY: THE INFORMATION PROCESSING SYSTEM

Encoding

Most contemporary theories of memory are based on the information processing view (e.g., Ashcraft & Radvansky, 2010). Information comes in, is processed (encoded), and stored. Once stored, information can also be retrieved from memory. In short, memory processes can be divided into encoding, storage and retrieval.

Encoding is a term used for the various processes that take place in the human brain to transform information from the environment into some kind of memory representation (e.g., Tulving & Thompson, 1973).

According to Atkinson and Shiffrin (1968) the memory system can be divided into three parts: sensory memory, short-term memory, and long-term memory. In sensory memory,

stimuli from the environment are converted into sensible information as a consequence of the attention we pay to these stimuli. For example, when one hears something, the sound enters sensory memory. When no further attention is paid to the sound, the process stops. When one continues to pay attention to the stimulus, it can be held in short-term memory for a very short period of time (i.e., seconds). According to Atkinson and Shiffrin (1968) rehearsal of the information in short-term memory will lead to storage of that information in long-term memory. This idea has been challenged, because in short-term memory information from long-term memory should also be accessed. Otherwise we would not be able to *interpret* the information in short-term memory. Hence, information from long-term memory is also processed to interpret the information in short-term memory. Baddeley and Hitch (1974) therefore introduced the concept of working memory to replace the concepts of short-term and sensory memory.

In working memory, pieces of information are held, but also organized, rehearsed, and integrated with existing information from long-term memory. For example, when one has to read and understand a sentence, the meaning of the whole sentence can only be derived if the meaning of the separate words and some knowledge about grammar is retrieved from long-term memory. As a consequence of working memory processes such as rehearsal information is stored in long-term memory. Different aspects, elements, features, or attributes of a stimulus are hence stored in a memory trace (e.g., Tulving & Thompson, 1973). The entire process of storing information in long-term memory is called encoding and the quality of the encoding process is considered to be very important to the strength of the memory trace and hence the long-term retention of information.

One of the main theories considering the quality of encoding is the Levels of Processing Theory of Craik and Lockhart (1972). They distinguished between shallow and deep levels of processing. Shallow processing refers to for example processing the physical aspects or sound of a word. Deep processing refers to the meaningful analysis of information, for example thinking about the meaning of a word by trying to come up with a synonym. Craik and Lockhart (1972) considered the level of processing of information as a continuum, with the stronger memory traces closer to the deep processing end of the continuum.

This view of Craik and Lockhart (1972) has been criticized over the years. Nairne (2002) for example claimed that the beneficial effect of deep processing is an artifact of retrieval conditions and hence can be explained by the match between the memory traces that are formed as a result of deep processing and the commonly used tests to assess retention. In contrast, tests to assess retention can be arranged to favor shallow processing, and in that case the beneficial effects of deep processing will be nullified. In fact, shallow processing can result in better memory performance than deep processing. For example, Morris et al. (1977, experiment one) let participants study sentences on a deep semantic level or

on a shallow (rhyme) level. A final recognition test was either a standard recognition or a rhyming recognition test. Results showed that when participants were given a standard recognition test, deep processing resulted in better performance than shallow processing. However when a rhyming recognition test was given, shallow processing resulted in better performance. Nairne (2002) therefore claimed that, although deep encoding is in general a beneficial processing strategy, the retrieval conditions determine which strategy is most favorable. He called this the diagnostic value of retrieval cues or the distinctiveness of the retrieval conditions. According to Nairne (2002) we should take this into account when making claims about the quality of encoding processes.

Storage in long-term memory

Whether shallow or deep, processing information will usually lead to storage in long-term memory. Generally speaking, long-term memory is divided into declarative memory and non-declarative memory (e.g., Eysenck, 2012). Declarative memory, also known as explicit memory, contains knowledge that can be declared or stated in words or symbols and can be subdivided into episodic memory, semantic memory, and autobiographic memory (e.g., Eysenck, 2012). Episodic memory is the memory for events, such as what one had for breakfast yesterday, or that one learned a list of French words in French class last Tuesday. Semantic memory is the memory for all kinds of factual information, such as the meaning of the French words mentioned above, or the names of the countries of the African continent.

Autobiographical memory is the memory for personal events that were of great importance in one's live, such as the memories of one's wedding, or a great holiday. It resembles episodic memory, but autobiographical memory is limited to the memories that are personally important.

Next to declarative memory there is non-declarative memory which is sometimes called implicit memory, because recollection of information from non-declarative memory does not involve conscious control. A form of non-declarative memory is procedural memory. For example, when we have learned how to drive a car, it is difficult to consciously remember and name the different steps of how to set a car in motion (put it in first gear, pull down the hand break, kick in the right pedal, etcetera) and drive it, but we can perfectly execute the procedure of driving a car. When we try to execute such automated procedures consciously they sometimes become more difficult and less fluent.

The present dissertation is concerned with declarative memory, more specifically with semantic memory. In order to use the information stored in long term memory, for instance when taking a test, it is necessary to retrieve that information.

Retrieval

Retrieval can be seen as some kind of backwards encoding, or as mimicry of the processes that were active during encoding. The neural pathways in the brain that are formed during encoding are 'revisited' or 'reactivated' (e.g., Roediger, Gallo, & Geraci, 2002). The strengths of those neural pathways determine the quality of the memory (i.e., how quickly it can be accessed and how complete the memory is). The neural pathways are a composition of linked elements. When a retrieval cue activates one of those elements, other linked elements are also reactivated and combined to reinstate the pattern that was encoded while forming the memory (e.g., Roediger et al., 2002).

When one speaks about retrieval, the earlier mentioned distinction between explicit and implicit memory is often used as a starting point to discriminate between different forms of retrieval. Explicit memory involves deliberate recall**,** whereas implicit memory is the result of an unconscious recall process, often provoked by some implicit cue (for example the smell of a flower that elicits the memory of a wonderful holiday). The present dissertation deals with explicit memory, for which two forms of retrieval are often distinguished: recognition and recall. The main difference between these two is the absence or presence of an external retrieval cue. Both recognition and recall are the result of a process of activating information based on a cue. Retrieval of the target depends on the associative strength between the cue and the target. Strength in turn is determined by the amount of overlap between the cue information and the target information (e.g., Raaijmakers & Shiffrin, 1992).

In recall, a cue leads to a sampling of possible targets with different relative strengths and the target with the strongest association with the cue is recalled. For instance, consider a situation in which an individual has to come up with the name of the village in Tuscany where s/he spent the summer holidays six years ago. S/he then generates cues, which in turn results in a sampling of targets. The cue target combination with the highest relative strength will result in recalling the name of the village. In recognition, the overlap between cue and target is 100% but there needs to be enough overall activation as a result of the cue. If there is enough overall activation, the accompanying memory representation of the target in memory is recognized (Raaijmakers & Shiffrin, 1992). For example, an individual has witnessed a robbery, is shown a photo of the possible thief, and has to decide whether s/he did see that person on the photo doing the robbery or not. When there is enough memory activation after seeing this photo, the individual will answer confirmatively. Because memory retrieval from explicit memory happens consciously, it can be accompanied by a form of memory awareness.

Memory awareness

According to Tulving (1985) the process of retrieval from episodic and semantic memory is associated with different forms of memory awareness. Imagine that you try to recall what the capital of the Netherlands is. You might have a specific recollection of when/where/from whom you learned this information. In this case, knowledge is retrieved from episodic memory and the accompanying awareness would be 'remembering' according to Tulving (1985). Alternatively, it may be possible that you simply know that 'Amsterdam' is the correct answer without having any idea when/where/from whom you learned this. Under this condition, knowledge is retrieved from semantic memory, and the accompanying awareness would be 'knowing'. The transition from remembering to knowing is often seen as an indication of knowledge schematization (e.g., Herbert & Burt, 2001) and knowledge schematization is in its turn seen as a beneficial condition for long-term retention of knowledge (e.g., Bath, 2004). Schematization is defined as a process of acquiring semantic representations of rules, concepts, and abstract stereotypes of the knowledge domain (e.g., Herbert, & Burt, 2001). The process starts with the forming of episodic knowledge representations which are rather concrete and bound to a specific learning situation. As the process develops, memory representations become more conceptual and generalised in nature due to repeated experiences with the to be learned information in various contexts (e.g., Bath, 2004; Conway, Gardiner, Perfect, Anderson, & Cohen, 1997). The repeated experiences with to be learned information can be for instance rereading the information, but especially in higher education, educators try to encourage students to elaborate on the to be learned information. It has been suggested that elaborative processing would also affect the memory awareness accompanying future memory recollections (Gardiner, 1988). In addition, learning experiences do not necessarily have to be study opportunities, but could for example also be retrieval opportunities (e.g., Karpicke & Blunt, 2011). Memory retrieval and its effect on knowledge schematization and long-term retention will be studied in this dissertation by using the aforementioned remember-know distinction of Tulving (1985). At a memory test, one tries to remember what s/he knows of the earlier learned information. If done successfully, long-term retention of that information can be strengthened. Therefore the title of this dissertation ('Do you remember what you know?) refers to the retrieval practice and to the remember-know distinction of Tulving (1985) which are both part of the current dissertation.

OVERVIEW OF THE DISSERTATION

As discussed in this introduction, memory performance is not only dependent on the encoding of information, but retrieval processes also play a role in how well information is retained in memory. Retrieval processes are thought to be more elaborate than restudy processes and are therefore thought to be more beneficial for memory performance.(e.g., Carpenter & DeLosh, 2006; Pyc & Rawson, 2010). However, a very recent study seems to contradict this explanation of the testing effect (Karpicke & Smith, 2012) and the reason why retrieval practice is more beneficial than restudying for future memory performance is still not completely clear.

The present dissertation is concerned with the following *research questions*: How can the testing effect be explained and how can we use retrieval practice to improve long-term retention and hence optimize its potential for educational practice? To answer these questions both the elaborative processing hypothesis (in studies 3 and 4) and the retrieval effort hypothesis (in study 5) are investigated. In addition, the present dissertation focuses on the potential role of memory schematization in explanations of the testing effect (in studies 2 and 3) and will look into the practical boundaries of the testing effect: do testing format (study 2) or test-taking strategies (study 5) make any difference in final test performance? Whereas studies 2 to 5 are all concerned with the retrieval aspect of long-term memory, study 1 focusses on encoding, because if information is not learned well initially, testing will have no beneficial effect, since retrieval will not be successful at an initial test to begin with.

The five studies of this dissertation contributed to answering the general research questions in the following manner.

Study 1 (Chapter 2) investigated what factors play a role in initial learning and hence long-term retention. We formulated a descriptive model to predict first-year university students' knowledge growth on the basis of level of initial learning, prior knowledge, class attendance, and study time.

Study 2 (Chapter 3) investigated the separate effects of two different testing formats, multiple choice (MC) and multiple choice justification (MC-justification) items, on memory awareness and on long-term retention. As a result of more elaborative processing, MC-justification items would have a stronger effect on the conceptual organization of knowledge and hence result in a more beneficial effect on knowledge schematization and long-term retention than MC-items.

Study 3 (Chapter 4) addressed the elaborative processing hypothesis of the testing effect, by including an elaborate restudy condition in a standard testing effect experiment. A second goal of this study was to investigate the role of recollection and familiarity in the testing effect using the remember-know procedure of Tulving (1985).

Study 4 (Chapter 5) also investigated the elaborative processing hypothesis of the testing effect by including an elaborate restudy condition in a standard testing effect experiment but differed from study 3 in that two different elaborate restudy conditions were compared to a testing condition. We hypothesized that participants need to be handed effective mediators to be able to elaborately study the word pairs. Therefore we compared two conditions were participants either received an elaborate mnemonic aid from the experimenter or had to come up with it by themselves.

Study 5 (Chapter 6) investigated two test-taking strategies (response generate versus immediate choice) within a MC-test format to test the retrieval effort hypothesis of the testing effect with direct possibilities for application in educational practice.

In Chapter 7, the main findings of this dissertation will be summarized and will be discussed in the light of theoretical explanations of the testing effect.

2 Assessing knowledge growth in a psychology curriculum: Which students improve most?*

* This chapter has been published in a modified version as: Schaap, L., Schmidt, H. G., & Verkoeijen, P. P. J. L. (2012). Assessing knowledge growth in a psychology curriculum: Which students improve most? *Assessment and Evaluation in Higher Education, 37*(7), 875-887. doi:10.1080/02602938.2011.581747

ABSTRACT

The purpose of this study was to gain insight into determinants of knowledge growth among first-year psychology students in a curriculum that uses the Progress Test (an assessment method for long-term retention of knowledge and knowledge growth) as its main assessment tool. To that end, the relation between level of initial learning, prior knowledge, class attendance, and individual study time, and Progress Test scores was analyzed. The data showed that level of initial learning was positively associated with prior knowledge, and class attendance. Further, level of initial learning was positively related to knowledge growth at the end of the first year of the curriculum. Students with higher levels of initial learning had a more extended knowledge base at the end of the first year of their curriculum than students with lower levels of initial learning. Prior knowledge, class attendance, and individual study time did not have a significant relation with knowledge growth.

INTRODUCTION

A central goal of any educational system is that students retain the information they learn during their study for future professional activity and that they expand their knowledge base during education and even afterwards. Next to that, a commonly accepted idea among teachers and in the literature is that assessment strongly influences student learning (e.g., Scouller & Prosser 1994). The first mentioned goal refers to knowledge growth: a mnemonic state in which the amount of retained and newly acquired information surpasses the amount of forgotten information. However, in educational practice students' knowledge growth is usually not monitored. Instead, knowledge is tested at the end of or during a particular course. After students pass the accompanying course test, the knowledge from this course will typically not be tested again. Consequently, students will possibly study in a way that will help them pass the test, but not in a way that will help them remember the knowledge for a long time. Therefore we often do not encourage students to study for long-term retention and we also do not know how long students retain their knowledge and to what extent their knowledge base is expanded during their study. It seems that knowledge growth is not a core topic in educational practice. From an extensive literature search it appears that it is not studied extensively in educational research either. With this study we want to take a first step in studying factors related to knowledge growth within the psychology domain.

The present study was conducted in a Dutch problem-based learning (PBL) bachelor program of psychology. When this program started, the goal was to emphasize the importance of long-term retention of knowledge and to choose an assessment system that was congruent with this goal. In the present PBL curriculum, the basic knowledge of central domains of psychology is covered in the first two bachelor years. Each of these two years comprises eight sequentially programmed, five-week courses that all end with a 'course test'. This course test is formative and gives students feedback on how well they have mastered the subject matter that was studied during the preceding five week course. Next to that, students take three summative Progress Tests (PTs) per year. More specifically, in each of the first two bachelor years, a PT is administered after the third course, the fifth course and the eighth course. A PT covers all (theoretical) topics from the first two years of the bachelor curriculum (i.e. 16 courses). The underlying idea for using this kind of assessment method is to avoid undesirable learning strategies as 'learning for the test' or 'cramming' (Van der Vleuten et al., 1996). When students are retested a few times a year on the learning material, students need to (re)study the material in a way that helps them to remember it for a longer period of time than only for the upcoming test. By assessing them after every course, they are also informed on the efficacy of their study behavior on a regular basis.

The PT has originally been developed in the context of medical education to assess knowledge growth (Van der Vleuten et al., 1996). Due to the specific scheduling of the PTs, a student's score on a PT reflects both initial learning, that is, knowledge of the preceding course, as well as long-term retention, that is, knowledge of the courses that were conducted prior to the preceding course. It should be noted that - ideally - the long-term retention component in the PT-score becomes increasingly important as students move through the bachelor program. To exemplify this, consider a student's score on the first PT in the first bachelor year and compare it with this student's score on the third PT in this year. A student's score on the first PT in the first year will be based on the initial learning of the third course and the long-term retention of the first and second course. By contrast, a student's score on the third PT of this year will be based on the initial learning of the eighth course in that year, and the long-term retention of the previous seven courses.

The notion that progress testing can measure knowledge growth has been empirically supported, albeit to a limited extent, in the medical domain. For example, Van Diest et al. (2004) demonstrated that medical students show a steady growth in knowledge during their pre-clinical years of studying. Students in this study showed a significant increase in percentage correct answers on subsequent progress tests during their study. A study by Verhoeven, Verwijnen, Scherpbier, and Van der Vleuten (2002) revealed similar results. Tan, Imbos, and Does (1994) compared medical students with different levels of knowledge growth and concluded that growth of knowledge in the first year of the curriculum has important predictive value towards the final level of knowledge at the end of the curriculum. Students with relatively large knowledge growth in their first year of college, tend to end-up with more knowledge of the basic curriculum than students with relatively small knowledge growth in their first year.

Given that the final level of knowledge a student obtains is related to knowledge growth in the first year, and considering that most programs in higher education aim at providing student with a strong final knowledge base, it is relevant for teachers to obtain insight into the factors that determine knowledge growth in the first year of the curriculum. To the best of our knowledge, no prior study has been directed at the identification of these factors. Therefore, in the present study we sought to fill this hiatus by identifying factors that correlate with knowledge growth. On the basis of relevant empirical evidence from the cognitive and educational psychology literature, we constructed a simple qualitative model, which contains a set of direct predictors of knowledge growth. Below, we will elaborate on these predictors and their relationship with knowledge growth.

As noted before, knowledge growth as measured by a PT taps on the sum of long- term retention of knowledge and initial learning. Thus, if we are to predict knowledge growth, our model should contain factors that affect long-term retention and factors that affect

initial learning. As will be shown in the next session, the existing literature suggests that long-term retention is positively related to level of initial learning. Furthermore, initial learning is known to be positively correlated with level of prior knowledge, class attendance, and (sometimes) study. Therefore, the model we will use to predict knowledge growth will contain the aforementioned predictors. Subsequently, we will provide a description of these predictors.

Long-term retention of knowledge

Long-term retention of knowledge learned in school has been studied empirically in four domains: foreign language acquisition, high school mathematics, cognitive psychology, and memory for novels studied in university art courses (e.g., Bahrick 1984; Conway, Cohen, & Stanhope, 1992). Most studies show a rapid decline in knowledge in the first few years after the knowledge is acquired and a stabilization of the retained knowledge from 6-25 years after acquisition. For example, Bahrick (1984) studied the retention of Spanish words learned in college over the course of fifty years. It appears that in the first 3-6 years after learning Spanish words, there is a sharp decline in retention. After that period losses stabilize and a substantial part of the knowledge (around 50% of the maximum score on a retention test on average) is retained until participants reach the age of 60. Around that age, the knowledge retained starts to decline again. A study of Bahrick and Hall (1991) showed a similar pattern of knowledge retention and loss. They studied the very long-term retention of algebra and geometry knowledge learned in high school with a retention period of fifty years. Their participants showed a rapid decline of algebra and geometry knowledge in the first 3-5 years after knowledge acquisition. After that period, knowledge retention stabilized.

Although most research suggests that very long-term retention of knowledge is better than usually expected, there is a substantial decline in knowledge in the first few years after it has been acquired (e.g., Conway, Cohen, & Stanhope, 1991; Semb, Ellis, & Araujo, 1993). From an educational perspective it is of importance to know how this rapid decline can be diminished, because knowledge growth depends on knowledge retention. The question, therefore, is: which factors facilitate long-term retention?

Level of initial learning

To investigate which factors facilitate long-term retention, Bahrick and Hall (1991), in their study, controlled for level of knowledge initially acquired after studying the materials for the first time (i.e. level of initial learning). They found that participants with the highest level of initial learning (at the beginning of the retention period) showed hardly any decline in knowledge, while participants with the lowest level of knowledge initially acquired showed

a great amount of knowledge loss during the retention period. Similar results were found by Conway et al. (1991). Semb et al. (1993) also investigated the correlation between the level of initial learning and long-term retention. After a retention period of four months, there was no difference in decline between students with higher versus lower levels of initial learning. But after 11 months, the decline was larger for students with lower levels of initial learning. These studies suggest that level of initial learning is an important determinant of long-term retention of knowledge. If initial learning is of a high level, it is plausible that knowledge growth will be more extensive than when the level of initial learning is relatively low. In the present study we will therefore investigate whether level of initial learning also has a positive effect on knowledge growth.

Prior knowledge

The influence of prior knowledge on student learning has abundantly been established (e.g., Shapiro, 2004; Thompson & Zamboanga, 2004). Prior knowledge differs from the level of initial learning by the fact that prior knowledge is the knowledge students possess prior to enrolling in a curriculum. Level of initial learning on the other hand is the direct result of student learning in a certain course. Possibly the best way to distinguish between prior knowledge and level of initial learning is to compare it with the terms that Alexander, Schallert, and Hare (1991) use: prior knowledge would then be "the sum what an individual knows" (Alexander et al., 1991, p. 333) and level of initial learning would best be compared to discipline knowledge, a "highly formal subset of domain knowledge; knowledge of an academic subject that is taught; a specialized field, or study, or particular branch of learning" (Alexander et al., 1991, p. 332). Prior knowledge seems to strongly determine the level of initial learning. For instance, Recht and Leslie (1988), showed that students with more prior knowledge of baseball were better at recalling and summarizing a text about a baseball game and at assorting passage sentences from the original text for level of importance (importance was defined by seven baseball experts). Also Bransford and Johnson (1972) showed that prior knowledge has a positive effect on comprehending and recalling prose passages. While there are several theories on how prior knowledge affects initial learning, they all endorse the idea of prior knowledge as a kind of cognitive structure that lays the foundation for new learning (Shapiro, 2004). However, the effect of prior knowledge on long-term retention is not known, and neither is it known whether prior knowledge also has an effect on knowledge growth. We will investigate prior knowledge as a possible determinant of knowledge growth, because of its effect on initial learning, which constitutes a part of knowledge growth.

Class attendance

Attending class is a factor that proved to have a positive association with initial learning in many studies (e.g., Gunn, 1993; Marburger, 2001; Romer, 1993; Van Blerkom, 1996). Because we expect initial learning to be a part of knowledge growth, we will include class attendance as another possible factor that influences knowledge growth. We are aware of the fact that class attendance can represent different things (for example conscientiousness or motivation), though it is very plausible that students who are more interested in the subject matter, who are more focused on obtaining high grades or motivated in some other way will be more likely to attend classes than students who are less interested in the subject matter or who are less focused on academic achievement. To consider class attendance as a pure indicator of motivation to learn (e.g., Pintrich, 1999; Romer, 1993; St. Clair, 1999; Van Berkel & Schmidt, 2000) is probably too strong a statement. However, in line with other research indicating a positive association between motivation, level of initial learning and academic success (e.g., Busato, Prins, Elshout, & Hamaker, 2000), we do expect students who attend more classes than others to show more knowledge growth.

Study time

A fourth factor that may influence knowledge growth is the amount of time students spent at their study. In some studies, study time was positively related to academic achievement (e.g., Schuman, Walsh, Olson, & Etheridge, 1985), whereas other studies revealed a negative effect or no effect on achievement (e.g., Plant, Ericsson, Hill, & Asberg, 2005). In educational practice it is rather common to attach importance to increasing students' individual study time, but research has still not resolved the issue of whether study time has a positive effect on study success. Also whether study time has an effect on knowledge growth is as yet unknown. For this reason we will investigate the relationship between study time and knowledge growth. Because of the contradictory results found in earlier research, we are not able to predict in which direction the possible influence of study time on knowledge growth will be.

To summarize, while knowledge growth is an important goal in educational practice, it has not been studied extensively in educational science. We do know that on average students show knowledge growth (Van Diest et al., 2004; Verhoeven et al., 2002) and we have some notion of what factors might influence long-term retention (Bahrick & Hall 1991; Conway et al., 1991, 1992; Semb et al., 1993) and level of initial learning (Recht & Leslie 1988; Thompson & Zamboanga 2004; Van Blerkom, 1996). Furthermore, knowledge growth in the first year is positively correlated with student knowledge at the end of a curriculum. Hence, we reasoned that it would be important to educators to know which factors are related to first-year knowledge growth. The purpose of the present study is to identify a

number of relevant factors. To this aim, we formulated a descriptive model to predict first year knowledge growth on the basis of level of initial learning, prior knowledge, class attendance, and study time.

METHOD

Participants

Participants were 224 (68 male and 156 female) Dutch students of the cohort enrolled in a problem-based learning (PBL) psychology curriculum in 2003. In the Netherlands, students can only enroll in a university bachelor curriculum if they have finished pre-university education (VWO in Dutch) or have finished at least one out of four years in higher vocational education. This is a requirement for all Dutch universities and makes the group of participants commensurable with other groups of Dutch first year (psychology) students. Of the 224 participants, 190 finished pre-university education and 34 went to an institute of higher vocational education before enrolling in the psychology bachelor curriculum at hand. In the Netherlands, psychology topics are not part of pre-university education programs. The majority of the participants did therefore not study any psychology topics in former education. The mean age of the participants was 19.52 years with a range of 17.08 to 26.92.

Educational context

As said before, the educational context of the study at hand is Problem-based Learning (PBL). PBL is an instructional approach that uses academically or professionally relevant 'problems' as a starting point for student learning. A problem usually consists of a realistic description of a phenomenon, event or for example a psychological case (Schmidt, 1993). Students meet twice a week in small groups of approximately ten. They first analyze the problem, generate possible explanatory hypotheses, build on one another's ideas, as well as identify key issues to be studied further. These activities, based on their prior knowledge, allow students to construct a shared initial explanatory theory or model explaining the problem-at-hand (Schmidt, 1993). After this period of teamwork, they disperse for a period of individual study to work on learning issues they have identified as a group. After three days they meet again and are expected to share and discuss their findings, as well as to refine their initial explanations based on what they have learned. Students then move on to analyze a new problem, or if new learning issues requiring further study are identified during this phase, the process described above would be repeated. During these meetings a tutor is present to guide students' learning in the problem analysis and reporting phases. The tutor's role is to facilitate the processes involved when students co-construct knowledge

through discussions and sharing of ideas (Hmelo-Silver & Barrows, 2006). Thus, PBL can be seen as a cyclical process consisting of three phases: initial problem analysis, self-directed individual learning, and a subsequent reporting phase (Barrows, 1988; Hmelo-Silver, 2004; Schmidt, 1993).

Instruments

Progress Tests (PTs) are administered three times in the first year of the bachelor program. Each PT consists of approximately 200 items covering the knowledge domain as a whole and reflecting the (final) objectives of a curriculum. For each administration, a new test is constructed. PT items are presented in a true/false format. This means that students have to judge propositions on their accuracy. An example of a true/false item in the category social psychology is: 'The results of the famous Darley and Latané (1968) experiment can be explained better by diffusion of responsibility than by pluralistic ignorance'. If students do not know the answer to a certain question, they can choose to answer with a question mark. To discourage guessing, students' scores on a PT are calculated by subtracting the number of incorrect answers from the number of correct answers. Questions that are answered with a question mark are not rewarded or penalized.

The PT is a test with proper construct validity and modest reliability (e.g., Blake et al., 1996). The reliability coefficients for the PTs used in this study as well as some measure of construct validity of the PT have been calculated. Cronbach's alpha for the first PT was $\alpha = .56$, for the second PT $\alpha = .67$ and for the third PT $\alpha = .76$. Because students have the opportunity to choose to answer questions with a question mark, the number of questions used to calculate reliability coefficients differ between the different PTs. This means that reliability coefficients are calculated on basis of the questions that are answered by the students with 'true' or 'false'. In line with other validation studies of the PT (e.g., Van Leeuwen,

Table 1
Number of participants (N) per PT, mean scores and standard deviations

Variable	N	M	SD
Progress Test 1	218	22.06	6.93
Progress Test 2	218	28.01**	9.93
Progress Test 3	212	37.07**	13.25

Note: ** = significant difference (P < 0.001) with preceding PT.

Pollemans, Mol, & Eekhof, 1995), construct validity was assessed by measuring growth. Mean scores of the three PTs were compared to each other. Mean performance increased

across the three PTs with the highest score associated with the third PT (*see Table 1*). This indicates mean growth of knowledge.

Knowledge growth was assessed by subtracting the score on the first PT from the score on the third PT for every student. The first PT was administered after 15 weeks of studying to make sure that the difference between prior knowledge and newly acquired knowledge would be clear. The third PT was administered at the last day of the academic year. In that way, the increase of knowledge after one year of study of every student is calculated. The reliability of such difference scores has been the topic of a thorough discussion. The reliability of a measure represents the ability of that measure to distinguish among people on a particular trait or true score, and differences between scores tend to be less reliable than the scores themselves (Lord, 1956). A low reliability of a measure reduces statistical power because the relationship with any other variable cannot be larger than the square root of this reliability. However, with respect to difference scores it has been demonstrated that the reliability is only problematically low when all individuals in a sample display nearly the same difference (Rogosa & Willet, 1983). In that case, the variation in difference scores attributed to 'true change' will be small. Rogosa, Brandt, and Zimowski (1982) showed that the variation in true difference scores is small when the correlation between the single constituent scores is high. In addition, Rogosa and Willet (1983) demonstrated that the decrease in reliability of difference scores due to an increase of the single-score correlation is smaller when the reliability of the single scores is high, that is a Cronbach's alpha $\geq .80$. In the present study, the reliability of the third PT was fairly high, whereas the reliability of the first PT was low to moderate. Furthermore, the correlation between the scores on the third and the first PT was moderate ($r = .60$). Hence, the power of the statistical analyses that involve the PT change score will be sufficient.

Level of initial learning was assessed with the formative course tests at the end of each course during the first year of the program. Formative tests are not rewarded with credits, but are used to give students feedback on their level of knowledge acquired at the end of the accompanying course. The mean grade (on a ten-point scale, with ten being the highest) the participants obtained during their first year of the curriculum at hand was 5.61 (standard deviation 1.35). This might seem rather low, but one should consider the fact that these course tests were formative, not rewarded with credits, and that a 5.5 is indicated as a satisfactory score.

Unlike US or UK bachelors, students in the curriculum of the present study spend most of their time studying core psychology topics. In every first year course, a different sub-domain of psychology is covered. Students start for example with a course on social psychology. Course tests reflect the learning objectives of the course and generally consist of a combination of a rather large number of multiple choice items, combined sometimes with some

essay questions or short answer questions. When essay or short answer questions are used to test knowledge, the questions are corrected on the basis of exemplary answers. None of the course tests consisted of solely essay questions or other 'open' test formats. Grades on course tests are expressed on a ten-point-scale with a 5.5 indicating a satisfactory score. For every student the mean grade on the eight formative first-year tests was calculated. The resultant mean grade was taken as a proxy for level of initial learning.

Study time was estimated by the students themselves. Directly after each course, students fill out a compulsory and anonymous course evaluation form in which they have to estimate the time spent on self-directed learning activities during the preceding 5-week period. In this evaluation form, there is also space for general comments of the students. Considering the content and number of remarks, one can assume that the participants did feel free to be honest in their evaluation of the courses. Research from Moust (1993) showed that for relative short periods of time (i.e. two months), students' estimates of their time spent at studying are a valid measure of the real time spent studying. For every student the mean estimate on the eight courses was calculated and used as a proxy for study time.

Class attendance was measured by the numbers of time students were absent from their tutorial groups as registered by their tutors. Every course consists of approximately nine meetings. Students are obliged to attend at least seven out of nine meetings per course to have the course registered. Nonetheless, students sometimes choose to attend fewer classes than are obliged. We added the number of meetings students missed during the first year to obtain an attendance-score.

Prior knowledge, finally, was measured directly after the students enrolled into the curriculum. They were required at that stage to take a training PT to get acquainted with the test procedure. The examination setting of this training PT is completely similar to the setting of a 'real' PT (including invigilators) to make sure students make this test as if it was a real test. The reliability coefficient for this PT was calculated as well. Cronbach's alpha for this training PT was $\alpha = .52$.

PROCEDURE

The data were routinely collected during the academic year 2003/2004. To investigate whether level of initial learning, prior knowledge, class attendance and study time predict knowledge growth, a regression analysis was conducted.

RESULTS

Regression model

We conducted a multiple regression analysis to determine the predictors of knowledge growth.

Predictors of knowledge growth

Table 2
Means and standard deviations for all variables in the regression analysis

Variable	N	M	SD
Knowledge growth	186	15.06	10.82
Level initial learning year 1	186	5.76	1.34
Absence year 1	186	3.50	3.94
Study time year 1	186	13.67	3.13
Prior knowledge	186	12.89	7.36

Table 3
Zero-order correlations of the variables in the regression analysis

	1	2	3	4	5
1. Knowledge growth	---	.547**	-.292**	.017	.134*
2. Level initial learning year 1		---	-.452**	.112	.250**
3. Absence year 1			---	-.099	-.234*
4. Study time year 1				---	-.118
5. Prior knowledge					---

*Note: * = significant at the 0.05-level, ** = significant at the 0.01-level*

The multiple regression analysis examined the effects of level of initial knowledge, prior knowledge, class attendance, and study time on knowledge growth. The means and standard deviations for all variables are displayed in Table 2 and zero-order correlations are reported in Table 3. Table 3 shows that knowledge growth was significantly correlated with level of initial knowledge, class attendance and prior knowledge. Level of initial learning was significantly correlated with class attendance and prior knowledge. Class attendance and prior knowledge were also significantly correlated. Study time did not significantly correlate with any of the other variables.

Assumptions for regression were checked and found tenable. There were no signs of multi-collinearity, as correlations between variables did not exceed (+/-) .80 and tolerance

coefficients ranged between .692 and .965. Tolerance coefficients lower than 0.2 indicate multi-collinearity. Errors were approximately normally distributed and independent as a Durbin-Watson value of 2.076 was obtained. According to Field (2005) this value should be between one and three to assume independent errors. The closer this value is to two, the more likely it is that the assumption of independent errors holds true.

Using the forced entry method, a significant model emerged, $F(4, 181) = 19.79$, $p < 0.01$, *Adjusted R square* = .289, *MSE* = 83.27 predicting knowledge growth. Knowledge growth was significantly predicted by level of initial learning (*bèta* = .524, $p < 0.001$). Students with a higher level of initial learning showed more knowledge growth during the first year of the bachelor curriculum than students with a lower level of initial learning. Prior knowledge (*bèta* = .005), class attendance (*bèta* = -.061), and study time (*bèta* = -.047) did not significantly predict knowledge growth.

DISCUSSION AND CONCLUSION

The aim of this study was to gain insight into determinants of knowledge growth of first year students within a psychology curriculum with a progress test as its main assessment instrument combined with formative course tests. To that end, the relationship between level of initial learning, prior knowledge, class attendance and individual study time, and knowledge growth was analyzed. The data showed that level of initial learning played an important role in predicting knowledge growth in the first year of the curriculum. Students with higher scores on formative course tests had a more extended knowledge base of psychology at the end of the first year of the curriculum than students with lower levels of initial learning. However, prior knowledge, class attendance and individual study time did not significantly predict knowledge growth.

The results of this study concerning level of initial learning are in line with previous research on long-term retention of knowledge. Bahrick and Hall (1991), Conway et al. (1991), and Semb et al. (1993) found better retention scores for students with higher levels of initial learning. Although we do not know how and if students prepare themselves for the course tests and PTs during the year, we do know that students with higher levels of initial learning have forgotten less (whether or not by relearning) and/or have acquired more knowledge at the end of the first year of the curriculum. Even though it might seem a rather obvious conclusion that gaining high grades on course tests often results in high grades on progress tests, it is of much interest for curricula using PTs or other assessment methods for retesting knowledge, but also when no tests for long-term retention are administered. It seems that students who do not obtain a high level of understanding the first time they study the

learning material, easily fall behind compared to students who do obtain a high level of understanding of the learning material the first time. These students show less gain or forget relatively more than students who start with a high level of initial learning and are not able to compensate this during the year with for example restudying the learning material. This can have implications for educators' decisions on assessment. Long-term retention and knowledge growth seem to be connected to a solid base of initial learning. It might therefore be beneficial to stimulate students to study on a regular basis in a meaningful way and retest them on their knowledge, to prevent them from cramming. The advice to study on a regular basis is in line with research on the spacing effect (e.g., Delaney, Verkoeijen, & Spirgel 2010; Dempster, 1988). The spacing effect refers to the finding that with the same amount of time spent on studying, spacing the learning episodes has a beneficial effect on learning over massed learning episodes. The advice to study in a meaningful way is in line with the fact that the study was conducted within a PBL-curriculum where students are encouraged to constructively process the learning materials (Hmelo-Silver & Barrows, 2006).

The results of this study concerning prior knowledge are not in line with other research. Where prior knowledge has a positive effect on learning in general (e.g., Thompson & Zamboanga, 2004), it did not predict the level of knowledge growth. Nevertheless, it did have a significant correlation with level of initial learning. We found the same pattern of results for the class attendance variable. Class attendance did not significantly predict knowledge growth, but it did significantly correlate with level of initial learning. The fact that prior knowledge and class attendance did significantly correlate with the level of initial learning could be explained by the educational context of this study. Students in problem-based learning schools are challenged to activate their prior knowledge while discussing problems and this will help students integrate new knowledge into their existing knowledge base (e.g., Schmidt, 1993). It is plausible that this will enhance the level of initial learning. Although it is unclear why prior knowledge and class attendance did not predict knowledge growth, it could be that the relation between prior knowledge and attendance on one hand and knowledge growth on the other hand, is mediated by level of initial learning. Future research is of course necessary to investigate this possibility.

Study time did not significantly predict knowledge growth, nor did it have a significant correlation with level of initial learning. This is in line with research from Kember, Jamieson, Pomfret, and Wong (1995). They investigated the relationship between learning approaches, study time and academic performance and concluded that there is no simple relationship between these three variables. Ineffective learning approaches often demand much study time and will probably result in lower academic performance, but an effective learning approach will not result in higher academic performance without investing a proper amount of time.

This study showed that level of initial learning is of predictive value for knowledge growth at the end of the first year of the psychology curriculum under study. Knowledge growth in the first year is important because it appears to be indicative for knowledge growth at the end of the curriculum (Tan et al., 1994). There are, however, still some issues unresolved. For example, we do not know how students prepare themselves for the course tests and PTs. We did not control for restudying. Students who show more knowledge growth than others, could for instance restudy the material more often than others (Driskell, Willis, & Copper, 1992). The study time measure used in this study indicated the amount of time students spent on the particular courses, during the courses. It did not assess the amount of time students spent on restudying study material (or summaries) from other courses. Furthermore, we do not know what explains the differences in level of initial learning. Prior knowledge and class attendance were positively associated with level of initial learning, but did not predict knowledge growth. Perhaps there are other factors, for instance type of learning strategy, which were not investigated in this study that could play a role. Future research will be necessary to address these questions.

The present study was conducted in an educational setting rather than in an experimental one. Results should therefore be interpreted with caution, also because correlations do not imply causality. In addition, the research was conducted in a PBL curriculum, so one should be careful with generalizing it to different learning environments. Nevertheless it was a first step in finding determinants of knowledge growth in a PBL psychology curriculum with an assessment instrument that focuses on long-term retention.

3 Effects of different testing formats on long-term retention and schematization of knowledge*

*A modified version of this chapter is submitted as: Schaap, L., Verkoeijen, P. P. J. L, & Schmidt. H. G. Effects of different testing formats on long-term retention and schematization of knowledge.

ABSTRACT

This study investigated effects of two different testing formats on memory awareness and long-term retention of knowledge. Participants took four subsequent knowledge tests on curriculum learning material that they studied at different retention intervals prior to the start of this study (i.e., prior to the first test). At the first and fourth (pre and post) test participants indicated which form of memory awareness (i.e., remember, know, familiar, and/or guess) accompanied their answer. On the two intermediate tests, testing format was manipulated: multiple choice (MC) or MC-justification, that is an MC question with the additional instruction to explain why the chosen alternative was correct. The results resembled earlier findings in that different forms of memory awareness could be distinguished. The study did not indicate (additional) knowledge schematization as a result of testing or testing format. However, independent of test format, the proportion of correct answers on the posttest was higher than on the pretest. This could indicate that the beneficial effects of testing can occur even when the learning episode was at a long retention interval prior to the first test.

INTRODUCTION

If someone asks you what the capital of France is, you most likely will immediately answer: Paris. Since the last two and a half decades, an important question in cognitive psychology has been whether different conscious states accompany such knowledge retrieval. Tulving (1985) suggested that the process of information retrieval from episodic and semantic memory is associated with different forms of memory awareness. With respect to the 'Paris' example, you might have a specific recollection of when/where/from whom you learned this information. In this case, knowledge is retrieved from episodic memory and the accompanying awareness would be 'remembering' according to Tulving (1985). Alternatively, it may be possible that you simply know that 'Paris' is the correct answer without having any awareness of the memory source. Under this condition, knowledge is retrieved from semantic memory, and the accompanying awareness would be 'knowing'.

The transition from remembering to knowing is often seen as an indication of knowledge schematization (e.g., Herbert & Burt, 2001) and knowledge schematization in its turn is seen as a beneficial condition for long-term retention of knowledge (e.g., Bath, 2004). The importance of knowledge schematization and long-term retention of knowledge in educational practice is rather obvious to most educators. To improve knowledge schematization and/or or long-term retention of knowledge, educators often focus on for example study methods or study strategies, whereas they use testing only to assess the effects of study. However, testing can also serve as a strong method to improve long-term retention of knowledge: taking tests after studying learning material can enhance performance on a final test to a greater extent than restudying this material, a phenomenon known as the testing effect (e.g., Karpicke & Roediger, 2007). In the present study we will investigate whether and how different formats of testing can enhance knowledge schematization and long-term retention of knowledge.

Memory awareness and knowledge schematization

As said before, Tulving (1985) proposed that the memory awareness that accompanies memory retrieval is an indication of different memory processes or memory sources, which are associated with knowledge schematization. Tulving's proposition has been investigated in many studies and the two forms of memory awareness (i.e., remembering and knowing) have often been found to be distinguishable (e.g., Gardiner & Java, 1993; Gardiner, Ramponi, & Richardson-Klavehn, 1998). However, while most researchers agree that a 'remember awareness' is always accompanied by recollection of episodic details, there is some disagreement on what a 'know awareness' represents exactly. A 'know awareness' has sometimes been shown to be associated with feelings of familiarity and sometimes with

feelings of just knowing (e.g., Gardiner, 2001). Because of the equivocal results concerning a 'know' response, Conway et al., (1997) made a distinction between remember, (just) know, familiar, or a pure guess as possible states of memory awareness accompanying responses. They acknowledged that there are different meanings of the word knowing. Knowing can be described as the feeling of familiarity someone has, caused by a feeling of a recent encounter with the information. However, there are no memories of details of this encounter that accompany this feeling of familiarity. For example, a student does remember learning something about Freud lately, but does not know exactly what it was that s/he learned or where s/he was when it was learned. However, the student knows for sure that there was an encounter with the information on Freud lately. Conway et al. (1997) called this 'familiarity'. Alternatively, one can know that something is just the case, but this feeling of just knowing is not accompanied by any feelings of recent encounters with the item. For example, a student just knows that Freud is the founder of psychoanalysis, but does not have the feeling that s/he had a recent encounter with this information, s/he just knows. Conway et al. (1997) called this latter form of knowing 'just know'.

Remember-to-know shift

Conway et al. (1997) used the above-mentioned forms of awareness to investigate the development of knowledge representations in student learning. Students of four lecture courses and of three research methods courses were asked to take a multiple choice (MC) examination at the end of their course and to indicate for every MC answer on the test which memory awareness accompanied their answer. Students of one lecture course were retested at a delayed interval of 25 weeks. Students' scores on the tests were divided into four categories (from lowest to highest performing students). Conway et al. (1997) found that on the immediate test the higher performing students on the lecture courses had more remember responses than lower performing students, while the lower performing students had more familiar and guess responses. At the delayed retest of the lecture course, these higher performing students had more know responses, while the lower performing students still had more familiar and guess responses. Conway et al. (1997) concluded that after knowledge acquisition, as learning proceeds, a shift in memory awareness takes place from remembering to just knowing. They called this the remember-to-know shift. In line with Tulving's (1985) ideas, this was interpreted as a shift in knowledge from episodic memory to semantic memory wherein knowledge becomes more schematized.

Knowledge schematization

Herbert and Burt (2001) tried to replicate the findings of Conway et al. (1997). They tested two groups of students with a multiple choice (MC) test. One group was tested at the end

of a lecture course and one group at the end of a research methods course. Both groups were retested at a delayed interval (of 24 in the lecture group and of 9 weeks in the research methods group). At both tests, students were asked to indicate which memory awareness accompanied their answer on the MC-test. Furthermore students were asked to make a confidence judgment for each MC-answer. They were also requested to answer a short-answer (SA) question to assess the degree of schematization of knowledge. Students' scores on the tests were (similar to the Conway et al. 1997 study) divided into four classes (from lowest to highest achieving students). Just as in the study of Conway et al. (1997) a remember-to-know shift between the two tests was found in the lecture condition for the high-achieving participants. In the research methods condition a remember-to-know shift was found for all four achievement classes.

Herbert and Burt (2001) related these findings to the SA questions they administered to assess the level of schematization of students' knowledge. Answering these SA questions comprised writing passages about certain concepts from the course. The answers on the SA questions were analyzed with Biggs' Taxonomy of Structure of Observed Learning Outcomes (SOLO; Biggs & Collis, 1982). By means of this taxonomy, knowledge can be classified into five levels of schematization (pre-structural, uni-structural, multi-structural, relational, and extended abstract). Along with the remember-to-know shift in the research methods condition, participants in this condition also showed a shift in schematization of knowledge from multi-structural to relational (this means from knowing several relevant aspects of the topic to knowing these aspects and also being able to integrate them in a meaningful way). Overall, in the lecture condition, students also improved their level of schematization, but the highest level of schematization achieved was multi-structural.

According to Herbert and Burt (2001, 2003) and Conway et al. (1997) the remember-to-know shift is an indication of knowledge schematization. These researchers defined schematization as the process of acquiring semantic representations of rules, concepts and abstract stereotypes of the knowledge domain. According to Conway et al. (1997) schematization in student learning is a process that starts with the forming of episodic knowledge representations which later develop into more conceptual, generalized semantic memory representations. They emphasize that it is not an all-or-none phenomenon. As learning progresses, the number of episodic representations declines while at the same time a conceptual organization of knowledge develops. This process is a result of repeated experiences with the to be learned information in various contexts (Bath, 2004).

Schematization of knowledge can take place after several different encounters (real or mental) with the to be learned information (Bath, 2004; Conway et al., 1997). Encounters with the learning material can comprise of restudy activities, but it is also possible to take tests as relearning opportunities.

Effects of testing on memory performance and knowledge schematization

Taking tests can enhance performance on a final test to a greater extent than restudying the material. This 'testing effect' has been demonstrated repeatedly in experimental settings (e.g., Roediger & Karpicke, 2006b) as well as in educational relevant settings (e.g., McDaniel et al., 2007; Roediger et al., 2011).

Just as taking tests can enhance memory retention, a study by Herbert and Burt (2003) showed that schematization of knowledge can be enhanced by the opportunity of reviewing the learning material with different test formats. In this study, three groups of students were tested in three sessions on their knowledge of a first year lecture-based psychology course. Students were tested with MC-tests and SA-tests. At the MC-tests students were asked to indicate for each answer which memory awareness (remember, familiar, just know, and guess) accompanied their answer. Answering the SA-test questions comprised writing passages about relevant concepts. For example students were asked "Would you please write a page or so on the language development of humans, and what you understand about the important principles involved in this topic." (Herbert & Burt, 2003, p. 91). The three test sessions took place directly, four weeks, and ten weeks after the course. For all three groups, the final test session consisted of an MC-test and an SA-test. Session one and two differed for the three groups. The first group of students received an MC-test at both session one and two (MC + MC condition). The second group of students received an MC-test at session one and the same MC-test with an SA-test at session two (MC + MC-SA condition). The last group of students received an MC-test with a SA-test at session one, and no test at session two (MC-SA + none condition). An overview of the three different conditions can be seen in Table 1. Herbert and Burt (2003) reasoned that these different reviewing opportunities would have different effects on knowledge schematization and on long-term retention of knowledge. They hypothesized that students in the MC + MC condition would make more remember responses at the final test than students in the MC + MC-SA condition. On the other hand, they thought that students in the MC + MC-SA condition would have more (just) know responses at the final test as a result of schematization. They supposed that the SA questions (at session two) required participants to think about interrelations between concepts and that this process would enhance schematization. In other words, Herbert and Burt (2003) distinguished between reviewing tests that emphasize the recognition of knowledge (MC questions) and reviewing tests that emphasize more elaborate recollection of the material (SA questions). They assumed that the transformation from episodic to semantic memory could be enhanced more by SA-questions than by MC-questions, because SA-questions would lead to a more enriched representation of the learning material, which in turn would lead to more knowledge schematization. Indeed, a study by Herbert (1999) showed that students who studied 'episodically rich' material showed a greater degree of

knowledge schematization compared to students who studied 'episodically poor' material. With 'episodically rich and poor' the author meant the number of episodic details in the learning material (e.g., an explanation of a statistical analysis wherein the experimental groups were described as 'group A with treatment X' and 'group B with treatment Z' or as 'a group of young children who received medication for treating their phobia' and 'a group of young children who received behavioral therapy for treating their phobia').

Indeed, in the Herbert and Burt (2003) study, participants in the MC + MC-SA condition

Table 1
Overview of Conditions of the Herbert and Burt (2003) Study (Number of Review Opportunities per Test Session in Parentheses)

Condition	Session 1	Session 2	Final Test	Number of review opportunities before final test
MC – MC	MC (1)	MC (1)	MC + SA	2 (1+1)
MC – MC + SA	MC (1)	MC + SA (2)	MC + SA	3 (1+2)
MC + SA – none	MC + SA (2)	- (0)	MC + SA	2 (2+0)

showed a remember-to-know shift and showed also a higher level of schematization (as tested with the SA test at the final session) than students in the MC + MC condition and students in the MC-SA + none condition. The students in the MC + MC condition and the MC-SA + none condition did not show a remember-to-know shift, and they attained a similar lower level of schematization compared to participants in the MC + MC-SA condition. Herbert and Burt concluded that when participants had the opportunity to review the learning material regularly and with different testing formats (MC and SA), this had a positive effect on the occurrence of a remember-to-know shift and on the level of schematization attained.

However, Herbert and Burt (2003) did not investigate the possible separate effects of the two different testing formats. For example, there was no condition with exclusively SA-tests. From an instructional perspective, this would be an interesting question to address, because teachers might want to know which testing format is more beneficial in promoting knowledge schematization and long-term retention. Next to that, Herbert and Burt (2003) did not control for time on task. It could be the case that students in the MC + MC-SA condition just had an extra "learning episode" compared to the other two conditions, because they took two tests in the second session (see Table 1). The remember-to-know shift and the higher level of knowledge schematization could therefore also be a consequence of this additional "learning episode".

Present study

Knowing what facilitates knowledge schematization is important, because knowledge that has been schematized is much more likely to be retained in long-term memory. When knowledge is not schematized, it will most likely be forgotten eventually (e.g., Bath, 2004). Different encounters with the learning material are thought to enhance knowledge schematization (e.g., Conway et al., 1997). These different encounters can occur in the form of restudy opportunities, but also in the form of retrieval practice which also has a beneficial effect on long-term retention of knowledge (e.g., Roediger & Karpicke, 2006b). According to Herbert and Burt (2003) different forms of retrieval practice can have different effects on memory schematization. However, in their study time on task was not equal for the different conditions and due to the fact that the different conditions contained combinations of test formats, the role of different test formats in isolation is unclear. In the present study we will therefore investigate the separate effects of two testing formats on knowledge schematization. We will not compare MC with SA questions, but compare MC questions with MC-justification questions instead. This will enable us to better monitor and compare the additional mental activities that the participants are asked to perform in MC-justification condition. MC-justification questions require first choosing the correct answer, and then justifying why the chosen answer is the correct answer according to the participant. This is thought to require the use of higher level thinking skills than answering regular MC questions (e.g., Fellenz, 2004). As an extension of the study of Herbert and Burt (2003), the effects of different test formats on long-term retention will be investigated as well.

As in other studies (e.g., Bath, 2004; Herbert & Burt, 2001, 2003), Tulving's (1985) distinction between remembering and knowing will be used to investigate the remember-to-know shift in memory awareness and will be seen as an indication of knowledge schematization.

We expect participants in the MC-justification condition to establish a larger remember-to-know shift, because of the more elaborative processing of the material during answering MC-justification items than during MC items. Retrieval practice in the MC-justification condition can be seen as a more elaborate relearning experience (e.g., Glover, 1989) which would have a stronger effect on the conceptual organization of knowledge than retrieval practice in the MC condition. We therefore also expect participants in the MC-justification condition to have a larger proportion of correct answers on a final recognition test than students in the MC condition.

METHOD

Participants and design

Participants were 26 Dutch psychology students (age $M = 20.8$ years, range: 18 to 30 years; 6 male) enrolled in a problem-based learning (PBL) psychology bachelor curriculum who participated for course credits. The original number of participants was 30, but four participants did not complete all four tests. Participants were randomly assigned to two conditions, an MC condition, and an MC-justification condition. The present study utilized therefore a 2x2 mixed factor repeated measures design. The independent variables were testing format (MC and MC-justification; between-subjects factor) and testing occasion (pretest and posttest; within-subjects factor). The dependent variables were the mean proportion of correct answers on the posttest and the relative accuracy of the different memory awareness categories on the posttest. How these variables were measured will be explained in the data analysis section of this paragraph. Participants in both conditions (MC versus MC-justification) were comparable in age, $t\,(24) = -.116$, $p = .908$, $r = .02$ and in their mean score on the course tests that were topic of interest in this study, $t\,(24) = -.260$, $p = .797$, $r = .05$.

Educational context

The educational context of the study is Problem-based learning (PBL). PBL is an instructional approach focusing on developing flexible knowledge, effective problem-solving skills, and self-directed learning skills (e.g., Loyens, Kirschner, & Paas, 2011).

Instruments

A knowledge test consisting of 45 true/false questions covering three courses of the first year of psychology was developed for the pretest and posttest, which were identical. In addition, two intermediate knowledge tests, each consisting of 45 different questions covering the same three courses of the first year of psychology were developed in two different formats (MC and MC-justification). In the MC condition participants received 45 true/false questions. In the MC-justification condition, participants received the same 45 true/false questions, with the difference that they had to explain in a few sentences why the statement in the question was true or false.

The three courses were social psychology, cognitive psychology, and developmental psychology. All participants took these courses prior to their participation in the present study, thus in advance of the pretest. All test questions were items that were made by a team of experts and that had been used before in other cohorts to test the knowledge acquired during these courses. The participants in the present study had not seen these questions before.

PROCEDURE

The procedure used was identical for both testing format conditions. A systematic overview of the procedure can be found in Table 2. There were four testing sessions over a period of nine weeks. The first testing session (pretest) was 1.5 weeks after the developmental psychology course ('short-term knowledge'), 6.5 weeks after the cognitive psychology course ('middle to long-term knowledge'), and almost 25 weeks after the social psychology course ('long-term knowledge'). The second testing session was three weeks after the first one, the third session four weeks later, and the fourth and last testing session (posttest) again two weeks later. Hence, the last testing session was 10.5 weeks after the developmental psychology course, 15.5 weeks after the cognitive psychology course, and almost 34 weeks after the end of the social psychology course. Every testing session lasted 30 minutes and took place in a lecture hall. Participants were told at the beginning of each session to take the test as serious as if it was a real exam and that they were not allowed to sit next to each other (at least one table should separate two participants), to talk with fellow participants, or to use books or other aids that could help answering the questions. Participants were asked not to leave any question unanswered and, if necessary, to guess if they did not know the answer. At each session, students received the appropriate test in a printed booklet. The booklet started with a standard instruction on how to answer the questions and (if applicable) a memory awareness instruction following a Dutch translation of the instruction used by Conway et al. (1997) and Herbert and Burt (2003) (see Appendix A). The score on each test was calculated by adding the total number of correct answers.

TABLE 2
Overview of the Lapse of Study Phase and Testing Occasions of the Present Study

Study Phase	Test 1 (pretest)	Test 2	Test 3	Test 4 (posttest)
	MC + memory awareness	*MC or MC-justification*	*MC or MC-justification*	*MC + memory awareness*
Social Psychology course	25 weeks after finishing the course	28 weeks after finishing the course	32 weeks after finishing the course	34 weeks after finishing the course
Cognitive Psychology course	6.5 weeks after finishing the course	9.5 weeks after finishing the course	13.5 weeks after finishing the course	15.5 weeks after finishing the course
Developmental Psychology course	1.5 weeks after finishing the course	4.5 weeks after finishing the course	8.5 weeks after finishing the course	10.5 weeks after finishing the course

Data analysis

As said before the dependent variables of the present study were the mean proportion of correct answers on the posttest and the relative accuracy of the different memory awareness categories on the posttest. Relative accuracy was defined as the probability of a correct response given a memory-awareness category. This was assessed by calculating the percentage remember, know, familiar, and guess responses that was correct as a proportion of the total number of answers (correct and incorrect) within each memory awareness category (for each participant). The mean relative contribution to accurate performance is the relative contribution of the different memory awareness categories to accurate performance. This is assessed by calculating the number of remember, know, familiar, and guess responses that were correct as a proportion of the total number correct responses (for each participant). These proportions add up to 100% per column.

RESULTS

Manipulation check

To test whether participants did elaborately explain their answers on the true/false questions in the MC-justification condition, the answers of the participants were analyzed. Participants' answers were divided into two categories. Answers that consisted of terms like "I just guessed, I don't know, it is just a fact, or I just remember it" were part of the first category (no elaboration). Answers that consisted of a more elaborate explanation than in the first category were part of the second category (elaborate answer). It appeared that 72.1% of the answers could be categorized as a second category answer, which indicates that on average participants gave an elaborate answer to 32.5 out of 45 questions. As an example of an elaborate answer, consider the following true/false question: "Excitation transfer can be better explained by James-Lange's theory of Emotions than by the Emotion theory of Schachter and Singer" with the justification: "The answer is false because excitation transfer can be explained by the cognitive evaluation of the arousal that takes place. In James-Lange theory cognitive evaluation does not play a role, while it does in the theory of Schachter and Singer".

Effects of testing conditions on long-term retention

The mean proportions of correct responses on the pretest and posttest are shown in Table 3. Table 4 shows the mean proportions of correct responses on the intermediate tests with different testing formats. To test the effects on long-term retention of the different testing formats, a two-way mixed ANOVA was carried out with testing occasion (pretest and

posttest) as within-subjects factor, testing format as between-subjects factor, and mean proportion correct answers as dependent variable. In this analysis, $p = .05$ was used as the threshold for statistical significance. This analysis showed a main effect of test occasion, $F(1, 24) = 6.206$, $p < 0.05$, *partial* $\eta^2 = .205$ indicating that the proportion of correct answers on the posttest was significantly larger than on the pretest.

The analysis did not show a main effect of testing condition, $F(1, 24) = 1.529$, $p > 0.05$, *partial* $\eta^2 = .060$, nor did the effect of occasion interact with testing condition, $F(1, 24) = .213$, $p > 0.05$, *partial* $\eta^2 = .009$. This indicated that both testing formats were equally beneficial for long-term retention of knowledge.

Table 3
Mean Proportions of Correct Responses as a Function of Test Occasion (Pretest and Posttest) and Testing Condition (MC and MC-Justification)

	MC	MC Justification	Total
Pretest	.76	.71	.74
Posttest	.80	.76	.78

Table 4
Mean Proportions of Correct Responses on the Intermediate Tests as a Function of Testing Condition (MC and MC-Justification)

	MC	MC Justification	Total
Intermediate Test 1	.78	.75	.76
Intermediate Test 2	.74	.68	.71

Effects on memory awareness

To test whether the mean accuracy of the different memory awareness categories differed from each other and whether this changed over time and/or as a consequence of testing condition a three-way mixed ANOVA was carried out. In this ANOVA memory awareness category and testing occasion were used as within-subjects factors, and testing format as between-subjects factor. Mean relative accuracy was the dependent variable. In this analysis, $p = .05$ was used as the threshold for statistical significance. Table 5 shows the mean accuracy as a function of testing condition across the two tests (pre and post).

If a remember-to-know shift would take place, an interaction effect should appear between memory awareness category and testing occasion. This ANOVA on the mean accuracy scores as dependent variable showed a main effect of memory awareness, $F(3, 72) = 27.656$, $p < 0.01$, *partial* $\eta^2 = .535$. Pairwise comparisons (Bonferroni) showed that overall the relative accuracy of remember and know responses was higher than the relative accuracy of the familiar and guess responses ($p < 0.01$). The remember responses and the know responses

did not significantly differ from each other ($p = 1.00$). The familiar and guess responses did also not differ significantly in relative accuracy ($p = 1.00$). This ANOVA did not show any other main or interaction effects.

To test whether a remember-to-know shift had taken place between the pretest and posttest and whether there were separate effects of the two testing conditions, another three-way mixed ANOVA was carried out. In this ANOVA memory awareness category and testing occasion were again used as within-subjects factors, and testing format as between-subjects factor. In this second ANOVA relative contribution to accurate performance was the dependent variable. Table 6 shows the mean relative contribution of the different memory awareness categories as a function of testing condition across the two tests.

The second ANOVA on the mean relative contribution to accurate performance as depen-

Table 5
Mean Accuracy of Remember, Know, Familiar, and Guess Responses as a Function of Test Occasion (Pre and Post), and Testing Condition (MC and MC-Justification)

	Pretest		Posttest	
	MC	MC Justification	MC	MC Justification
Remember	.84	.81	.88	.89
Know	.88	.82	.86	.82
Familiar	.62	.59	.68	.57
Guess	.57	.46	.61	.63

Table 6
Mean Relative Contributions of Correct Remember, Know, Familiar, and Guess Responses as a Function of Test Occasion (Pre and Post), and Testing Condition (MC and MC-Justification)

	Pretest		Posttest	
	MC	MC Justification	MC	MC Justification
Remember	.34	.37	.35	.33
Know	.29	.28	.27	.29
Familiar	.21	.23	.24	.23
Guess	.16	.12	.15	.15

dent variable also showed a main effect of memory awareness, F (1.864, 44.731) = 5.696, $p < 0.05$, *partial* $\eta^2 = .192$. Mauchley's test of Sphericity showed a violation for the factor memory awareness. Therefore, the more conservative Greenhouse-Geisser estimates were used. Pairwise comparisons (Bonferroni) showed that out of all correct responses, most were remember and know responses. Remember, know, and familiar responses did not

significantly differ in their contribution to the total number of correct responses, but the contribution of all three responses to correct responses was significantly larger ($p < 0.05$) than the contribution of guess responses. However, no interaction effect between memory awareness and testing occasion was found, $F(3, 72) < 1$, $p > .05$, *partial* $\eta^2 = .007$. Therefore, no indication of a remember-to-know shift was found. This ANOVA also did not show any other main or interaction effects

DISCUSSION AND CONCLUSION

The aim of this study was to investigate the separate effects of two different testing formats, MC and MC-justification questions on memory awareness and on long-term retention of knowledge. The shift from remembering to knowing is called the remember-to-know shift and has been found to be an indication of knowledge schematization (e.g., Herbert & Burt, 2004). It was expected that MC-justification items would have a more beneficial effect on knowledge schematization and long-term retention than MC-tests. Since knowledge that has been schematized is much more likely to be retained in long-term memory, it is important to know for educators whether there would be a differential effect from both test formats, so that they might incorporate more MC-justification items in educational practice if these would show advantageous for schematization.

The findings from this study did not reveal a comparable remember-to-know shift to that previously established by Conway et al. (1997) and Herbert and Burt (2001, 2003). In fact, it did not show a shift at all, because there was no interaction-effect between memory awareness and testing occasion for either test type. The relative contribution of remembering and knowing to the correct responses was not influenced by retrieval practice itself or by one of the two testing formats. It was expected that the contribution of remember responses to correct answers would become less prominent over the test occasions and that knowing would increase in relative contribution due to knowledge schematization. These results were not found. In addition, the relative accuracy of the different memory awareness categories remained stable over time and was not influenced by test type.

Although non-significant results should always be interpreted with caution, there might be several possible explanations of these results. First, they might be due to the fact that the knowledge had already been schematized. In other words, that the shift had already taken place, before students started participating in the experiment. Dewhurst, Conway, and Brandt (2009) tried to establish a remember-to-know shift in an experimental design. They also distinguished between remember, (just) know, and familiar, and guess responses. Participants studied lists of non-related words and were (re)tested several times with re-

tention intervals of five minutes, four weeks, eight weeks, and six months. In this study, evidence for the remember-to-know shift was found by comparing the relative contribution of the different awareness categories to correct responses after five minutes with the relative contribution after six months. The relative contribution of remember responses decreased, while the relative contribution of know responses increased (as did the relative contribution of familiar and guess responses). However, if the relative contribution to correct responses after four weeks would have been compared with those after six months, no remember-to-know shift might have been found in the Dewhurst et al. (2009) study. The relative contribution to correct responses of the different memory awareness categories at their retention test after four weeks, eight weeks, and six months were of comparable sizes as those reported in our study.

Also in a study of Dudukovic and Knowlton (2006) a remember-to-know shift became apparent between a retention test after ten minutes and one after a week. The relative contribution of remember and know responses to correct answers after one week were comparable to our findings too. This supports the idea that in our study, the knowledge had possibly already been schematized even after 1.5 weeks. This could be a consequence of the educational context, problem-based learning (PBL). In PBL students are encouraged to become active learners, to construct their own knowledge, and to discuss and elaborate the to be learned material (e.g., Loyens et al. 2011). This is comparable to the Conway et al. (1997) study where the rather large number of know responses among the correct responses directly after a research methods course also indicated that the knowledge already had been schematized. Conway et al. (1997) suggested that this could be explained by the fact that the research methods course led to more conceptual processing instead of episodic processing and therefore enhanced the number of know responses. On the other hand, it could also be the case that the remember awareness is only apparent a very short time after studying the material, irrespective of the educational context. Perhaps, even a retention interval of 1.5 weeks is too long to have many episodic memories of the encoding event. This is in line what Conway (2009) described. Participants had to list as many specific episodic memories for yesterday, two days before, three days before etc. It seemed that after a retention interval of three days memories became more general and schema-like instead of specifically episodic in nature.

Second, another possible explanation for the unexpected result of not finding a remember-to-know shift is the fact that earlier studies found this shift only for high performing students. Because of the relatively small number of participants in this study, we could not discriminate between different performance groups. We were able to compare the participants' scores on the original course tests of the three psychology courses under study with the total cohort scores on these course tests. The mean scores of the total cohort

were respectively 6.1, 5.9, and 6.4 (on a ten point scale, with ten being the highest) and the mean scores of the participants were respectively 6.3, 5.5, and 6.2. Three independent one sample t-test showed that these mean scores were not significantly different from the mean scores of the cohort the participants were part of, t (25) = .475, p = .639; t (24) = -1.017, p = .320; t (25) = -.669, p = .509. This indicated that the participants were representative of the cohort with respect to their level of knowledge of the different study topics. If the remember-to-know shift is only apparent in higher performing students, it is possible that the schematization is mediated by, for example, learning strategies, and meta-cognitive skills that help students conceptualize their knowledge. Perhaps studying by understanding the learning material or by using rote memorization to learn new information, will have differential effects on a possible remember-to-know shift. Future research could investigate this possibility.

Third, the fact that we did not find a remember-to-know shift might possibly be explained by the fact that participants encountered the intermediate tests as episodic learning events. Even though knowledge may have been more schematic at the posttest, the proportion remember awareness might have stayed high because participants were thinking of the retrieval practice tests and retrieved episodic details about those events in stead of the original learning event. This explanation can be related to the results of a study by McDermott (2006). She investigated the effect of number of retrieval practice tests on final test performance and also used the remember-know distinction of Tulving (1985). McDermott (2006) found that "the greater the number of tests intervening between the encoding and the final retrieval episodes, the higher the probability that people would claim to be able to remember the *initial study* episode" (McDermott, 2006, p. 264).

Finally it is possible that participants experienced difficulties in deciding which memory awareness accompanied their memory retrieval. Recent research from McCabe, Geraci, Boman, Sensenig, and Rhodes (2011) showed that remember responses are almost always associated with episodic details, but so are know responses to a certain extent. McCabe et al. (2011) recommend that it is of great importance that participants strictly follow the given instructions to indicate their memory awareness. Possibly, in the Dutch language, the connotations that are evoked by the terms representing the different forms of memory are different or less distinctive than in English. Perhaps participants did not quite capture the difference between remembering and knowing or between familiarity and guessing. They might have indicated 'remember' when retrieving a semantic concept and 'knowing' when retrieving a episodic concept. This is in line with the present finding that remembering and knowing as well as familiarity and guessing did not differ from each other in accuracy. Future research should try to incorporate some control variable in these kinds of experiments, to make sure that instructions are well understood and followed by the participants.

As for the long-term retention of knowledge, we did find a main effect of testing occasion for mean proportion of correct answers. These findings suggest that taking intermediate tests on previously acquired knowledge increased memory performance on a final test. A limitation of this study is that it did not include a no-intermediate test condition. Nevertheless, it seems fairly reasonable to suggest that the intermediate tests are responsible for the increase in performance, because without such tests, one would expect a decrease in accuracy over time as an effect of decay, especially when the rather large retention interval between pre- and posttest in the present study is taken into account. Still, one could wonder why memory performance increased instead of remained stable as a result of intermediate testing. Although there were no explicit signs of it, we cannot exclude the possibility that the participants in the present study prepared themselves for the intermediate or final tests. If this was the case in the present study, the increase in performance could be explained by the so called 'indirect effect' of testing: taking tests gives students feedback on their performance which in turn could guide their future learning (e.g., Roediger & Karpicke, 2006b).

In sum, this study showed that taking intermediate tests consisting of either MC or MC-justification items can have beneficial (indirect) effects on performance on a final MC test. This is interesting for educations in general, but especially for educators working with progress tests (see Schaap, Schmidt, & Verkoeijen, 2011, i.e., Chapter 2 of this dissertation) as it suggests that the long-term learning outcomes of students can be affected by testing practices.

Appendix A

Instruction on how to choose between the different memory awareness categories (based literally on the instructions of Conway et al. 1997, p. 397-398):

Please indicate for each answer what the memory awareness was that you had while answering the question. Indicate whether you had a Remember, Just know, Familiar or Guess awareness by encircling the most appropriate awareness. You will find a description of every awareness category below:

Remember: You remembered a specific episode/incident when you learned the specific item of information. In this case you might have images and feelings in mind relating to the recalled information. Perhaps, you virtually 'hear' or 'see' again the situation you were in when learning the item of information. Alternatively you might have a specific memory of reading or talking about the topic. Answers such as these are called Remember-answers.

(Just) Know: You might 'just know' the correct answer and the alternative you have selected 'stood out' from the two choices available. In this case you would not recall a specific episode and instead you would simply know the answer. Answers with this basis are called Know- answers.

Familiar: It may be, however, that you did not remember a specific instance, nor do you just know the answer. Nevertheless the alternative you have selected may seem or feel more familiar than the other alternative. Answers made on this basis are called Familiar-answers.

Guess: Finally, you may not have remembered, known, or felt the choice you selected to have been familiar. In this case you may have a made a guess, possibly an informed guess, e.g., you have selected the answer that looked least unlikely. This is called a Guess-answer.

4 | Investigating the processes underlying the testing effect: The role of elaborative processing, familiarity, and recollection*

*A modified version of this chapter is submitted as: Schaap, L., Verkoeijen, P. P. J. L, & Schmidt. H. G. Investigating the processes underlying the testing effect: The role of elaborative processing, familiarity, and recollection.

ABSTRACT

This study investigated the elaborative processing (EP) hypothesis of the testing effect (i.e., the finding that testing participants' memory after an initial learning phase improves their performance on a subsequent memory test even when compared to restudying). In line with the EP hypothesis, elaborately restudying, like testing, should result in better memory performance than plain restudying. Participants ($N = 34$) learned Swahili-Dutch word pairs, and took a final cued-recall test after a testing, elaborately restudying, or a restudy control condition. Although we did find a general testing effect, we did not find support for the EP hypothesis, because elaborately restudying did not result in better memory performance than the restudy control condition. Another goal of this study was to further elaborate on the role of recollection and familiarity in testing by using the remember-know procedure of Tulving (1985) after a longer retention interval than previously studied (e.g., Gardiner, 1988). Our results were in line with these earlier studies and additional results indicated that over time, in the restudy control condition, the relative contribution of familiarity based answers to correct answers diminished while they remained stable in the elaborate restudy and testing condition. The role of recollection processes remained stable over time in all three conditions.

INTRODUCTION

In the last decade, psychologists have gained renewed interest in the testing effect, that is, the empirical finding that testing participants' memory after an initial learning phase will improve their performance on a subsequent memory test (e.g., Glover, 1989). This effect holds even when compared to restudying and is most often (or sometimes only) found when the final test is administered after a multi-day retention interval (Roediger & Butler 2011; for a review, see Roediger & Karpicke, 2006b). Because many experiments on the testing effect will be discussed in this chapter, an overview of a standard testing effect design is presented in Figure 1.

Although the testing effect has been established with different types of tests and with

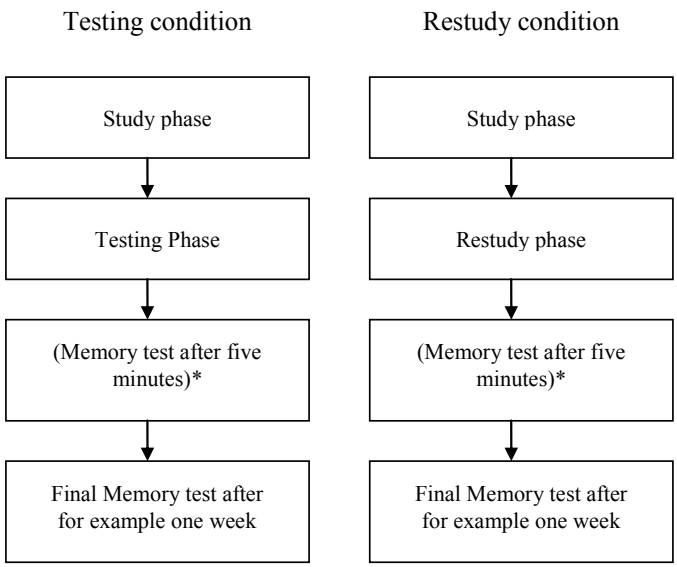

Figure 1
Schema of a standard testing experiment.

** This part is sometimes excluded from the experiment for (half of the) participants*

several different types of materials, such as word pairs (e.g., Karpicke & Roediger, 2008), word lists (e.g., Wheeler et al., 2003), facts (e.g., Carpenter et al., 2008) and prose passages (e.g., Roediger & Karpicke, 2006a) it is still not completely clear, which cognitive processes underlie the effect. One of the main hypotheses which has recently been put forward is the elaborative processing hypothesis (e.g., Carpenter, 2009). The present study will test this

hypothesis and by doing so, we will aim at contributing to the line of research on explaining the testing effect. Another goal of this study is to gain more insight in the relative contribution of familiarity and recollection processes to the performance on a final memory test (Chan & McDermott, 2007) by using the remember-know procedure (Tulving, 1985).

Explanations of the testing effect

Nowadays, two theoretical accounts of the testing effect are dominant in the literature: The transfer appropriate processing (TAP) view (e.g., Morris et al., 1977) and the elaborative processing (EP) view (e.g., Glover, 1989). The TAP view explains the benefit of testing in terms of the match between the processes that take place during the initial test and the final test (e.g., Roediger & Karpicke, 2006b). According to the TAP view, the cognitive processes required at the initial test and at the final test show much more overlap than the processes at restudying and at the final test. As a result the final test performance will be better after initial testing than after restudying.

The EP view on the other hand emphasizes the role of the cognitive processes that operate during the initial test. During an initial test, learners are assumed to be more actively engaged in elaborately reprocessing of the material than during restudy of the material. At an initial test, the act of retrieval, but not, or to a much lesser extent restudying, will result in the activation of information that is linked to the tested material, resulting in multiple retrieval routes to this material. Retrieving information will therefore result in more elaborate memory traces and thus in better retention on a final test than restudying information (e.g., Carpenter, 2009).

Carpenter and DeLosh (2006) investigated these two possible explanations of the testing effect in three experiments. The first two experiments included a restudy control condition. In experiment one, the similarity of testing format between the initial test and the final test was varied. For example, the initial test and final test were both free recall tests (high degree of similarity), or the initial test was a recognition test and the final test was a free recall test (low degree of similarity). In experiment two and three, the amount of elaboration evoked by testing was varied in terms of the number of cues given during the initial cued-recall tests (fewer cues were assumed to require more elaboration). With respect to these experiments, the TAP view predicts that memory performance will be best when the initial test and final test show a high degree of similarity compared to a low degree of similarity. By contrast, the EP view predicts that the mnemonic effect of testing will increase as a function of the elaboration required during the initial tests. For instance, an initial cued-recall test that gives only the first letter as a cue of a to-be-recalled word requires more elaboration, than a three-letter cued-recall test and will therefore benefit memory performance on a final free recall test more. The results of the experiments strongly supported the EP view and not the TAP view.

That is, similar testing formats did not result in a better final retention when compared to dissimilar testing formats. On the contrary, Carpenter and DeLosh (2006) found an increase in memory performance as a function of the number of cues. The final test performance was best in the condition with the fewest numbers of cues.

The EP view was investigated further by Carpenter (2009) who conducted two experiments in which participants encoded related word pairs of different semantic associative strengths (strong versus weak). Carpenter (2009) for example stated that the words 'toast' and 'bread' are semantically strongly related, whereas the words 'basket' and 'bread' are weakly related. After the encoding phase, participants were either tested with a cued-recall test or they were asked to restudy the word pairs. After a retention interval of five minutes, participants were given a free recall test for the target words (i.e., the second words from the word pairs). The EP view predicts that at an initial test, the retrieval of semantically strongly related word pairs is much easier and therefore requires less elaborate effort than that of weakly related word pairs. Due to more elaborate processing, targets from weakly related word pairs should be better retained at the final test (at five minutes delay) than targets from strongly associated word pairs. This was exactly what Carpenter (2009) found in both experiments, in addition to a general testing effect (the tested word pairs were overall better retained than the restudied word pairs). Carpenter (2009) explained her results in terms of the EP hypothesis, focusing on the elaborate part of retrieval. Another hypothesis, posed by Pyc and Rawson (2009), is that it is not so much the amount of elaboration, as the amount of effort associated with retrieval that is the key to explaining the testing effect. This retrieval effort hypothesis predicts that the more effortful retrieval is at an initial test, the better memory performance will be at a final test, provided that retrieval is successful at an initial test. They confirmed this hypothesis in their study (experiment 1). Participants had to study Swahili-English word pairs, followed by practice comprising cued-recall tests and restudy trials. If a target was correctly recalled at a cued-recall test, it was dropped from further study. If a target was incorrectly recalled, the word pair had to be studied again, until it was correctly recalled. Furthermore, the number of times a target should be correctly recalled varied between 1, 3, 5, 6, 7, 8, or 10 times. As soon as participants reached the assigned criterion level for an item, the item was dropped from further practice. This manipulation was assumed to reflect a variation of retrieval difficulty, because successive retrieval of a target becomes easier when it has been retrieved before. Moreover, this facilitation increases with the number of successful retrievals. So, the last retrieval of nine successful earlier attempts requires less effort than the last retrieval of five successful attempts. Moreover, inter stimulus interval (ISI) was manipulated, with longer intervals between study and test also reflecting more effortful retrieval. To investigate the effects of retrieval effort on the short and long term, a final cued-recall test was administered either after 25 minutes or after one

week. The results of this experiment indicated that retention on the final tests (for the short and long retention interval) was enhanced most by effortful and successful retrieval. This means that final test performance was better after a long ISI than after a short ISI. Next to that, the positive influence of testing decreased as a function of the number of times a target was correctly retrieved. Pyc and Rawson (2009) therefore showed that the harder successful retrieval is, the stronger the testing effect is.

Thus far it seems that the elaborative processing hypothesis of the testing effect is quite plausible. More or less the same can be said for the retrieval effort hypothesis; because it is likely to presume that effortful retrieval will also include elaborative processing (e.g., Roediger & Butler, 2011). In the standard testing experiments testing is indeed more elaborate than restudying. But what would happen if restudy would be made more elaborate? If the elaborative processing hypothesis is correct, a restudy condition that requires elaborate processing should benefit memory performance more than normal restudying, and perhaps even to a comparable degree as testing. The first purpose of this study was to test the elaborative processing hypothesis as explanation of the testing effect, by including an elaborate restudy condition in a standard testing effect experiment. The second goal of this study was to study the memory awareness processes associated with the testing effect, which will be discussed next.

Remember-Know procedure and retrieval processes in the testing effect

From the above mentioned studies it has become clear that the processes during retrieval play an important role in explaining the testing effect, but there are still some uncertainties about how testing improves memory. A different approach towards investigating the processes accompanying the testing effect is to use Jacoby's (1991) dual process framework, which states that recognition performance is determined by a conscious recollection process as well as a more automatic familiarity process. This is often investigated by using the remember-know procedure of Tulving (1985)[1]. Conscious recollection would be associated with a so called remember awareness and the more automatic familiarity responses by a know awareness.

According to the elaborative processing hypothesis of the testing effect, testing is more elaborate in nature than restudying information. It has been suggested that elaborate processing would also affect the memory awareness accompanying future memory recollections (Gardiner, 1988). Gardiner (1988) investigated the effects of encoding processes on the relationship between recognition memory and memory awareness. Participants in

1 In chapter 3 of the current dissertation an adaptation of this procedure was used to stay close to the most relevant research that addressed the issues of chapter three. In the present chapter, we will use the original procedure.

this study were asked to elaborate to various degrees (phonetic versus semantic processing and read versus generate) on a learning task. At a final recognition test after a one hour retention interval participants had to indicate whether they recognized words from the study phase and whether or not their recognition was accompanied by conscious recollection (remember response) or not (know response). When subjects reported a remember awareness accompanying their recollection on the recognition test, the preceding learning task was more likely to be elaborate (semantic and generate) in nature. When subjects reported a know awareness accompanying their recollections, no differences were found in the amount of elaboration in the preceding learning tasks. Thus, this study suggests that elaborate processing affects memory awareness accompanying recognition decisions on a later memory test. This was also demonstrated by Gardiner, Gawlik and Richardson-Klavehn (1994) who found that elaborate rehearsal of words in a study list resulted in more remember responses at a final recognition test; whereas maintenance rehearsal of words in a study list resulted in an increase of know responses at a final recognition test.

Hypotheses

Participants in the present experiment studied Swahili-Dutch word pairs and after studying these words, they were either tested with a cued-recall test or they restudied the word pairs in an elaborate way with help from a mnemonic aid, or simply restudied the word pairs (control condition). For example the mnemonic aid for the word pair *ardhi – grond* (*soil* in English) was: '*ardhi*, sounds like *aarde* (*earth* in English), which is another Dutch word for *grond* (*soil*)'. Subsequently, participants were retested once (immediately) or twice (immediately plus after a one week retention interval) with a cued-recall test. At every cued-recall test, participants were also asked to indicate their accompanying memory awareness to every answer.

After a one week retention interval, one would expect a better memory performance in the testing condition compared to the restudy control condition, because testing has been shown to have a beneficial effect on subsequent memory performance. However, according to the elaborative retrieval hypothesis, the elaborate restudy condition should also improve memory to a greater extent than the restudy control condition. The final test in the present study will not only be administered directly after studying/testing, but for about 65% of the participants also after a one week retention interval. Administering a cued-recall test directly and after a retention interval of one week, will give us the opportunity to exploratory investigate the memory awareness that accompanies memory retrieval directly and after a retention interval of one week, and possible changes in memory awareness between these two tests.

Following the elaborative retrieval hypothesis, we are interested in possible differences between the proportions of correct remember responses at the direct test in the three relearning conditions. It is expected that participants in the testing and elaborate restudy condition participants would elaborate more than in the plain restudy condition and would therefore show more remember responses. The opposite would be expected to be the case for the know responses, because in the restudy control condition, participants would probably rely more strongly on familiarity processes.

METHOD

Participants

Participants were 34 Dutch bachelor students enrolled in a Problem-based Learning (PBL) program who participated for course credits (age *M* = 20.61 years, range 18-31; 5 male).

Materials and procedure

Forty-eight Swahili – English word pairs were selected from the Nelson and Dunlosky (1994) study. These word pairs were subsequently translated into 48 Swahili – Dutch word pairs. For every word pair a mnemonic aid was developed. For example for the word pair: *ardhi – grond* (*soil* in English) was: '*ardhi*, sounds like *aarde* (*earth* in English), which is another Dutch word for *grond* (*soil*)'.In the first phase of the study, participants were instructed to study 48 word pairs. Each word pair was shown in the middle of a computer screen on a green, blue, red or yellow background for 5 seconds. The color of the background of the computer screen was varied to insert episodic details into this phase. We would then be able to measure at a later point in this study whether participants were able to recall episodic details of this phase. Participants were instructed to memorize the word pairs for a later moment in the experiment. Each participant received a different random presentation order of the word pairs. In phase two, participants received different instructions for three subsets of word pairs which were randomly presented per subset (16 word pairs). These subsets were obtained by creating three random sets from the original 48 pairs. For one subset of 16 word pairs, participants were instructed to restudy every word pair for 8 seconds (re-read them silently as often as they could), for 16 word pairs to elaborately restudy each pair with help from the mnemonic aids for 8 seconds. For the remaining 16 word pairs participants were instructed to test themselves with a cued-recall test with the Swahili words as a cue. Each Dutch target word had to be typed in within 7 seconds. In the 8th second, the correct answer was shown as feedback. In this second phase, word pairs were again shown in the middle of a computer screen, but on a white background instead of a colored background,

to make sure that participants would refer to the initial learning phase when asked for episodic details of this phase. After this phase, 12 participants were dismissed and asked to return one week later, the other 22 participants were asked to take a cued-recall test for all 48 word pairs (phase 3). Before starting with the cued-recall test, participants received an instruction on how to answer the questions and how to indicate their memory awareness. This instruction (see below) consisted of a Dutch translation and adaptation of the instruction used by Conway et al. (1997) and Herbert and Burt (2003):

"From the 48 studied word pairs you are going to see the Swahili cues. Type in the Dutch translation for every cue. You have 8 seconds to do so for every cue. Even if you do not know the Dutch translation, do type in a word (take a guess if necessary). After you have typed your answer, indicate within 3 seconds which memory awareness accompanied your answer. Choose between Remember, Just Know and Guess (R/K/G). You will find a description of every awareness category below:

Remember: You remember a specific detail of the environment in which you learned the specific item of information. In this case you might have images and feelings in mind relating to the recalled information. Perhaps, you virtually 'see' the background color of the computer screen when you saw the word pair the first time. If something like this is the case, you have a so called "Remember response".

Just Know: You might 'just know' the Dutch translation of the Swahili word that was just presented to you. You know the translation but you would not recall any specific details of the situation wherein you learned the word pair. If something like this is the case, you have a so called "Know response".

Guess: Finally, when you don't know the translation of the Swahili word and you don't have any feeling of recognition of the Swahili word or any idea what the translation might be, you answer with a "Guess response".

After participants had indicated their memory awareness for a recalled target, they were asked to indicate what color the background of the computer screen had when they studied the word pair the first time. They had to choose (guess if necessary) between B (for blue, "blauw" in Dutch), R (for red, "rood" in Dutch), Ge (for yellow, "geel" in Dutch) or Gr (for green "groen" in Dutch) within 4 seconds. Finally, participants were asked which (if any) association they had during learning this particular word pair and to write it down within 8 seconds. They were explicitly told not to repeat the Dutch translation of the Swahili word. After this question, the next Swahili cue appeared on screen and the procedure repeated

itself until the last cue had appeared. The order of appearance of the cues was random for every participant.

After one week, all participants received a final cued-recall test, which was exactly the same as the test in phase three. Participants were again asked to indicate their memory awareness accompanying their answer, the color of the background of the computer screen and (if any) their associations during phase 1. Correct answers on the cued-recall tests were scored by hand to ensure that misspellings or mistakes in plural/singular form (e.g. the answer "clouds" should have been "cloud") were scored as 'correct'.

RESULTS

Table 1 presents the mean proportion of correct answers at the direct test (phase 3) and the final test as a function of condition for the participants who received both tests ($n = 22$). Next to that the mean proportion of correct answers on the final test for the participants who did not receive a direct test after 5 minutes ($n = 12$) are shown. If not reported otherwise, $p = .05$ was used as the threshold for statistical significance for the analyses reported.

To test whether a testing effect had taken place at a retention interval of one week, a one way repeated measure ANOVA was conducted with condition (restudy control, elaborate

Table 1
Mean Proportion Correct Answers on the Tested Items in Phase 3 and on the Final Test as a Function of Condition (standard deviations in parentheses)

	Restudy control	Elaborate Restudy	Test
Mean Performance Phase 3 (n=22)	.60 (.25)	.53 (.22)	.52 (.15)
Mean Performance Final Test (n=22)	.40 (.18)	.34 (.22)	.39 (.16)
Mean Performance Final Test (n=12)	.35 (.25)	.21 (.12)	.43 (.19)

restudy, or testing) as within-subjects variable and proportion of correct answers as dependent variable on the data of the 12 participants who received only the final test. This analysis revealed an effect of relearning condition, $F\ (2, 22) = 6.994$, $p < 0.05$, *partial* $\eta^2 = .389$. Mauchley's test of Sphericity showed a violation for the factor condition. Therefore, the more conservative Greenhouse-Geisser estimates were used. Pairwise comparisons (Bonferroni) showed that testing resulted in a significantly higher mean proportion of correct answers than in the elaborate restudy condition ($p < 0.01$). The other pairwise comparisons were not significant ($p > 0.05$).

To test whether a testing effect had taken place directly after the study phase, a one way repeated measures ANOVA was conducted with condition (restudy control, elaborate restudy, or testing) as within-subjects variable and proportion correct answers (phase 3) as dependent variable. For this analysis the data of the 22 other subjects were used, who received a test directly after the study phase and another one after one week. This analysis did not show an effect of condition, $F(2, 42) = 1.571$, $p = .220$, *partial* $\eta^2 = .070$. Hence, directly after the study phase, the three conditions did not differ in the mean proportion of correct answers.

At the immediate test (phase 3) and at the final test, participants were asked to indicate what the background color of the screen was in phase 1. This was a control variable to check whether participants named the correct color after a remember response and an incorrect color after know or guess responses.

It appeared that participants were not quite able to correctly indicate the background color of the computer screen. In on average 21.4 percent of the cases participants named the right background color of the screen (independent of memory awareness), which is close to chance level given that there were four alternatives. Therefore we decided to exclude this variable from any further analysis. The relative proportions of correct answers of the three different forms of memory awareness (Remember, Know, or Guess) as a function of condition (restudy control, elaborate restudy, or test) are shown in Table 2. These relative proportions of correct answers, are obtained by calculating per participant the proportion of remember, know, or guess responses within the total number of correct responses. One would expect that these proportions add up to 100%. In the current study participants sometimes did not indicate their memory awareness, which resulted in a non-response occasionally. This explains why the proportions do not always add up to 100%.

To test whether there was a difference in remember and know responses between the three conditions at the direct test, two one-way repeated measures ANOVA were performed on the data of the 22 participants who received the direct cued-recall test as well as the

Table 2

Relative Proportions of Correct Remember, Know, and Guess Responses as a Function of Condition (Restudy Control, Elaborate Restudy, or Test) at the Direct Test and after a Retention Interval of 7 Days

	Restudy Control	Elaborate Restudy	Test
Remember (directly)	.18	.30	.22
Remember (after 7 days)	.30	.31	.20
Know (directly)	.69	.53	.65
Know (after 7 days)	.51	.54	.63
Guess (directly)	.10	.11	.07
Guess (after 7 days)	.12	.07	.14

cued-recall test at a 7 days retention interval. Condition (restudy control, elaborate restudy, and test) was the within-subject variable in both ANOVA's and the proportions of correct remember responses or know responses were the dependent variables in the two ANOVA's. Although the first ANOVA did not show an effect of condition, there was a tendency towards this effect in favor of the elaborate restudy condition, $F(2, 42) = 2.889$, $p = 0.067$, *partial* $\eta^2 = .121$. At the direct test, given a correct answer, the mean proportion of remember responses was .30 for the elaborate restudy condition, as compared to .18 for the restudy control condition and .22 for the testing condition.

The second ANOVA showed an effect of condition, $F (2, 42) = 3.553$, $p < .05$, *partial* $\eta^2 = .145$. Pairwise comparisons (Bonferroni) showed that at the direct test, given a correct answer, the proportion correct know responses in the restudy control condition ($M = .69$) was higher than in the elaborate restudy ($p = .10$). This was marginally significant. Given a correct answer, the proportion know responses was lower in the elaborate restudy condition ($M = .53$) than in the testing condition, but this was not significant ($p = .13$). Neither was the difference significant that was found between the testing condition ($M = .65$) and the restudy control condition ($p = 1.00$).

To test whether a shift had taken place in memory awareness as a function of condition between the direct test and the test at 7 days retention interval, two two-way repeated measures ANOVA's were performed on the data of 22 participants who received both tests. Retention interval (direct, and after one week) and condition (restudy control, elaborate restudy, and test) were the within-subject variables and the proportion correct remember (ANOVA 1) and know (ANOVA 2) responses of the total number of correct answers was the dependent variable.

The first ANOVA did not show any main and/or interaction effects, indicating that the relative contribution of remember responses to the correct answers, remained stable over time and for every relearning condition.

The second ANOVA neither showed a main effect of retention interval, $F (1, 21) = 1.879$, $p = .185$, *partial* $\eta^2 = .082$. The main effect of condition approached significance, $F (2, 42) = 2.546$, $p < .10$, *partial* $\eta^2 = .108$, indicating that, on average over time, the mean proportion of know responses to correct answers, differed for the three conditions. Pairwise comparisons (Bonferroni) showed that the difference between the restudy control condition (.60) and elaborate restudying condition (.53) and the difference between the restudy control condition and testing (.64) was not significant (smallest $p = .532$), but that the difference between elaborately restudy and testing approached significant ($p = .09$), indicating that the mean contribution of know responses to correct answers was larger for testing than for elaborately restudying. Finally, the interaction effect between retention interval and condition also approached significance, $F (2, 42) = 2.940$ $p = .06$, *partial* $\eta^2 = .123$, indicat-

ing that there were differences in the changes over time among conditions. Over time the contribution of know responses to the correct answers, diminished in the restudy control condition, while it remained stable in the elaborate restudy and testing condition.

In sum, the mean contribution of remember and know responses to correct answers remained stable over time, but tended to differ between conditions for the know responses. Since the interaction between retention interval and condition approached significance, the results indicated that the mean contribution of know responses in the restudy control condition decreased as a function of retention interval while the mean contributions of know responses to correct answers remained stable over time in the elaborate restudy condition and the testing condition.

DISCUSSION AND CONCLUSION

This study was designed to investigate the elaborative processing hypothesis of the testing effect, by including an elaborate restudy condition in a standard testing effect experiment.

In our study, a testing effect was found at a retention interval of one week, but no testing effect was found directly after the learning phase. At the cued-recall test after one week, memory performance was best in the testing condition, compared to the restudy control condition and to the elaborate restudy condition. However, the testing condition only significantly differed from the elaborate restudy condition and not from the restudy control condition (although there was a non-significant difference in favor of the testing condition). These results are somewhat in line with our expectations. We did not expect a testing effect at the direct test considering earlier research on the testing effect (e.g., Roediger & Karpicke, 2006b). If the elaborative processing hypothesis was true, elaborately processing should result in better performance than the restudy control condition and perhaps even in comparable performance as in the testing condition. This result has not been found. On the contrary, the testing condition only significantly outperformed the elaborate restudy condition and not the restudy control condition.

The elaborative processing hypothesis as explanation of the testing effect was not supported, because elaborately restudying did not result in better memory compared to the restudy control condition. This in line with recent research from Karpicke and Smith (2012) who also found results opposing the elaborative processing hypothesis.

Although we did not find support for the elaborative processing hypothesis, more research would be necessary before we could reject this hypothesis with any certainty, because there are some limitations to our study. For example, the relearning conditions were manipulated within subjects, and although there were no explicit signs of it, it cannot be ruled

out that participants elaborated on the learning material in the plain restudy condition. This was only requested from them in elaborate restudy condition, but they might have thought that is was a better way of restudying than plain restudying and therefore also elaborately restudied the material in the plain restudy condition. So even though testing resulted in better memory performance than both restudy conditions, in a between-subjects design, plain restudying could lead to lower performance than elaborately restudying. Future research could address this.

A second goal of this study was to further elaborate on role of recollection and familiarity in the testing effect using the remember-know procedure of Tulving (1985). At the direct cued-recall test, the mean proportion of correct remember responses was highest for the elaborate restudy condition. This effect was marginally significant. In addition, we found that the mean proportion of correct know responses differed significantly between conditions at the direct test. That difference was largest and marginally significant between the restudy control condition and the elaborate restudy condition. The other two possible comparisons were not significant. Thus, at the direct test the proportion of correct remember responses was highest in the elaborate restudy condition, while the proportion of correct know responses was highest in the restudy control condition. This is in line with the research done by Gardiner (1988) and by Gardiner et al. (1994).

We wanted to take this a step further and to exploratory investigate what would happen to the proportion of correct remember and know responses on the longer term and whether that would be different for the three conditions. Our results showed that the mean contribution of remember responses to correct answers remained stable over time, while there was a tendency towards an interaction effect concerning the contribution of know responses. This indicated that the contribution of know responses to correct answers diminished over time in the restudy control condition, while the contribution of remember responses to correct answers remained stable in the elaborate restudy and testing condition. This is an interesting finding, since this proportion of know responses was highest in the restudy control condition at the direct test. It seems that the accuracy of familiarity after plain restudying diminished over time, while this accuracy remained stable in the testing and elaborate restudying condition. On the contrary, the accuracy of more conscious recollection processes remains stable over time in general.

This is in line with research from Chan and McDermott (2007, experiment 2) who investigated the effect of testing on remember-know awareness in line with the dual processing account (e.g., Jacoby, 1991; Yonelinas, 2002). In this experiment participants studied six word lists. Three lists of words were followed by a free recall test and a distracter task, while the other three lists were followed by a distracter task only (no-testing condition). After the distracter task, participants were given a final recognition test, that required them to choose

between three judgments (remember, know, or new). The results of the experiment showed no testing effect for tested items in the hit rates. It did show more remember responses (for hits) in the testing condition compared to the no-testing condition and also less know responses (for hits) in the testing condition compared to the no-testing condition. Chan and McDermott (2007) concluded from this that testing (as compared to no-testing) changed the manner in which participants decided whether they recognized an earlier studied item. After testing they based their judgment on recollection rather than on familiarity. In this experiment by Chan and McDermott (2007), testing was not compared to a restudy control condition, but it could be the case that participants would base their judgments more often on familiarity than on recollection after a restudy control condition. That is, Chan and McDermott (2007) considered their results as an endorsement of their ideas that the testing effect can be masked because of this greater reliance on familiarity processes that compensates for lack of recollection. An implicit assumption that can be deduced from this idea is that recollection is more accurate than familiarity, especially after a longer retention interval between initial studying/testing and the final test. This could explain why the testing effect is often not found after a short retention interval when the chance of being correct on basis of familiarity is rather high (e.g., Roediger & Karpicke, 2006b). In contrast, after a longer retention interval, as in the present study, a beneficial effect from testing as compared to restudying is often found. This could be explained by the idea that on the long term, decisions on basis of familiarity become less accurate. The fact that in our study the mean contribution of correct know responses to correct answers diminished only in the restudy control condition, is in line with this idea.

In sum, this study contributed to the idea that the elaborative processing hypothesis of the testing effect might not be tenable, but further research is needed to establish this. Next to that, this research contributed to the line of research investigating recognition memory in terms of dual-processing accounts (e.g., Jacoby, 1991; Yonelinas, 2002).

5 | Further evidence that the elaborative processing hypothesis cannot account for the testing effect*

* A modified version of this chapter is submitted as: Schaap, L., Verkoeijen, P. P. J. L, & Schmidt. H. G. Further evidence that the elaborative processing hypothesis cannot account for the testing effect.

ABSTRACT

The "testing effect" is a well-known phenomenon in cognitive psychology and refers to the beneficial effect on long-term memory of taking tests compared to restudying the material. One of the main theoretical explanations of this effect is the elaborative processing hypothesis. A prediction following from this hypothesis is that elaborately restudying of the material would enhance memory performance, perhaps even to a similar extent as testing does. We therefore compared an elaborate restudy condition with a testing condition. Two different elaborate restudy conditions were investigated. In one condition participants had to come up with their own elaborate mediator to study word pairs and in the other condition participants received a mediator from the experimenter. We did not find support for the elaborative processing hypothesis in either of the restudy conditions. This is in line with recent research by Karpicke and Smith (2012) and suggests that the elaborative processing hypothesis does not seem to explain the testing effect. Possible alternative explanations of the testing effect in terms of cue diagnosticity are discussed.

INTRODUCTION

The last two decades the so called testing effect has gained renewed interest (e.g., Cull, 2000; Glover, 1989; Karpicke & Roediger, 2007) in cognitive and educational psychology. The testing effect refers to the finding that taking an initial test leads to better memory performance than restudying after a multiday retention interval. This effect has often been studied experimentally in a standard design (see Chapter 4 of this dissertation), in which an initial study phase is followed by one or several restudy phases which are compared with an equivalent number of test phases or a combination of restudy and test phases (often called a test-restudy condition). After a retention interval of several minutes or multiple days, all participants receive a final memory test on the content of the material studied in the initial learning phase (e.g., Roediger & Butler, 2011). Although the testing effect has been demonstrated repeatedly (e.g., Roediger & Karpicke, 2006b) it is still unclear what mechanism underlies the effect. The present study aims to address one of the hypotheses regarding the underlying mechanisms of the testing effect: the elaborative processing hypothesis.

Elaborately processing

Nowadays, the elaborative processing hypothesis is often mentioned as underlying explanation of the testing effect (e.g., Roediger & Butler, 2011). This hypothesis states that the act of retrieval is more elaborate than the processes involved in restudying. Hence, taking an initial test will lead to more elaborate memory traces and therefore to more retrieval cues than restudying, which in turn will improve memory performance for tested information (e.g., Carpenter, 2009). On the other hand, a negative side effect from testing is the greater amount of erroneous information that is also recalled at a final test. This erroneous information is often semantically associated with the target (e.g., McDermott, 2006). This can easily be explained in terms of elaborative processing. That is, the more elaborate memory traces will not only enhance the number of retrieval cues, but also enhance the activation of related concepts which are therefore more easily recalled. Anderson (1976) gave an excellent example of a possible elaborate process of studying the word pair *dog-chair*. Imagine a dog that loves his master, but also loves to sit on his master's chair. One day the dog climbed on the master's black chair and his white hairs were all over it. The dog is punished for sitting on this black chair. As a consequence of this imagination, an elaborate structure is activated around the word pair *dog-chair*. As a result, multiple pathways between dog and chair are activated, because the activation that would have been confined solely to the word pair *dog-chair*, is now spread to other concepts as well (love, master, sit, hair, punishment). What other concepts are activated exactly and how this helps future memory performance, however, does the elaborative processing hypothesis not explain. The mediator effective-

ness hypothesis, posed by Pyc and Rawson (2010), tries to be more explicit about that point than the elaborative processing hypothesis.

Mediator effectiveness hypothesis

According to the mediator effectiveness hypothesis (Pyc & Rawson, 2010), testing can improve memory by means of creating effective mediators during encoding. A mediator is defined as a word, phrase, or concept that links the cue to the target (e.g., the cue word and target word in a word pair) and when retrieved, it is assumed to strengthen the connection between cue and target. To obtain empirical evidence for the mediator effectiveness hypothesis, Pyc and Rawson (2010) instructed their participants to study a list of Swahili-English word pairs. After this initial study phase, three blocks of practice trials followed. In a practice trial, participants either restudied the word pairs or were given a cued-recall test followed by restudy. During the initial study and the restudy trials, participants had to think of a mediator that could facilitate their learning. After one week, a final cued-recall test was administered. At this final test the target was cued in three different ways. Either only the original cue was given (C-group), or the cue plus the generated mediator was given (CM-group), or only the cue was given with the instruction that participants had to recall their generated mediator first, before recalling the target word (CMR-group). Pyc and Rawson (2010) predicted that mediators that were generated during a test plus restudy condition would be more likely to improve memory performance than mediators that were generated during a restudy only condition. The rationale for this prediction was that the mediator would be retrieved when the cue is presented and would therefore aid retrieval of the target from memory. Mediators that were retrieved during testing (as compared to no-testing) would be more likely to be retrieved during a subsequent test, which would increase the chance of recalling the target at that subsequent test, which would therefore lead to increased memory performance. In this study a general testing effect was found at the final cued-recall test in all three cueing conditions (C, CM, and CMR group). At the cued-recall test in the C-group, memory performance was almost three times better in the test-restudy condition than in the restudy condition. In the CM-group a general testing effect was also found, though it was smaller. The reason for the smaller effect becomes apparent when comparing the final test scores of the C-group and the CM-group. It was found that memory performance was equal in the testing condition between these two groups, but in the restudy condition the CM-group outperformed the C-group. This finding indicated that providing the mediators was beneficial in the restudy condition, but seemed to be superfluous in the test-restudy condition. In other words, providing mediators (CM-group) resulted in a smaller testing effect (smaller difference between the test-restudy condition and the restudy condition) than in the C-group. When the final test performance in the CMR-group was examined, the testing

effect was of a similar size as in the C-group. However, recall of mediators was greater in the test-restudy condition than in the restudy condition (51% versus 34%). For the mediators that were actually retrieved, memory performance (indicated by the percentage of correctly recalled targets after correctly recalling the mediator) was best for the test-restudy condition, but the restudy condition also benefited from the mediator retrieval. Pyc and Rawson (2010) concluded from this that in the test-restudy condition more successful mediators were generated than in the plain restudy condition, because when a mediator was retrieved at the final test, retrieving the accompanying target was more successful in the test-restudy condition than in the restudy condition. Pyc and Rawson (2010) therefore reasoned that successful retrieval of mediators during tests may strengthen the memory paths between the cue and the mediator, and between the mediator and the target. They also suggested that unsuccessful retrieval of mediators may incite participants to think of another more successful mediator, which in turn increases memory performance on the long term. This was supported by the finding that participants changed their mediators more often in the test-restudy condition than in the restudy condition (in 25% versus 19% of the trials).

The fact that in 51% of the cases (at best) the mediator was successfully retrieved and that participants tended to change their mediators could be interpreted as a sign that participants found it hard to come up with effective mediators. This idea inspired us to investigate the role of elaborate retrieval practice a bit further. Even more, because recently some other studies have claimed that another mechanism than elaborate retrieval might be responsible for the testing effect (e.g., Karpicke & Smith, 2012). This is interesting since several studies have found results in line with the elaborative processing hypothesis (e.g., Carpenter, 2009; Pyc & Rawson, 2009), and the mediator effectiveness hypothesis (e.g., Carpenter, 2011).

Elaborate retrieval practice

Karpicke and Smith (2012) extended the line of research Pyc and Rawson initiated and conducted two experiments in which participants used an imagery-based keyword method or a verbal elaboration method as elaborate restudy conditions. The imagery-based keyword method is a study method to learn word pairs aided by a keyword mnemonic and to form an image of the word pair with the mnemonic. For example for the word pair *loggia – balcony*, the mnemonic would be *log* and the learning instruction would be "loggia sounds like log and means balcony". Subsequently, participants are told to form a mental image of a log lying on a balcony. The verbal elaboration method comprised the instruction to think of a word that could help remember the word pair (comparable to the mediator in Pyc and Rawson's [2010] study). For example, when a participant has to learn the word pair *wingu – cloud*, the participant may elaborate and think of the word bird to relate *wingu* (which resembles the word '*wing*') to *cloud*. Participants in the Karpicke and Smith study (2012) had to study

and test themselves on word pairs until they could correctly recall the target with the cue. After a target was recalled correctly for the first time, three possible conditions could follow. Either nothing happened and the item was dropped from further practice, or the item was elaborately restudied, or it was retested. A final cued-recall test was administered after a one week retention interval.

From this study of Karpicke and Smith (2012) it appeared that testing benefited memory more than elaborately restudying, but this may have been due to the fact that elaboration only occurred after accurate recall. Therefore, in a second experiment they varied elaboration during encoding in the initial learning phase. Elaborately processing was only beneficial when it happened at the initial learning phase before successful retrieval. That is, participants in the elaborate learning condition performed better in the initial learning phase than participants in the no-elaboration condition. However, no improvement was found as a result of elaborately restudying after the first successful retrieval attempt. This means that testing was more beneficial than elaborately restudying after the participants were able to correctly recall the word pair for the first time. Although this study seems to indicate that elaboration is not the underlying explanation of the testing effect, one can wonder whether the participants in this study had to elaborate during restudy to a similar extent as when they were retrieving the target. The elaborate restudying condition comprised generating a visual image or a keyword to help remembering the word pair. Although the imagery-based keyword method and the verbal elaboration method have proven to be beneficial for vocabulary learning (e.g., Pressley, Levin, & Delaney, 1982), none of those keyword methods use semantic elaboration strategies. They are meaningful, but not semantic. For instance, in the example of the word pair *wingu – cloud*, participants have to think of a single keyword that relates to the cue and the target. A more semantic approach of elaboration would be: *wingu* contains the word *wing*, which can mean a part of an airplane. When an airplane flies through the sky, the *wing* of the airplane touches the *cloud*. According to the Levels-of-Processing Model of Craik and Lockhart (1972), one would expect better retention after deep processing as compared with shallow processing. The level of processing is dependent on the degree of semantic or cognitive analysis of the material (Craik & Lockhart, 1972). The imagery and/or verbal keyword method do not seem to require the deepest possible levels of processing. It would therefore be interesting to investigate whether a more semantic/comprehensive restudy strategy could be as elaborate as testing.

Present study

In the present study, participants will start with an initial elaborate study phase in which they study 40 word pairs either with a given mediator or with a self-generated mediator. These are called the mediator conditions. In the given mediator condition, participants are given

mediators which require quite a lot of semantic elaboration. In the self-generated mediator condition participants have to think of mediators themselves. After the initial study phase, half of the word pairs will be tested and the other half of the word pairs will be elaborately restudied (again, either with a given or a self-generated mediator, dependent of mediator condition). In sum, within mediator conditions, initial elaborately studying is followed by either elaborately restudying (eS-eS-eS) or testing (T-eS-T) to test the elaborative processing hypothesis of the testing effect, continuing the work of Karpicke and Smith (2012). At the same time it is investigated whether it makes a difference if participants study by aid of given or self-generated mediators. For an overview of the design of the study see Table 1.

The study by Karpicke and Smith (2012) suggested that elaborately processing does not explain the testing effect, because memory did not improve equally as a consequence of elaborately restudying and testing. However, as explained above, this could also be due to

Table 1
Overview of the Design of the Study

Mediator Condition	Practice Condition	Initial learning phase	Practice	Practice	Practice	Final test phase
Given	Test	eS	T	eS	T	Final Test
	Restudy	eS	eS	eS	eS	Final Test
Self-generated	Test	eS	T	eS	T	Final Test
	Restudy	eS	eS	eS	eS	Final Test

the fact that the restudying task did not ask for an optimal level of semantic elaboration. That is, the elaborate restudy condition is less optimal than the testing condition to begin with and for that reason the elaborative processing hypothesis cannot be tested properly. In the present study two different elaborate initial learning conditions (i.e., mediator conditions) are therefore compared and within these mediator conditions, elaborately restudying is compared with testing to investigate the elaborative processing hypothesis of the testing effect.

From the Pyc and Rawson (2010) study it seemed that it is difficult for participants to come up with successful mediators in a short time (e.g., 8 seconds). In their study, at best, only half (51%) of the mediators were recalled at a final test. Moreover, if the mediator was recalled, this led to recall of the accompanying target in max 70% of the cases. In the third experiment of Karpicke and Smith (2012) participants also had to come up with their own keyword/mediator to elaborately study a word pair. In this experiment, it took participants more time to think of a mediator in the elaborate restudy condition than to recall the target in the testing condition. This might indicate that participants find it hard to come up with

an effective mediator. The two mediator conditions of the present study make it possible to investigate the differential effects from given or self-generated mediators on memory performance.

In sum, assuming that the elaborative processing hypothesis is correct and that participants would need to be handed effective mediators to be able to elaborately study the word pairs, the following hypotheses can be stated: the elaborate restudy condition with given mediators would result in comparable final test scores as the testing condition, while the elaborate restudy condition with self-generated mediators would result in worse performance on a final test than the testing condition (i.e., a testing effect would be found).

METHOD

Participants

Seventy adults (*M* age = 29.51 years, range 18-49; 19 male) participated in this study for course credits, payment, or no compensation. Of the participants, 52 were higher educated (bachelor or master degree) adults recruited via the social network of the first author and 18 participants were students of eight different bachelor and master programs of a Dutch University (psychology, international business administration, culture studies, marketing, law, medicine, and public administrations).

Materials and procedure

Forty Swahili – English word pairs were selected from the Nelson and Dunlosky (1994) study and translated into 40 Swahili – Dutch word pairs. For every word pair a mnemonic aid was constructed. For example for the word pair: *ardhi – grond* (*soil* in English) was: '*ardhi*, sounds like *aarde* (*earth* in English), which is another Dutch word for *grond* (*soil*)'. The study consisted of an initial learning phase followed by three practice sessions. Half of the participants (randomly assigned) studied every word pair in the initial learning phase with a given mnemonic aid, the other half of the participants was instructed to come up with a mnemonic aid that would help them remember the word pair. In both mediator conditions participants studied every word pair for eight seconds and the word pairs were presented in random order. In the three following practice sessions all participants elaborately restudied 20 of the word pairs while the other 20 word pairs were tested, elaborately restudied and tested again. Which 20 word pairs were restudied or tested was counterbalanced across participants. Again, eight seconds were reserved to study or test each word pair and word pairs were presented in random order. After a one week retention interval a final cued-recall test was administered. This final test was self-paced. Participants randomly received one of

four versions of the final test that differed in the order of the word pairs. Participants were not only asked to recall the target, but also the mnemonic aid that they had used to study the word pair.

RESULTS

Check of assumptions

To check whether our assumption that it is difficult to come up with an effective mediator in a very short time was correct, we conducted a repeated measures ANOVA with mediator condition (given or self-generated) as between-subjects factor, practice condition (restudy or testing) as within-subjects factor, and the number of recalled mediators at the final test as dependent variable.

This analysis showed a main effect for mediator condition, $F(1, 68) = 9.378$, $p < .01$, *partial* $\eta 2 = .121$, indicating that, at the final test, participants from the given mediator condition ($M = 13.38$, $SD = 5.74$) recalled significantly more mediators than participants in the self-generated mediator condition ($M = 9.22$, $SD = 6.90$).

This analysis showed no main effect for practice condition, $F(1,68) = 2.969$, $p = 0.89$, *partial* $\eta 2 = .042$, indicating that on average and irrespective of mediator condition, participants in the testing condition recalled the same number of mediators ($M = 12.17$, $SD = 6.29$) as participants in the elaborate restudy condition ($M = 11.24$, $SD = 6.66$). According to the elaborative processing hypothesis, the process of testing is more elaborate than the process of restudying. At the same time is there a negative side effect from testing. That is, after retrieval practice a great amount of erroneous information is recalled at a final test. This seems plausible in terms of elaborately processing, because this erroneous information is often semantically associated with the target (e.g., McDermott, 2006) and can therefore be a result of co-activation of related concepts of the target during testing. To check our assumption that the given mediator condition would be comparable to the testing condition in terms of elaborately processing, we compared within the given mediator condition, the incorrect answers of the restudy condition with the incorrect answers in the testing condition. It would be likely to assume that if the two conditions are comparable in terms of elaborately processing, the amount of semantically associated but erroneous information that is recalled at a final test in the two conditions is also comparable. We therefore scored one point for every incorrect target that was semantically related to the mediator. That is, for every incorrect answer that contained (part of) the mediator or an incorrect target that had been incorrectly deduced from the given mediator, we gave participants one point. To investigate whether there was a difference between the given mediator restudy condition and

the given mediator testing condition, a paired-samples t-test was conducted. On average, participants in the given mediator testing condition recalled the same number of incorrect but semantically related targets ($M = 2.47$, $SD = 1.98$) as participants in the given mediator restudy condition ($M = 2.88$, $SD = 2.47$), $t\ (33) = .845$, $p > .05$, $r = .16$. This indicated that the negative side effects (in terms of erroneously recalling semantically related information instead of the correct target) were the same for the testing condition and the elaborative restudy condition within the given mediator condition.

Interaction between mediator condition and testing condition

The mean numbers of correct answers on the final test for each within-subjects condition and between-subjects condition are shown in Table 2. A mixed ANOVA with mediator condition (given or self-generated) as between-subjects factor, practice condition (restudy or testing) as within-subjects factor, and the number of correctly recalled targets at the final test as dependent variable was conducted to test the hypothesis that the elaborate restudy condition with given mediators would result in a comparable memory performance on the final test as the testing condition and that a general testing effect will be found between the testing condition and the elaborate restudy condition where the participants need to invent their own mediators. In other words, an interaction effect is expected between mediator condition and practice condition.

This analysis showed a main effect for practice condition, $F\ (1, 68) = 22.703$, $p < .01$, *partial* $\eta 2 = .250$, indicating that testing was more beneficial to long-term memory of the

Table 2
Mean number of correct targets on the final test for each condition (standard deviations in parentheses).

	Self-generated mediator condition	Given mediator condition
Restudied items	7.61 (5.38)	6.68 (4.89)
Tested items	9.78 (4.88)	9.35 (5.33)

Dutch-Swahili word pairs than elaborately restudying (irrespective of mediator condition). In contrast to our expectations, no interaction effect was found between mediator condition and practice condition, $F\ (1, 68) = .252$, $p = .618$, *partial* $\eta 2 = .004$, indicating that the difference in final test performance between testing and restudying in the given mediator condition was of the same size as the difference in final test performance between testing and restudying in the self-generated mediator condition. In other words, testing was more beneficial to memory performance irrespective of mediator condition.

DISCUSSION AND CONCLUSION

The present study was designed to investigate the elaborative processing hypothesis of the testing effect. This hypothesis states that the act of retrieval is more elaborate than the processes involved in restudying. Hence, taking an initial test will lead to more elaborate memory traces and therefore to more retrieval cues than restudying, which in turn will improve memory performance for tested information (e.g., Carpenter, 2009). Although several studies have found results that seem to confirm the elaborative processing hypothesis (e.g., Carpenter, 2009; Pyc & Rawson, 2009), recently other researchers have claimed that elaboration is probably not the explanatory factor (e.g., Karpicke & Smith, 2012).

A prediction that follows from the elaborative processing hypothesis is that elaborately restudying the learning material will dissolve or at least diminish the testing effect. The present study investigated this prediction by comparing a testing condition with an elaborate restudy condition after an initial elaborate learning phase. The results from our study showed that the testing condition still outperformed the restudy condition on a final retention test at a one week retention interval. The study by Pyc and Rawson (2010) already made clear that elaboration can foster learning, but that it is difficult for participants to come up with effective mediators. If elaborate restudy conditions are not effective because participants have difficulties to come up with a mediator to link the cue and the target, the comparison between an elaborate restudy condition and a testing condition is not correct. That is, the elaborate restudy condition is less optimal than the testing condition to begin with and for that reason the elaborative processing hypothesis cannot be tested properly. The present study therefore compared two conditions of elaborate restudy, namely a condition in which participants had to come up with a mediator and a condition where they were given a mediator to help them remember a target when presented with a cue. If comparisons between elaborate restudy and testing were influenced by the fact that studying word pairs with a self-generated mnemonic aid is too difficult, one would expect a differential effect from self-generated and given mediators. Although participants in the given mediator condition did on average recall more mediators at the final test than the participants in the self-generated mediator condition, no differential effect of mediator condition was found on the mean number of correct targets that was recalled at the final test in either restudy condition. This could indicate, in line with the results from the Karpicke and Smith (2012) study, that the testing effect cannot be explained by elaborately processing.

The question that arises, then, is what can explain the testing effect? Karpicke and Smith (2012) do not answer this question, and neither does the present study provide a conclusive answer, but we can speculate on one potential explanation based on all of the studies. One possible explanation is that during testing the only memory path that is strengthened is the

path between cue and target, and hardly any other memory path is activated. This strengthening might enhance the all or none chance of recalling the target. Although the participants in our study recalled on average the same number of mediators in the restudy condition as in the testing condition, in the testing condition the number of correctly recalled targets was still higher than in the restudy condition. In addition, participants in the given mediator condition recalled significantly more mediators at the final test than participants in the self-generated mediator condition, but the number of targets recalled on the final test was the same for both mediator conditions. For some reason, the additional number of recalled mediators at the final test in the given mediator condition (as compared to the self-generated mediator condition) did not help the participants in that condition to recall more targets at the final test than participants in the self-generated mediator condition either in the restudy or in the testing condition. These results do not completely fit the mediator effectiveness hypothesis (e.g., Pyc & Rawson, 2010). According to this hypothesis, a retrieved mediator will aid retrieval of the target from memory, which will lead to an increased memory performance. However, in the two mediator restudy conditions of this study, no significant difference in memory performance was found, while the numbers of retrieved mediators did significantly differ between the two conditions. Moreover, the number of recalled mediators at the final test was comparable between the testing and restudying condition. Even though the same number of mediators was retrieved, still a beneficial effect from testing was found. Thus it seemed that either the effectiveness of the mediator is higher after testing, or the role of mediators in the testing effect is not as important as is presumed by the mediator effectiveness hypothesis. It could be that retrieval practice results in other benefits than adding extra information to the memory path. We should note, however, that the testing condition in the present study was not a 'pure testing' condition (i.e., it was a test-restudy condition), therefore it cannot be ruled out that the sequence of testing, restudying, testing has boosted the results in the testing condition.

We also looked into the *incorrect* answers participants gave in the given mediator condition. A negative side effect from testing can be the greater amount of erroneous information that is recalled after retrieval practice at a final test. This erroneous information is often semantically associated with the target (e.g., McDermott, 2006). We investigated whether the amount of semantically related information that was recalled at the final test differed between the testing and restudy condition, assuming that it would be the same as both conditions are comparable in terms of elaborately processing. The negative side effects (in terms of erroneously recalling semantically related information instead of the correct target) were the same for the testing condition and the elaborate restudy condition. This is an interesting finding, because participants in the testing condition obtained a higher final correct recall score than participants in the restudy condition. In other words, in both condi-

tions participants recalled the same number of incorrect targets, but in the testing condition the amount of correct recall of targets was higher than in the elaborate restudy condition. We cannot explain this in terms of the elaborative processing hypothesis. It seemed that our elaborate restudy condition was indeed elaborate, but still participants benefited more from retrieval practice than from elaborately restudying in terms of correct final cued-recall scores. Apparently, the chance to recall a target when presented with a cue is higher after retrieval practice than after elaborately restudy. This is what Nairne (2002) called the diagnostic value of a cue. When presented with a cue, we use that cue to decide which of viable candidate targets was the one studied. When distinctive features of a target are linked to a specific cue, the diagnostic value of that cue is high (Nairne, 2002). Retrieval presumably enhances this diagnostic value of the cue, while elaborately restudy does not or to a smaller extent. This was also suggested by Karpicke and Blunt (2011). It seems that it is not so much the elaborate nature of the retrieval process, but the strengthening of the path between cue and target as a result of successful retrieval that is responsible for increasing the diagnostic value of a particular cue. In case of elaborate restudy multiple pathways are activated, which not only increases the chance to correctly recall the target, but also the chance to recall one of the other information parts that are activated during encoding. However, the diagnostic value of the cue may not be increased, because the correct memory path between cue and target has not been strengthened in a retrieval process. In other words, instead of adding features to the memory path (which probably happens during elaboration), retrieval practice constrains the set of features that may help to find the target based upon a certain cue. Future research could focus on testing this idea.

6 | Test-taking strategies that require more effortful retrieval do not influence the testing effect*

*A modified version of this chapter is submitted as: Schaap, L., Verkoeijen, P. P. J. L, Coppens, L. C., Nugteren, M., & Schmidt. H. G. Test-taking strategies that require more effortful retrieval do not influence the testing effect.

ABSTRACT

This study was concerned with testing the retrieval effort hypothesis of the testing effect, which states that the more effortful retrieval is the more beneficial testing as relearning opportunity will be. This hypothesis was tested by comparing the effect of two different test-taking strategies and a restudy condition on long-term retention. In a pilot study and a main experiment, participants studied a text and after this study phase, participants were randomly assigned to one of three relearning conditions: a direct choice multiple choice (MC) condition, a generate response MC condition (GR-MC) in which participants had to first answer the question without seeing the MC alternatives and subsequently choose an alternative, and a restudy condition (RC). These three conditions were compared on final MC test performance after a three day (pilot) or a one week (main experiment) retention interval. The GR-MC test-taking strategy was indeed perceived as more effortful, but did not result in higher final test scores than in the MC- testing condition. Although the two testing conditions did improve over time, the final test scores of participants in the two testing conditions were not significantly higher than the final test scores of the participants in the RC. In other words, no testing effect was found.

INTRODUCTION

The "testing effect" is a well known phenomenon in cognitive psychology and refers to the finding that after an initial learning phase, taking tests has a beneficial effect on memory performance compared to restudying the material (e.g., Roediger & Butler, 2011; Roediger & Karpicke, 2006a). This phenomenon has often been studied experimentally in a standard design, in which an initial study phase is either followed by one or more restudy phases or an equivalent number of test phases. After a retention interval all participants receive a final memory test on the content of the material studied (e.g., Roediger & Karpicke, 2006a). The length of this retention interval has been varied in different studies (for example 5 minutes, 2 days, or 7 days). After very short intervals (i.e., 5 minutes) mostly no positive testing effect is found: in fact many studies have observed a memory advantage of restudying over testing. After a longer retention interval (i.e., 2-7 days) on the other hand, testing generally produces a better test performance than restudying (e.g., Roediger & Karpicke, 2006a; Wheeler et al., 2003).

Many experimental studies have demonstrated this positive effect of testing (for an overview, see Roediger & Karpicke, 2006b) and over the years, the cognitive processes that are involved in the testing effect have gradually become clearer (e.g., Carpenter, 2009). Furthermore, the materials that are used in the experiments have become more relevant for educational practice (e.g., Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Butler & Roediger, 2007) and some studies have made a successful attempt to transfer the testing effect to the classroom (e.g., McDaniel et al., 2011; McDaniel et al., 2007; Roediger et al., 2011). This is an important step, because for many years assessment and test-taking have been considered primarily a tool for assessing student learning. Test-taking as an act of learning has not been widely implemented in educational practice yet. However, before it can be widely implemented, further research is needed on the conditions under which the benefits of testing occur in educational practice, and on the mechanisms underlying the testing effect (Roediger & Butler, 2011). This study aims to contribute to this line of research by investigating the influence of different test-taking strategies on memory performance at a final test.

Retrieval effort

Nowadays, most of the theoretical explanations of the testing effect focus on how the act of retrieval affects memory (Roediger & Butler, 2011). The elaborative processing hypothesis (e.g., Carpenter, 2009) states that retrieval processes are more elaborative than restudy processes and this leads to a more elaborative memory trace and therefore to more retrieval cues. A related explanation is the retrieval effort hypothesis (e.g., Pyc & Rawson, 2009).

This hypothesis explains the testing effect by the fact that retrieving information takes more effort than restudying, leading to more elaborate search processes, which in turn results in more activation of related information, and therefore leads to better future memory performance. Although several experiments have found results in line with the elaborative retrieval/retrieval effort hypotheses (e.g., Carpenter, 2009; Pyc & Rawson, 2009), there are still some uncertainties about what happens during these retrieval processes that explains the long-term memory improvement.

Despite those uncertainties, researchers in general agree that for an optimal testing effect to occur, retrieval effort should be high and also successful. This shared opinion fits within the "desirable difficulties framework" (e.g., Bjork, 1994; Pyc & Rawson, 2009). In search of these desirable difficulties, earlier studies already varied retrieval effort by varying test format, using for instance free recall and recognition tests, with free recall presumably being more effortful than recognition. These studies indeed showed a stronger testing effect after a free recall test than after a recognition test (e.g., Butler & Roediger, 2007; Glover 1989, experiment 4). A recommendation that follows from this research is that testing should be effortful, given that it results in successful retrieval during testing. In other words, if testing is effortful, but participants do not succeed in actually retrieving the information asked for by the test, memory performance is not enhanced (e.g., Roediger & Butler, 2011). McDaniel et al. (2007) investigated this recommendation in the classroom and they found a stronger testing effect (as measured with a final MC test) for short-answer quizzing compared to multiple choice (MC) quizzing. This is in line with the desirable difficulties view, because short answer questions in general require more retrieval effort than MC questions and will therefore have a more beneficial effect on memory performance than MC questions. Not only do different testing formats ask for different amounts of retrieval effort, but the way a test format is approached (i.e., test-taking strategy) can also affect the amount of retrieval effort spent. In other words, different strategies to take a certain test can ask for different amounts of retrieval effort, and as such test-taking strategies form an interesting alternative approach towards investigating the testing effect in terms of desirable difficulties.

Test-taking strategies

Research on test-taking strategies showed that students differed in their approaches to answering MC questions. For example McClain (1983) investigated the test-taking behavior of students of different ability levels (A, C and F students). She found that students differed in the number of times they generated an answer to the question before even reading the alternatives. This test-taking strategy of generating an answer before reading the alternatives is comparable to free recall and probably requires more retrieval effort than the strategy of immediately reading the alternatives and selecting the correct one. That is, in

general, generating an answer from memory is considered more effortful than choosing/recognizing the correct answer from different given alternatives (e.g., Kang, McDermott, & Roediger, 2007).

To the best of our knowledge, only two studies have investigated the effect of different test-taking strategies on subsequent memory performance with MC tests (Crocker & Schmitt, 1987; Sensenig, 2010, experiment 2a and 2b). Crocker and Schmitt (1987) investigated the effectiveness of two test-taking strategies on a MC test that assessed the knowledge of statistical concepts three weeks after a statistics course. Participants were students with different levels of test anxiety who were randomly assigned to one of the test-taking strategy conditions. In one condition participants had to generate an answer before reading the alternatives of a MC-question (the response generate condition) and in the other condition they had to directly read the alternatives (MC condition). Response generating was found to be a more effective test-taking strategy than the MC test-taking strategy for the low test-anxious students. No positive or negative effect was found for the students who were in the mid-range of test-anxiousness and for high test-anxious students, the response generate strategy even led to a lower score on their MC-test. Crocker and Schmitt (1987) did not investigate the effect of testing as compared to a restudy condition or the long-term effects of different test-taking strategies on memory performance, but they did show that different test-taking strategies had different immediate effects on performance.

Sensenig (2010) did examine different test-taking strategies in the light of the testing effect. She conducted an experiment with two different MC testing conditions and a restudy condition. The MC-tests assessed factual knowledge. In one testing condition participants were asked to recall the right answer before choosing the correct alternative. In the other testing condition, participants were asked to look directly at the alternatives and to choose the correct answer. In the restudy condition the questions from the testing conditions were stated as facts and participants were asked to re-read them and to mark them on an answer sheet when read. The final test was administered at a five minutes retention interval. Sensenig (2010) did not find any differences between the three conditions after a short retention interval, which is often the case (e.g., Roediger & Karpicke, 2006b), however Sensenig (2010) did not include a final test after a long retention interval (i.e. 2-7 days). Therefore, it remains an open question whether these different test-taking strategies have differential effects on long-term retention.

The current study replicates Sensenig's (2010) design but then with a long-term retention interval, because test-taking strategies are relevant for educational practice and offer possibilities for direct application by MC tests. In educational practice standardized tests are often MC-tests and students practice many of these items in the preparatory process for the

final test. When a certain strategy is more beneficial for long-term memory (and thus for the final test) than another strategy, students' final test performance can be improved.

Present study

In the current study we conducted a standard testing experiment with two different testing conditions and a restudy condition. The two testing conditions differ in the test-taking strategy participants used (generate response versus direct choice). The experiment ends with a final MC-test after a retention interval of seven days. As mentioned above, seen from a retrieval effort perspective, one would expect the generate response strategy to require more retrieval effort than the direct choice strategy. Because the amount of retrieval effort is thought to be positively related to memory performance (e.g., Pyc & Rawson, 2009), the generate response strategy would be expected to have a more beneficial effect on long-term memory performance after a retention interval of at least a few days than the direct choice strategy. Before this experiment, we first carried out a pilot study in order to assess whether participants had enough time to follow our instructions in the different phases of the experiment. This pilot study will be described first.

PILOT STUDY

METHOD

Participants and design

In the pilot study 25 undergraduate students (*M* age = 21.44, range 19-25, 4 male), mainly from a Dutch bachelor program in psychology (n = 20), volunteered for course credits, payment or no compensation. They were randomly assigned to one of the conditions by computer software (E-prime): nine participants were placed in the restudy condition, eight in the direct choice condition and eight in the generate response condition.

Materials and procedure

The experiment consisted of three phases: a learning phase, a testing/relearning phase, and a final test. During the learning phase participants read a text on the 1991 eruption of Mount Pinatubo, a volcano at The Philippines. The text was printed on paper and consisted of 3255 words. It was acquired from the website www.wikipedia.nl. (a free online encyclopedia that anyone can edit). Participants were instructed to study the text for twenty minutes and were informed they were going to be tested on the facts mentioned in this text at a certain point

in the future. For every time they started to reread the text, participants were asked to draw a line on top of the first page. After the study phase, participants were randomly assigned to one of three conditions: a restudy condition (RC), a generate response MC (GR-MC) condition, or a direct choice MC condition (MC). In the RC, participants restudied 18 facts (the same facts that were tested in the other two conditions) from the Mount Pinatubo-text on the computer. Every fact was shown for 22 seconds. An example of a fact from the text is: "The original people who lived on Mount Pinatubo were called the Aeta". In the MC condition, participants were tested on 18 facts from this text with 18 MC-questions with four alternatives. The questions were shown for 19 seconds, irrespective of whether the participant had typed an answer. After 19 seconds, feedback was provided by showing the right answer on the screen for three seconds. In the GR-MC condition, participants received the same 18 MC-questions, but they were first instructed to respond to the question without seeing the alternatives. After giving a response or after 14 seconds, participants were shown the four alternatives and had five seconds to choose the correct answer. After these five seconds participants received feedback for 3 seconds. The order of the questions changed randomly for every test for every participant. In every condition, the total time per trial (for restudying a fact or answering a question and receiving feedback) was 22 seconds. At the end of phase two, all participants were asked to indicate on a scale ranging from 1 (very little) to 7 (very much) how much effort they invested in the previous phase (Van Gog & Paas, 2008). Three days later, the final MC-test was administered. This was the same test participants had received in the second phase with respect to the content. To prevent any pure memory effects of the order of questions or the alternatives, the order of the items and alternatives was changed in the final test. No time constraint was given for answering each question at the final test. After the final test, participants were thanked, debriefed and dismissed.

RESULTS

To check our manipulation that the generate response MC condition would be more effortful than the other two conditions, a one-way ANOVA with condition (restudy, generate response MC, and direct choice MC) as between-subjects factor and effort indication as dependent variable was conducted. This analysis showed an almost significant main effect for effort, $F(2, 22) = 2.874$, $p < 0.08$, partial $\eta 2 = .21$. Pairwise comparisons (LSD) showed that the difference between the GR-MC condition and the MC condition was significant ($p < 0.05$). The other conditions were not significantly different from each other, but the difference between the GR-MC condition and the restudy condition approached significance ($p = .058$). Mean reported effort indications on a 1-7 point Likert scale were 3.78 ($SD = 1.30$)

for the restudy condition, 3.63 (*SD* = 1.69) for the MC condition, and 5.13 (*SD* = 1.13) for the GR-MC condition. This indicated that participants in the generate response MC condition experienced the task as more effortful than participants in the direct choice condition, and in the restudy condition (the difference approached significance).

The mean numbers of correct answers on the final test for each condition are shown in Table 1. A one-way ANOVA with condition (restudy, GR-MC, and MC) as between-subjects factor and the number of correct answers on the final test as dependent variable was conducted to test the hypothesis of this pilot that the GR-MC condition would result in a bigger testing effect than the direct choice MC condition. This analysis, however, showed that final test performance did not differ for the three conditions, $F(2, 22) = 0.33$, $p = 0.719$, $\eta^2 = .03$. In other words, no testing effect was found, let alone additional beneficial effects from a certain test-taking strategy.

Because this was a pilot study, we asked the participants explicitly to comment on the

Table 1
Pilot: Mean number of correct alternatives on the final and initial test for each condition

	Condition					
	Restudy		MC		GR - MC	
	M	SD	M	SD	M	SD
Initial test performance	--	--	12.88	2.53	12.38	2.45
Final test performance	14.67	2.65	14.50	1.93	13.88	1.36

study and we also investigated the raw data in order to find an explanation of these unexpected results. One possible explanation of the results is the length of the retention interval we used. We initially choose to use a retention interval of three days, because after three days a testing effect is often found, but longer retention intervals can lead to a stronger testing effect (e.g., Butler & Roediger, 2007; Carpenter et al., 2009). For the main experiment (see below) we therefore changed the length of the retention interval from three days to seven days. In addition, 16 out of 25 participants (six from the RC, four from the MC condition and six from the GR-MC condition), indicated that they were not able to finish reading the whole text. This might have influenced the results, especially with such small groups (e.g., there were only two participants left in the GR-MC condition who had finished reading the complete text). Possibly, the text was not studied properly and therefore the three conditions (one restudy and two testing) were not so different at all. The facts to restudy were formulated in manner comparable to the MC-questions; with the difference that one factual element was stated in one of four alternatives. For example the question in the testing conditions "In the year .... there was a great eruption of the volcano" was stated as the following

fact in the restudy condition "In the year 1991 there was a great eruption of the volcano". When participants in the initial study condition were not able to finish reading the text, the testing conditions might have been very similar to the restudy condition.

A third point that might have influenced the results is the relatively short time to answer the open questions for the participants in the GR-MC. Four participants admitted to having trouble answering the question as an open question, because they often only had finished reading the question and had not been able to type an answer before the computer program progressed to the alternatives. Therefore, the effect of the strategy in the GR-MC condition may have been lost, because half of the participants were unable to apply the strategy due to the speed at which the questions had to be answered. The aforementioned three points were taken into account to improve the design of the main experiment, which is described in the next section.

MAIN EXPERIMENT

METHOD

Participants

In the main experiment 69 adults (*M* age = 23.00 years, range 19-38, 15 male), volunteered to participate; 59 were students from 14 different bachelor and master studies of a Dutch University (psychology, pedagogy, (international) business administration, biomedical sciences, international communication and media, sociology, financial economics, health sciences, culture studies, marketing, law, medicine, and public administration) and 10 were colleagues. They received course credits, payment, or no compensation for participation in the study. Computer software (E-prime) randomly assigned participants to one of the conditions: 23 participants were placed in the restudy condition, 25 in the MC condition and 21 in the GR-MC condition.

Materials and procedure

The experiment again consisted of three phases: a learning phase, a testing/relearning phase, and a final test. During the learning phase participants read the same text as in the pilot study. They were instructed to study the text for twenty-five minutes (instead of twenty minutes in the pilot) and were informed they were going to be tested on the facts of this text at a certain point in the future. For every time they started to reread the text, participants were asked to draw a line on top of the first page. After the study phase, participants

were randomly assigned to one of the three conditions. The conditions were the same as in the pilot study with the exception that in the GR-MC condition, participants were given 17 seconds (instead of 14 in the pilot) to type in their generated answers before they were shown the four alternatives and had five seconds to choose the correct answer. In both testing conditions participants received feedback after every question (for 3 seconds). The order of questions for every test and for every participant was changed randomly. In every condition, total time per trial (for restudying a fact or answering a question and receiving feedback) was 25 seconds (instead of 22 in the pilot study). At the end of phase two, all participants were asked to rate invested mental effort as in the pilot study. One week later (instead of three days in the pilot), the final MC-test was administered. This was the same test participants had received in the second phase with respect to the content. To prevent any pure memory effects of the order of questions or the alternatives, the order of the items and alternatives was changed in the final test. No time constraint was given for answering each question at the final test. After the final test, participants were thanked, debriefed and dismissed.

RESULTS

To check our manipulation that the generate response MC condition would be more effortful than the other two conditions, a one-way ANOVA with relearning condition (restudy, GR-MC, and MC) as between-subjects factor and effort indication as dependent variable was conducted. This analysis showed a significant main effect for effort, $F(2, 66) = 4.676$, $p < 0.05$, $\eta^2 = .12$. Pairwise comparisons (LSD) showed that participants in the restudy condition and the MC condition reported investing significantly less effort than participants in the GR-MC condition ($p < 0.05$). The effort ratings in the restudy condition and the MC condition did not significantly differ from each other. Mean effort ratings were 3.17 ($SD = 1.27$) for the restudy condition, 3.04 ($SD = 1.59$) for the MC condition, and 4.24 ($SD = 1.37$) for the GR-MC condition. Moreover, when we look at the generated responses (i.e., before participants in the GR-MC condition were shown the possible alternatives), it seems that they were not very successful: they scored on average 7.48 out of 18 possible points. This is a score of 41.6%, which, compared to their scores after seeing the alternatives (MC-score: 70.4%) was rather low. It seemed that the generation of a response was perceived as an effortful task which often did not lead to a correct response.

The mean numbers of correct answers on the final test for each condition are shown in Table 2. A one-way ANOVA with relearning condition (restudy, GR-MC, and MC) as between-subjects factor and the number of correct answers on the final test as dependent

variable was conducted to test the hypothesis that the GR-MC condition would result in a bigger testing effect than the MC condition. As in the pilot study, this analysis showed, however, that final test performance did not differ for the three conditions, $F(2, 66) = 0.471$, $p = 0.626$, $\eta^2 = .01$.

To investigate whether participants' scores changed (differentially) over time, a mixed ANOVA was conducted. The between-subjects factor was relearning condition (MC versus GR-MC) and the within-subjects factor was test phase (initial versus final test). The number of correct answers on both tests was the dependent variable. This mixed ANOVA showed that participants performed significantly better on their final test than on the initial test, $F(1, 44) = 5.946$, $p < .05$, $\eta_p^2 = .119$. However, no interaction effect was found, $F(1, 44) = 1.975$, $p = .167$, $\eta_p^2 = .043$.

Table 2
Mean number of correct alternatives on the initial and final test for each condition

	Condition					
	Restudy		MC		GR - MC	
	M	SD	M	SD	M	SD
Initial test performance	--	--	13.12	2.46	12.67	2.61
Final test performance	13.13	2.58	13.44	2.08	13.86	2.80

DISCUSSION AND CONCLUSION

The present study was concerned with the retrieval effort hypothesis of the testing effect. This hypothesis states that the more effortful retrieval is, the more beneficial testing as relearning opportunity will be. In the present study participants studied a text and after this study phase, participants were randomly assigned to one of three relearning conditions, a restudy condition, an MC condition, a GR-MC condition. These three conditions were compared on the final test performance after a one week retention interval. The present study indicated that the response generate test-taking strategy was indeed perceived as more effortful than restudying or a direct choice MC strategy, but this more effortful way of testing did not result in higher final test scores. Although the two testing conditions did improve over time (between the initial and final test), the final test scores of participants in the two testing conditions were not significantly higher than the final test scores of the participants in the restudy condition. In other words, no testing effect was found. Although this was in line with the results from the study by Sensenig (2010), we had expected different results on basis of studies showing that more retrieval effort enhances subsequent memory performance (e.g., Pyc & Rawson, 2009).

The testing effect has been demonstrated repeatedly and is therefore considered to be a robust effect. However, in the present study we could not establish a testing effect. A few possible explanations of the unexpected results will be pointed out.

First of all, in the present study the restudy phase differed from the initial learning phase while in other testing studies, restudying resembles initial learning. In the present study, participants had to restudy 18 facts extracted from the original text from the initial learning phase. Perhaps, if the participants had to restudy the original text again, the difference between restudying and testing would have become more apparent.

Second, the retrieval effort hypothesis of the testing effect predicts a beneficial effect from testing if retrieval effort is high as well as successful. It seems that in our study, (retrieval) effort was perceived as rather high in the GR-MC condition, but in this condition participants were not so successful in generating responses. On the other hand it might have been too easy to choose the correct answer from the presented alternatives, given the rather high scores on the MC-questions in both testing conditions. Retrieval practice after being presented with the alternatives may therefore have been successful but not very effortful. The fact that there was no time delay between studying the text and retrieval practice might have contributed to the relative easiness of the MC testing and could have made the MC condition (and the MC-part in the GR-MC condition) comparable to a restudy condition. For example, Pyc and Rawson (2009) compared a time lag between study and test of approximately one minute with a time lag of approximately six minutes and concluded that a time lag of six minutes resulted in a bigger testing effect than a time lag of only one minute. In our study participants had 25 minutes to study the text. Therefore the time lag between study and test was not equal for all items. For the items that were studied at the end of the 25 minutes study phase, the time lag might have been too short, while for other items the time delay might have been long enough.

Another factor might be that retrieval practice consisted of only one test session. The testing effect often gets stronger after repeated retrieval as compared to single retrieval (e.g., Roediger & Butler, 2011). This, in combination with the MC test which asks for less effort than for example free recall (e.g., Glover, 1989), might explain why we did not find a beneficial effect from the direct choice MC condition as compared to the restudy condition. This is in line with the first experiment of Kang et al. (2007). They also did not find a beneficial effect from initial MC retrieval on a final MC test compared to a restudy condition.

In contrast, in the GR-MC condition retrieval practice was probably difficult enough, but not successful enough. The participants in the generate MC condition did indicate that they perceived the test as rather effortful with a mean score of 4.24 on a 7 point scale, which was significantly higher than in the two other conditions. However, participants free recalled on average 7.48 targets (out of 18), which is only 41.6% correct during the generation phase.

It is important for a testing effect to occur that retrieval effort is high, but retrieval should at the same time be successful, in order to enhance memory performance. Because participants in the generate MC condition were not quite successful in recalling the targets, but could directly choose between the possible alternatives afterwards, the two different testing conditions might not have been so different after all.

Crocker and Schmitt (1987) also conducted a study with a GR-MC condition and found that this test-taking strategy was beneficial for students with low test-anxiety, but that it was even a bit detrimental with high test anxiety. An alternative interpretation of the results of Crocker and Schmitt (1987) is that the students low in test-anxiety were the better students (e.g., Culler & Holahan, 1980). If the low test-anxiety students in the Crocker and Schmitt (1987) study were indeed the better students, they benefited from this response generate test-taking strategy because they were just better able to successfully recall the correct answer before seeing the MC alternatives. This would endorse our explanation of not finding a testing benefit in the generate MC condition of our study, because in our study the proportion recalled answers in this condition was rather low (e.g., perhaps comparable to low performing students).

To conclude, the present study was designed to investigate the retrieval effort hypothesis of the testing effect, which we approached by varying the test-taking strategies used with a commonly used testing format (i.e., MC). Since we did not find a testing effect, we cannot say whether the retrieval effort hypothesis should be rejected or maintained. Our results do seem to connect to earlier research that suggests that retrieval practice should not only be effortful, but that retrieval practice should also be successful in order for it to be beneficial to future memory performance (e.g., Carpenter, 2009). As a consequence we suggest that future research could investigate the GR-MC test-taking strategy in light of individual differences. If retrieval practice is only beneficial if it is effortful and successful, this strategy will probably only be beneficial for high performing students and not for low performing students.

7 | Summary and Discussion

The studies presented in this dissertation were concerned with long-term memory and the testing effect. Researchers generally agree on the idea that long-term memory can be improved as a result of beneficial encoding strategies (e.g., Craik & Tulving, 1975), however research on the testing effect shows that future memory performance can also benefit from retrieval of information from memory (e.g., Roediger & Karpicke, 2006b). In other words, taking a test is not only useful to assess what people know, but can enhance learning as well (e.g., Roediger & Butler, 2011).

The testing effect is the empirical finding that testing students' memory after an initial learning phase will improve memory performance on a subsequent memory test. The effect holds even when compared to restudying the information and is most often found after a multiday retention interval (e.g., Roediger & Butler, 2011). The testing effect has been studied extensively and seems to be very promising for educational practice. However, despite the large number of studies on this topic, the mechanisms underlying the testing effect are still not completely clear. One of the most cited hypotheses nowadays to explain the testing effect is the elaborative processing hypothesis. This hypothesis states that the processes that occur during retrieval of information are more elaborate than the processes during restudying the information, hence resulting in better memory performance (e.g., Carpenter, 2009). However, recently some studies have been published that strongly challenge this hypothesis (e.g., Karpicke & Smith, 2012) and as a result the discussion on what explains the testing effect is still unresolved.

Moreover, the applications of the testing effect in educational practice have not been fully explored. Because the testing effect seems so promising for educational practice, it is important to explore the various ways in which testing can be used to improve long-term retention of knowledge. The current dissertation contributed to the body of research on these two issues.

The following *research questions* were addressed in this dissertation:

What explains the testing effect and how can we optimize retrieval practice to improve long-term retention and hence optimize the use of it in educational practice? These questions were addressed by studying the elaborative processing hypothesis (chapter 4 and 5) and the retrieval effort hypothesis (chapter 6) of the testing effect. In addition, the role of memory schematization in explanations of the testing effect was studied (chapter 3 and 4). Also some practical boundaries of the testing effect were addressed (chapter 3 and 6). Does testing format (chapter 3) or test-taking strategies (chapter 6) make any difference in the magnitude of the testing effect? Chapters 3 to 6 were all concerned with the retrieval aspect of long-term memory. Chapter 2, on the other hand, focused on the encoding part of long-term memory. If information is not learned well initially, testing will have no beneficial effect, because retrieval will not be successful at an initial test to begin with.

In the remainder of this chapter, the main results of the studies in chapters 2 to 6 will be described and discussed in terms of the theoretical explanations and practical applications of the testing effect. The chapter will end with the main conclusions and suggestions for future research.

SUMMARY OF THE MAIN RESULTS

The study reported in **chapter 2** was conducted within a psychology curriculum where the Progress Test is used as its main assessment tool. The Progress Test is a method to assess long-term retention of curriculum knowledge and was administered three times a year assessing the content of all 16 basic courses of a Dutch psychology curriculum. In addition, students in this curriculum were assessed with a formative test on their basic course knowledge at the end of each course (i.e., after five weeks). The scores on this course test can be considered to represent the level of initial learning at the end of the course. Although favorable effects from testing (e.g., formative course test every five weeks) are to be expected in this curriculum, large differences in long-term retention of knowledge (as measured with the Progress Test) between students existed. The aim of this study was therefore to gain insight into the determinants of long-term retention performance of first-year students in the curriculum at hand. To that end, the relationship between level of initial learning (measured with the formative course tests), prior knowledge, class attendance and individual study time, and Progress Test scores (as a measure of long-term retention of curriculum knowledge), was analyzed. The data showed that level of initial learning played an important role in predicting long term-retention at the end of the first year of the curriculum. Students with higher scores on formative course tests had a more extended knowledge base of psychology at the end of the first year of the curriculum than students with lower levels of initial learning. However, prior knowledge, class attendance and individual study time did not significantly predict knowledge growth. The results of this study concerning level of initial learning are in line with previous research on long-term retention of knowledge. Bahrick and Hall (1991), Conway et al. (1991), and Semb et al. (1993) also found better retention scores for students with higher levels of initial learning. Moreover, they can be interpreted in the light of testing effect studies showing that initial testing should be successful will it be beneficial for long-term retention (e.g., Pyc & Rawson, 2009). That is, students with higher levels of initial learning at course tests were more successful in recalling information on the Progress Test than students with lower levels of initial learning. There are, however, some limitations to this study. For example, we did not control for restudying. Students, who show more knowledge growth than others, could for instance restudy the material

more often than others (Driskell et al., 1992). This point relates to the study time measure used in this study which indicated the amount of time students during the particular courses but it did not assess the amount of time students spent on restudying study material (or summaries) in advance of the upcoming Progress Tests. Furthermore, we do not know what explains the differences in level of initial learning. Prior knowledge and class attendance were positively associated with level of initial learning, but did not predict knowledge growth. Perhaps there are other factors, for instance type of learning strategy, which were not investigated in the study of this chapter that could play a role. Future research will be necessary to address this question. The aim of the study in **chapter 3** was to investigate the separate effects of two different testing formats, being Multiple Choice (MC) and MC-justification questions, on memory awareness and on the long-term retention of knowledge. MC-justification questions are answered by first choosing the correct answer from MC-alternatives and then justifying why the chosen answer is the correct answer. Memory awareness was measured with the remember-know approach proposed by Tulving (1985) where a 'remember awareness' accompanying memory retrieval is an indication of retrieval from episodic memory and a 'know awareness' is considered to be an indication of retrieval from semantic memory. The shift from remembering to knowing over consecutive retrieval episodes has been interpreted as an indication of knowledge schematization (e.g., Herbert & Burt, 2004). It was expected that MC-justification items would have a more beneficial effect on knowledge schematization and long-term retention than MC-items. Participants took four subsequent knowledge tests on curriculum learning material they studied at different retention intervals prior to the first test of the study. At the first and final test, participants reported their accompanying memory awareness when answering the questions of the test. At the two intermediate tests, test-format was manipulated (MC versus MC-justification) and no memory awareness was reported. Although a general improvement in test scores over time was found, the findings from this study did not reveal a remember-to-know shift as had previously been established by Conway et al. (1997) and Herbert and Burt (2001, 2003). This could for example be the result of the fact that the knowledge had already been schematized as in the research methods course condition of the study by Conway et al. (1997) or that 1.5 weeks is too long to have many episodic memories of the encoding event as Conway (2009) suggested. The fact that memory performance was significantly higher at the final test needs to be interpreted with some caution, since there was not included a no-intermediate test condition in this study. Nevertheless it seems fairly reasonable to suggest that the intermediate tests are responsible for this increased test scores, since the rather long retention interval between the pre- and posttest in this study would have been expected to result in a decrease of memory performance as an effect of decay.

The study in **chapter 4** was designed to investigate the elaborative processing hypothesis of the testing effect. A prediction that follows from this hypothesis is that elaborately restudying, like testing, should result in better final memory performance than plain restudying. Participants learned Swahili-Dutch word pairs and were randomly assigned to one of three conditions: self-testing, elaborately restudying or plain restudying the word pairs. The elaborate restudy condition consisted of restudying the word pair with help from a mnemonic aid. For example the mnemonic aid for the word pair *ardhi – grond* (*soil* in English) was: '*ardhi*, sounds like *aarde* (*earth* in English), which is another Dutch word for *grond* (*soil*)'. The experiment ended with a final cued-recall test at a one-week retention interval. In this study a general testing effect was found. At the cued-recall test after one week, memory performance in the self-testing condition was better compared to elaborately restudying, but not compared to plain restudying the material. This indicated that elaborately restudying did not result in comparable memory performance as testing or in better memory compared to plain restudying. The elaborative processing hypothesis as explanation of the testing effect was therefore not supported by this study. However, more research is of course necessary to seriously decline this hypothesis, even more because this study was not flawless. For example, the relearning conditions of this study were manipulated within subjects, and although there were no explicit signs of it, it cannot be ruled out that participants elaborated on the learning material in the plain restudy condition.

The study in **chapter 5** was also designed to investigate the elaborative processing hypothesis of the testing effect. Although several studies have found results in line with the elaborative processing hypothesis (e.g., Carpenter, 2009; Pyc & Rawson, 2009), recently some other studies have shown that elaboration is probably not the explanatory factor (e.g., Karpicke & Smith, 2012, and see also chapter the study described in chapter 4). From a study by Pyc and Rawson (2010) it became clear that elaboration in the form of the use of mediators (e.g., a word, phrase, or concept that links a cue to a target) can foster learning, but that it is difficult for participants to come up with effective mediators. If participants in an elaborate restudy condition are not sufficiently able to come up with an effective mediator to link the cue and the target, the question rises how elaborate this restudy condition actually is. Therefore, the study described in chapter 5 compared two conditions of elaborate restudy, a condition in which participants had to come up with a mediator versus a condition in which they were given a mediator to help them remember a target when studying a word pair. If comparisons between elaborate restudy and testing were influenced by the fact that studying word pairs with a self-generated mnemonic aid is too difficult, one would expect a different effect from self-generated and given mediators. Both elaborate restudy conditions were compared with a testing condition on a cued-recall test at a one-week retention interval. The results from this study showed that the testing condition still outperformed

the restudy condition on a final retention test at a one-week retention interval. Although participants in the given mediator condition on average recalled more mediators at the final test than the participants in the self-generated mediator condition, no differential effect was found on the mean number of targets that was recalled at the final test. It seemed that the given mediator condition was truly more elaborate than the self-generated mediator condition. However, both conditions were outperformed by the testing condition. There should be noted however, that the testing condition was not a 'pure' testing condition (i.e., it was a test-restudy condition). Therefore it cannot be ruled out that the sequence of test-restudy-test has boosted the results in the testing condition. Nevertheless, the results of this study could again indicate, in line with the results from the Karpicke and Smith (2012) study, that the testing effect cannot be explained by the elaborative processing hypothesis.

The study described in **chapter 6** investigated the retrieval effort hypothesis, which was approached not by varying the test format, but the test-taking strategies. The retrieval effort hypothesis states that the more effortful retrieval is, the more beneficial testing as a relearning opportunity will be. Participants studied a text and after this study phase, they were randomly assigned to one of three conditions: a direct choice MC condition (MC condition), a response generate MC (GR-MC) condition, or a restudy condition. In the GR-MC condition participants were asked to recall the right answer, before seeing the alternatives and choosing the correct one. In the MC condition, participants were asked to look directly at the alternatives and to choose the correct answer. In the restudy condition participants were asked to restudy a list of facts selected by the researcher. These three conditions were compared on final test performance after a one-week retention interval. The GR-MC test-taking strategy was indeed perceived as most effortful, but this more effortful way of testing did not result in higher final test scores. Although memory performance in the two testing conditions did improve over time (between the initial and final test), the final test scores of participants in the two testing conditions were not significantly higher than the final test scores of the participants in the restudy condition. These results were unexpected, but could be due to some characteristics of the experiment. For instance, the fact that retrieval practice consisted of only one trial might have undermined the strength of the effect of retrieval practice. In line with this is the fact that retrieval effort in the GR-MC condition was indicated as high but that retrieval practice was not so successful might have influenced the results of the study described in this chapter. According to the retrieval effort hypothesis, retrieval should not only be effortful, but it should also be successful.

In the next section, the main results will be discussed in terms of the theoretical explanations and practical applications of the testing effect.

GENERAL DISCUSSION

One of the most notable findings of the present dissertation was that the elaborative processing hypothesis of the testing effect could not be supported. A prediction following from this hypothesis is that elaborately restudying the material will lead to comparable memory performance as testing. This prediction was investigated but not confirmed in this dissertation. Although quite some studies have found confirmation for the elaborative processing hypothesis (e.g., Carpenter, 2009; Carpenter & DeLosh, 2006), very recent research from Karpicke and Smith (2012) as well as from Karpicke and Blunt (2011) could not endorse it. Results from the studies presented in chapters 4 and 5 were in line with these latter studies and therefore contributed to the idea that the elaborative processing hypothesis of the testing effect might not be tenable.

Another main finding from this dissertation relates to the role of schematization of knowledge in explaining the testing effect. The results from the studies in chapters 3 and 4 indicated that schematization of knowledge does not seem to play a role in explaining the testing effect. Schematization is thought to be the results of different encounters with the to be learned information. Different encounters with the to be learned information are sometimes also seen as some kind of elaboration (e.g., Conway et al. 1997). We therefore investigated the relation between two testing formats differing in the amount of elaboration and knowledge schematization to test the elaborative processing hypothesis. The results did not show a relation between the more elaborate testing format and knowledge schematization and were therefore not in line with the elaborative retrieval hypothesis. Finally, the study presented in chapter 6 also failed to find support for a hypothesis comparable to the elaborative processing hypothesis, that is, the retrieval effort hypothesis.

The results from this dissertation combined with the aforementioned studies by Karpicke and Smith (2012) and Karpicke and Blunt (2011) seem to have strong implications for the research on the explanations of the testing effect. The elaborative processing hypothesis does not seem right. Therefore, the focus should be shifted: instead of trying to understand the elaborate nature of retrieval processes, it seems more interesting to investigate the differences between encoding and retrieval and focus more on what happens during retrieval that is different from encoding. Perhaps it is not so much the elaborate nature of the retrieval process that is responsible for improved memory performance after testing, but the strengthening of the path between cue and target as a result of successful retrieval. When presented with a cue, one uses that cue to decide which of viable candidate targets was the one studied. When distinctive features of a target are linked to a specific cue, the diagnostic value of that cue is high (Nairne, 2002). Retrieval presumably enhances this diagnostic value of the cue, while elaborate restudy does not or to a smaller extent. In future attempts to

explain the testing effect, we could aim at gaining better understanding of the processes and conditions during retrieval practice that lead to high cue diagnosticity (e.g., Karpicke & Schmit, 2012; Nairne, 2002).

The other main research question in this dissertation was not so much concerned with the theoretical explanations of the testing effect, but with the practical use of tests to improve long-term retention and hence optimize its use in educational practice. To answer this research question we investigated the effect of two different testing formats, Multiple Choice (MC) and MC-justification questions, on memory awareness and on long-term retention of knowledge. We also looked into possible different effects of two different test-taking strategies, generate response MC and direct choice MC, on long-term retention. Although we know that testing it self is a powerful tool to improve memory performance (e.g., Karpicke & Roediger, 2006b), from the studies in chapters 3 and 6 we concluded that different test formats (MC or MC-justification) or test-taking strategies (generate response MC or MC) do not have differential effects on long-term retention of knowledge. Next to that the results from chapter 6 do seem to connect to earlier research that suggests that retrieval practice should not only be effortful, but that retrieval practice should also be successful in order to be beneficial for future memory performance (e.g., Carpenter, 2009). As a consequence we conclude that the generate MC test-taking strategy is only beneficial if it is successful. This strategy will therefore probably only be beneficial for high performing students and not for low performing students, since high performing students are likely to perform better at initial tests than low performing students. Educators should therefore be cautious with advising their students in general to test themselves on learned information. For some students that might be very beneficial, but if (as our research in chapter 1 also suggests) initial knowledge is not sufficient no benefits from testing are to be expected. Even worse, students might encounter negative effects in study motivation or performance. Students need to make sure that they study information well enough in that they are able to recall it on an initial test. If that is the case, they can use testing to improve retention of their knowledge and become more successful in their studies. It is also known that it is beneficial to provide correct answer feedback (instead of only 'right' or 'wrong') after retrieval practice and to expand the interval between subsequent tests to improve memory performance to keep retrieval practice effortful (Roediger & Butler, 2011).

Using retrieval practice in education should be promoted when retrieval practice can be effortful, successful, followed by correct answer feedback and when the intervals between retrieval attempts are of increasing length to ensure that the retrieval practice remains effortful. When these guidelines are taken into account, retrieval practice can form a useful tool to enhance learning in educational practice.

CONCLUSION

It can be concluded from the studies presented in this dissertation, that the level of initial knowledge of students plays an important role in long-term retention of that knowledge (Chapter 1). Although testing was found to have a beneficial effect on long- term retention of knowledge (Chapter 4 and 5), the present dissertation could not answer the question what the underlying mechanism of the testing effect is. It did, however, suggest that the elaborative processing hypothesis is not a very plausible explanation of the testing effect. A novel aspect of this dissertation was that it investigated the link between knowledge schematization and the testing effect, however, no support was found for the idea that testing improves memory performance as a consequence of an enhanced level of knowledge schematization. The second goal of this dissertation was to explore various ways in which testing could be used in educational practice to improve long-term retention of students' knowledge. From the studies described in chapters 3 and 6, it it was concluded that different test formats (MC or MC-justification) or test-taking strategies (generate response MC or MC) did not have differential effects on long-term retention of knowledge.

SUGGESTIONS FOR FUTURE RESEARCH

From the studies presented in the current dissertation as well as other research (e.g., Karpicke & Smith, 2012) it can be concluded that the elaborative processing hypothesis cannot account for the beneficial memory effects of testing. Future research should continue developing theories and investigating hypotheses that could account for the testing effect. Even if retrieval is an elaborate process, it seems that it is not this elaborate nature of the retrieval process that explains the beneficial effects from testing. Perhaps the strengthening of the path between cue and target as a result of successful retrieval is responsible for increasing the diagnostic value of a particular cue, which in turn leads to increased memory performance after testing and not after restudying. In the case of elaborately restudying for example, multiple pathways are activated which may increase the chance to correctly recall the target, but also the chance to recall one of the other information parts that are activated during encoding. However, the diagnostic value of the cue may not be increased, because the correct memory path between cue and target has not been confirmed in a retrieval process.

Moscovitch and Craik (1976) concluded something similar from their study investigating the relation between retrieval and different levels of encoding. In their study (experiment 2) participants studied words with encoding questions at different levels of encoding (deep versus shallow) on which they were tested later on with a cued-recall test. The cues were the

encoding questions. Some encoding questions were used for only one word and some for 10 words, varying the uniqueness of the cue-target relationship (unique versus shared). Recall was lower for the shared cues and the detrimental effects were largest for words studied at a deep level of processing. Moscovitch and Craik (1976) concluded from this that encoding processes determine the ceiling of potential memory performance, with deeper encoding leading to better memory performance. The extent to which the set ceiling can be reached is dependent on the retrieval conditions, with more unique cues leading to better memory performance. These ideas of Moscovitch and Craik can be used to investigate the testing effect. If cue diagnosticity or distinctiveness plays a role in the testing effect, the testing effect would be more profound with non-distinctive cue-target relations than with distinctive cue-target relations. Larger differences in final test performance between testing and restudying conditions would be expected after testing/restudying with non-distinct cue-target relations than after testing/restudying with distinct cue-target relations. This would be expected since the restudy condition with distinct cue-target relations would benefit from the cue distinctiveness while the restudy condition with non-distinct cue-target relations would not. In addition, restudy conditions that differ in elaboration could also interact with this, just as it did in the Moscovitch and Craik (1976) study, because more elaboration would lead to even more potential targets belonging to a particular cue (i.e. lower cue distinctiveness).

Another suggestion for future research is taking the next step towards investigating individual differences that explain the (magnitude of) the testing effect. A study by Bouwmeester and Verkoeijen (2011) using latent class analysis showed that three different groups could be distinguished in terms of magnitude of the testing effect. Participants studied word lists which were either restudied or tested afterwards. After one week, a final recognition tests with targets and distracters was administered. The latent class analysis showed that one group of participants did not benefit from testing at all, while the other two groups benefited both, but one group significantly more than the other. It was suggested that this difference could be explained by the way in which these participants processed the semantic meaning of the target words, since the groups differed in the amount of falsely recognized distracters that were semantically related to the target words. From chapters 1 and 6 of this dissertation, the suggestion arose that individual differences in levels of initial knowledge could influence the magnitude of the testing effect. When we discover the individual differences that interact with differences in the magnitude of the testing effect, the underlying explanations and applications of the testing effect might become clearer, and guidelines for educational practice might become more specific.

References

REFERENCES

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. doi:10.1002/acp.1391

Alexander, P. A., Schallert, D. L., & Hare, V. C. (1991). Coming to terms: How researchers in learning and literacy talk about knowledge. *Review of Educational Research, 61*, 315-343. doi:10.3102/00346543061003315

Anderson, J. R. (1976). *Language, memory, and thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior, 22*, 261-295.

Ashcraft, M. H., & Radvansky, G. A. (2010). *Cognition* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence, J. T. Spence (Eds.), *The Psychology of Learning and Motivation: Advances in Research and Theory, Volume 2.* (pp. 89-195). NY: Academic Press

Baddeley, A. D., & Hitch, G. (1974). Working memory. In Bower, G. H. (Ed.), *The Psychology of Learning and Motivation.* (pp. 47-89). NY: Academy Press.

Bahrick, H. P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology: General, 113*, 1-29. doi:10.1037/0096-3445.113.1.1

Bahrick, H. P., & Hall, L. K. (1991). Lifetime maintenance of high school mathematics content. *Journal of Experimental Psychology: General, 120*, 20-33. doi: 10.1037 /0096-3445.120.1.20

Barrows, H. S. (1988). *The tutorial process.* Springfield: Southern Illinois University School of Medicine.

Bath, D. M. B. (2004). Rembering, knowing and schematisation: Theoretical and practical perspectives. In S. P. Shohov (Ed.), *Advances in Psychological Research* (pp. 23-45). New York: Nova Science Publishers, inc.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy.* New York: Academic Press.

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

Blake, J. M., Norman, G. R., Keane, D. R., Mueller, B., Cunnington, J. & Didyk, N. (1996). Introducing progress testing in McMaster University's problem-based medical curriculum: Psychometric properties and effect on learning. *Academic Medicine, 71*, 1002-1007.

Bouwmeester, S., & Verkoeijen, P. P. J. L. (2011). Why do some children benefit more from testing than others? Gist trace processing to explain the testing effect. *Journal of Memory and Language, 65*, 32-41. doi:10.1016/j.jml.2011.02.005

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*, 717-726. doi:10.1016/S0022-5371(72)80006-9

Busato, V. V., Prins, F. J., Elshout, J., & Hamaker, C. (2000). Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. *Personality and Individual Difference, 29*, 1057-1068. doi:10.1016/S0191-8869(99)00253-6

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*, 514-527. doi:10.1080/09541440701326097

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. doi:10.1037/a0017021

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the
testing effect. *Journal of Experimental Psychology: Language, Memory, and Cognition, 37*, 1547-1552. doi:10.1037/a0024140

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*, 268-276. doi:10.3758/BF03193405

Carpenter, S. K., Pashler, H., (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14*, 474-478. doi: 10 .3758/BF03194092

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*, 438-448. doi: 10 .3758 /MC .36.2.438

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*, 633-642. doi:10.3758/BF03202713

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 431-437. doi:10.1037/0278-7393.33.2.431

Conway, M. A. (2009). Episodic memories. *Neuropsychologia, 47*, 3205-2313. doi:10 .1016/j.neuropsychologia.2009.02.003

Conway, M. A., Cohen, G. C., & Stanhope, N. (1991). On the very long-term retention of knowledge acquired through formal education: Twelve years of cognitive psychology. *Journal of Experimental Psychology: General, 121*, 382-384. doi: 10. 1037/0096-3445.121.3.382

Conway, M. A., Cohen, C. G., & Stanhope, N. (1992). Very long-term memory for knowledge acquired at school and university. *Applied Cognitive Psychology, 6*, 467-482. doi:10.1002/acp.2350060603

Conway, M. A., Gardiner, J. M., Perfect, T. J., Anderson, S. J., & Cohen, G. M. (1997). Changes in memory awareness during learning: The acquisition of knowledge by psychology undergraduates. *Journal of Experimental Psychology: General, 126*(4), 393-413. doi:10.1037/0096-3445.126.4.393

Coppens, L. C., Verkoeijen, P. P. J. L., & Rikers, R. M. J. P. (2011). Learning Adinkra symbols: The effect of testing. *Journal of Cognitive Psychology, 23*, 351-357. doi: 10. 1080/20445911.2011.507188

Craik, F. I. M., & Lockhart, R. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-84. doi:10.1016/S0022-5371(72)80001-X

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268-294. doi:10.1037/0096-3445.104.3.268

Crocker, L., & Schmitt, A. (1987). Improving multiple-choice test performance for examinees with different levels of test anxiety. *The Journal of Experimental Education, 55*, 201-205.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14*, 215-235. doi:10.1002/(SICI)1099-0720(200005/06)14:3<215::AID-ACP640>3.0.CO;2-1

Culler, R. E., & Holahan, C. J. (1980). Test anxiety and academic performance: The effects of study-related behaviors. *Journal of Educational Psychology, 72*, 16-20. doi:10.1037/0022-0663.72.1.16

Darley, C. F., & Murdock, B. B. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology, 19*, 66-73. doi:10.1037/h0031836

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation: Advances in Research and Theory, 53*, 63-148. doi:10.1016/S0079-7421(10)53003-2

Dempster, F. N. (1988). The spacing effect. A case study in the failure to apply the results of psychological research. *American Psychologist, 43*, 627-634. doi: 10 .1037/0003-066X.43.8.627

Dewhurst, S. A., Conway, M. A., & Brandt, K. R. (2009). Tracking the R-to-K shift: Changes in memory awareness across repeated tests. *Applied Cognitive Psychology, 23*, 849-858. doi:10.1002/acp.1517.

Driskell, J.E., Willis, R. P., & Copper, C. (1992). Effect of overlearning on retention. *Journal of Applied Psychology, 77*, 615-622. doi:10.1037/0021-9010.77.5.615

Dudukovic, N. M., & Knowlton, B. J. (2006). Remember-Know judgments and retrieval of contextual details. *Acta Psychologica, 122*, 160-173. doi:10.1016 /j.actpsy .2005.11.002

Eysenck, M. W. (2012). *Fundamentals of Cognition*. NY: Psychology Press.

Fellenz, M. R. (2004). Using assessment to support higher level learning: The multiple choice item development assignment. *Assessment & Evaluation in Higher Education, 29*, 703-719. doi:10.1080/02602930420002 27245.

Field, A. (2005). *Discovering Statistics using SPSS: And sex and drugs and rock 'n' roll (third edition)*. London: Sage.

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition, 16*, 309-313. doi:10.3758/BF03197041

Gardiner, J. M. (2001). Episodic memory and autonoetic consciousness: A first-person approach. *Philosophical Transactions of the Royal Society B, 356*, 1351-1361. doi:10.1098/rstb.2001.0955

Gardiner, J. M., Gawlik, B., & Richardson-Klavehn, A. (1994). Maintenance rehearsal affects knowing, not remembering; elaborative rehearsal affects remembering, not knowing. *Psychonomic Bulletin & Review, 1*, 107-110. doi:10.3758 /BF03200764

Gardiner, J. M., & Java, R. I. (1993). Recognition memory and awareness: An experiental approach. *European Journal of Cognitive Psychology, 5*, 337-346. doi:10.1080/09541449308520122

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (1998). Experiences of remembering, knowing, and guessing. *Consciousness and Cognition, 7*, 1-26. doi: 10.1006/ccog.1997.0321

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6* (40).

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*, 392-399. doi:10.1037/0022-0663.81.3.392

Gunn, K. P. (1993). A correlation between attendance and grades in a first-year psychology class. *Canadian Psychology, 34*, 201-202. doi:10.1037/h0078770

Herbert, D. M. B. (1999, November). *What do students learn from lectures? The role of episodic memory in early learning*. Paper presented at the combined Australian Association for Research in Education – New Zealand Association for Research in Education Conference, Melbourne, Australia. Retrieved from http://www.aare.edu.au/99pap/her99084.htm

Herbert, D. M. B., & Burt, J. S. (2001). Memory awareness and schematisation: Learning in the university context. *Applied Cognitive Psychology, 15*, 617-637. doi:10 .1002/acp.729

Herbert, D. M. B., & Burt, J. S. (2003). The effects of different review opportunities on schematisation of knowledge. *Learning and Instruction, 13*, 73-92. doi:10.1016 /S0959-4752(01)00038-X

Herbert, D. M. B., & Burt, J. S. (2004). What do students remember? Episodic memory and the development of schematization. *Applied Cognitive Psychology, 18*, 77-88. doi:10.1002/acp.947

Hmelo-Silver, C. E. (2004). Problem-Based Learning: What and how do students learn? *Educational Psychological Review, 16*(3), 235-266. doi:10.1023 /B:EDPR .0000034022.16470.f3

Hmelo-Silver, C. E., & Barrows, H. S. (2006). Goals and strategies of a problem-based learning facilitator. *The Interdisciplinary Journal of Problem-based Learning, 1*, 21-39. doi:10.7771/1541-5015.1004

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. doi:10.1016/S0022-5371(71)80029-4

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513-541. doi: 10.1016/0749-596X(91)90025-F

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101*, 621-629. doi: 10.1037/a0015183

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*, 528-558. doi:10.1080/09541440601056620

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*, 772. doi: 10.1126/science.1199327

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*, 151-162. doi:10 .1016/j.jml.2006.09.004

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

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67*, 17-29. doi:10 .1016/j.jml.2012.02.004

Kember, D., Jamieson, Q. W., Pomfret, M., & Wong, E. E. T. (1995). Learning approaches, study time and academic performance. *Higher Education, 29*, 329-343. doi:10.1007/BF01384497

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16*, 421-437. doi:10.1177/001316445601600401

Loyens, S. M. M., Kirschner, P. A. & Paas, F. (2011). Problem-based learning. In K.R. Harris, S. Graham & T. Urdan (Eds.), *APA educational psychology handbook: Vol. 3. Application to learning and teaching* (pp. 403-425). Washington DC: American Psychological Association.

Marburger, D. R. (2001). Absenteeism and undergraduate exam performance. *Journal of Economic Education* 32, 99-109. doi:10.1080/00220480109595176

McCabe, D. P., Geraci, L., Boman, J. K., Sensenig, A. E., & Rhodes, M. G. (2011). On the validity of remember-know judgments: Evidence from think aloud protocols. *Consciousness and Cognition, 20*, 1625-1633. doi:10.1016/j.concog.2011.08.012

McClain, L. (1983). Behavior during examinations: A comparison of "A", "C", and "F" students. *Teaching of Psychology, 10*, 69-71. doi:10.1207/s15328023top1002_2

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103*, 399-414. doi:10.1037/a0021782

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*, 494-513. doi:10.1080/09541440701326154

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16*, 192-201. doi:10.1016/0361 -476X(91)90037-L

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. doi:10.3758/BF03202614

McDermott, K. B. (2006). Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition, 34*, 261-267. doi:10.3758/BF03193404

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. doi:10.1016/S0022-5371(77)80016-9

Moscovitch, M., & Craik, F. I. M. (1976). Depth of processing, retrieval cues, and uniqueness of encoding as factors in recall. *Journal of Verbal Learning and Verbal Behavior, 15*, 447-458. doi:10.1016/S0022-5371(76)90040-2

Moust, J. H. C. (1993). *De rol van tutoren in probleemgestuurd onderwijs* [the role of tutors in problem-based learning]. PhD diss., Rijksuniversiteit Limburg, Maastricht, The Netherlands.

Nairne, J. S. (2002). The myth of the encoding-retrieval match. *Memory, 10*, 389-395. doi:10.1080/09658210244000216

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory, 2*, 325-335. doi:10 .1080/09658219408258951

Pintrich, P.R. (1999). The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research, 31*, 459-470. doi: 10 .1016/S0883-0355(99)00015-4

Plant, E.A., Ericsson, K. A., Hill, L., & Asberg, K. (2005). Why study time does not predict grade point average across college students: Implications of deliberate practice for academic performance. *Contemporary Educational Psychology, 30*, 96-116. doi:10.1016/j.cedpsych.2004.06.001

Pressley, M., Levin, J. R., & Delaney, H. D. (1982). The mnemonic keyword method. *Review of Educational Research, 52*, 61-91. doi:10.3102/00346543052001061

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. doi:10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. Science, 330, 335. doi:10.1123/science.1191465

Raaijmakers, J. G. W., & Shiffrin, R. M. (1992). Models for recall and recognition. *Annual Reviews of Psychology, 43*, 205-234. doi:10. 1146 /annurev.ps.43 .020192.001225

Recht, D. R., & Leslie, L. (1988). Effect of prior knowledge on good and poor readers' memory of text. *Journal of Educational Psychology, 80*, 16-20. doi:10.1037 /0022-0663.80.1.16

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, 17, 382-395. doi:10 .1037 /a00026252

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Science*, 15, 20-27. doi: 10.1016 /j.tics.2010.09.003

Roediger, H. L., Gallo, D. A., & Geraci, L. (2002). Processing approaches to cognition: The impetus of the levels-of-processing framework. *Memory, 10*, 319-332. doi: 10.1080/09658210224000144

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning. Taking memory tests improves long-term retention. *Psychological Science, 17(3)*, 249-255. doi:10 .1111/j.1745-6916.2006.00012.x

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory. Basic research and implications for educational practice. *Perspectives on Psychological Science, 1(3)*, 181-210. doi:10.1111/j.1745-6916.2006.00012.x

Rogosa, D., Brandt, D.,& Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*, 726-748. doi:10.1037/0033-2909.92.3.726

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement, 20*, 335-343. doi:10.1111/j.1745-3984.1983.tb00211.x

Romer, D. (1993). Do students go to class? Should they? *Journal of Economic Perspectives, 7*, 167-174. doi:10.1257/jep.7.3.167

Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory & Cognition*, 11, 641-650. doi:10.3758/BF03198289

Schaap, L., Schmidt, H. G., & Verkoeijen, P. P. J. L. (2012). Assessing knowledge growth in a psychology curriculum: Which students improve most? *Assessment and Evaluation in Higher Education, 37(7)*, 875-887. doi: 10.1080/02602938. 2011.581747

Schmidt, H. G. (1993). Foundations of problem-based learning: Some explanatory notes. *Medical Education*, 27, 422-432. doi:10.1111/j.1365-2923.1993.tb00296.x

Schuman, H., Walsh, E., Olson, C, & Etheridge, B. (1985). Effort and reward: The assumption that college grades are affected by quantity of study. *Social Forces, 63*, 945-966. doi:10.1093/sf/63.4.945

Scouller, K.M., & Prosser, M. (1994). Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education*, 69, 267-279. doi: 10 .1080/03075079412331381870

Semb, G.B., Ellis, J. A., & Araujo, J. (1993). Long-term memory for knowledge learned in school. *Journal of Educational Psychology, 85*, 305-316. doi:10.1037/0022-0663.85.2.305

Sensenig, A. E. (2010). *Multiple choice testing and the retrieval hypothesis of the testing effect* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database. (Accession Order No. AAT 3428630).

Shapiro, A.M. (2004). How including prior knowledge as a subject variable may change outcomes of learning research. *American Educational Research Journal, 41*, 159-189. doi:10.3102/00028312041001159

Spitzer, H. F. (1939). Studies in retention. *The Journal of Educational Psychology, 30*, 641-656. doi:10.1037/h0063404

St. Clair, K.L. (1999). A case against compulsory class attendance policies in higher education. *Innovative Higher Education, 29*, 171-180. doi:10.1023/A :1022942400812

Tan, E. S., Imbos, T., & Does, R. J. J. M. (1994). A distribution-free approach for comparing growth of knowledge. *Journal of Educational Measurement, 31*, 51-65. doi:10.1111/j.1745-3984.1994.tb00434.x

Thompson, R.A., & Zamboanga, B. L. (2004). Academic aptitude and prior knowledge as predictors of student achievement in introduction to psychology. *Journal of Educational Psychology*, 96: 778-784. doi:10.1037/0022-0663.96.4.778

Tulving, E. (1967). The effects of presentation and recall of material in free recall learning. *Journal of Verbal Learning and Verbal Behavior, 6*, 175-184. doi:10 .1016/S0022-5371(67)80092-6

Tulving, E. (1985). Memory and Consciousness. *Canadian Psychology, 26*, 1-12. doi:10 .1037/h0080017

Tulving, E. & Thompson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*, 352-373. doi:10.1037/h0020071

Van Berkel, H. J. M., & Schmidt, H. G. (2000). Motivation to commit oneself as a determinant of achievement in problem-based learning. *Higher Education, 40*, 231-242. doi:10.1023/A:1004022116365

Van Blerkom, M. (1996). Academic perseverance, class attendance, and performance in the college classroom. Paper presented at the annual meeting of the American Psychological Association, August 9-13, in Toronto, Canada.

Van Diest, R., Van Dalen, J., Bak, M., Schruers, K., Van der Vleuten, C., Muijtjens, A., & Scherpbier, A. (2004). Growth of knowledge in psychiatry and behavioural sciences in a problem-based learning curriculum. *Medical Education, 38*, 1295-1301. doi:10.1111/j.1365-2929.2004.02022.x

Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist, 43*, 1-11. doi: 10.1080 /00461520701756248

Van Leeuwen, Y. D., Pollemans, M. C., Mol, S. S. L., & Eekhof, J. A. H. (1995). The Dutch knowledge test for general practice. *The European Journal of General Practice, 1*, 113-117.

Van der Vleuten, C. P. M., Verwijnen, G. M., & Wijnen, W. H. F. W. (1996). Fifteen years of experience with progress testing in a PBL-curriculum. *Medical Teacher, 18*, 103-109.

Verhoeven, B. H., Verwijnen, G. M., Scherpbier, A. J. J. A., & Van der Vleuten, C. P. M. (2002). Growth of medical knowledge. *Medical Education, 36*, 711-717. doi:10 .1046/j.1365-2923.2002.01268.x

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*, 571-580. doi:10.1080 /09658210244000414

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441–517. doi:10.1006 /jmla .2002.2864

Samenvatting en discussie

SAMENVATTING EN DISCUSSIE

De studies gepresenteerd in dit proefschrift hebben betrekking op het langetermijngeheugen en het *testing effect*. Hoewel onderzoekers het over het algemeen eens zijn dat het langetermijngeheugen kan worden versterkt als een gevolg van gunstige strategieën voor encodering (bijv. Craik & Tulving, 1975), laat onderzoek naar het *testing effect* zien dat geheugenprestaties op de lange termijn voor bepaalde informatie ook verbeterd kunnen worden door het ophalen van die informatie uit het geheugen (bijv. Roediger & Karpicke, 2006b). Anders gezegd zijn toetsen die het ophalen van informatie uit het geheugen vereisen niet alleen bruikbaar voor het beoordelen wat mensen hebben geleerd, maar kunnen ze ook worden ingezet om leren te bevorderen (bijv. Roediger & Butler, 2011).

Het *testing effect* betreft de empirische bevinding dat het testen van het geheugen van mensen na een initiële leerfase, hun geheugen voor het geleerde in die leerfase op een latere test zal verbeteren. Dit effect houdt zelfs stand als het wordt vergeleken met het herbestuderen van de informatie en wordt meestal gevonden na een retentie interval van enkele dagen (bijv. Roediger & Butler, 2011). Het *testing effect* is uitgebreid bestudeerd en lijkt veelbelovend voor de onderwijspraktijk omdat toetsen als effectief leermiddel kunnen worden ingezet. Echter, ondanks het grote aantal studies naar dit fenomeen, zijn de mechanismen die dit effect kunnen verklaren nog niet geheel duidelijk. Een van de meest genoemde hypotheses is de elaboratieve verwerkingshypothese. Deze hypothese stelt dat de processen die plaatsvinden tijdens het ophalen van informatie uit het geheugen meer bewerking (elaboratie) behelzen dan processen die actief zijn tijdens herbestuderen van die informatie. Door elaboratie wordt de kans op vergeten kleiner en om die reden zou testen volgens de elaboratieve verwerkingshypothese leiden tot een betere geheugenprestatie dan herbestuderen (bijv. Carpenter, 2009). Echter, zeer recent zijn enkele studies gepubliceerd die deze hypothese in twijfel trekken (bijv. Karpicke & Smith, 2012) met als gevolg dat de discussie rondom de verklaringen voor het *testing effect* nog steeds actueel is. Bovendien zijn de toepassingen van het *testing effect* in de onderwijspraktijk nog niet uitputtend onderzocht. Aangezien het *testing effect* veelbelovend lijkt te zijn voor de onderwijspraktijk, is het van groot belang de verscheidene manieren te onderzoeken waarop het effect kan worden gebruikt om het langetermijngeheugen voor geleerde informatie te verbeteren. Het huidige proefschrift heeft een bijdrage geleverd aan beide bovengenoemde onderwerpen van onderzoek.

De volgende *onderzoeksvragen* werden gesteld in het huidige proefschrift: wat verklaart het *testing effect*? Hoe kunnen we het ophalen van informatie uit het geheugen het beste gebruiken om het langetermijngeheugen te verbeteren met als gevolg ook de toepassing ervan in het onderwijs te optimaliseren? Deze vragen werden beantwoord door de *elaborative retrieval* (elaboratieve vewerkings-) hypothese (hoofdstuk 4 en 5) en de *retrieval effort* (inspanning

als gevolg van het ophalen van informatie uit het geheugen) hypothese (hoofdstuk 6) van het *testing effect* te bestuderen. Bovendien werd de rol van schematisatie van kennis in het geheugen in de verklaringen voor het *testing effect* bekeken (hoofdstuk 3 en 4). Ook werd een aantal potentieel praktische grenzen aan het *testing effect* onder de loep genomen. Varieert bijvoorbeeld de grootte van het *testing effect* met de gebruikte toetsvorm (hoofdstuk 3) of testafnamestrategie (hoofdstuk 6)? Hoofdstuk 2 had als onderwerp factoren die samenhangen met de opslag van informatie in het geheugen. Als informatie in eerste instantie al niet goed wordt opgeslagen in het geheugen zal er geen of nauwelijks informatie op te halen zijn uit het geheugen, en zal testen niet kunnen resulteren in een positief effect op de geheugenprestatie. De hoofdstukken 3 tot en met 6 hielden zich allemaal bezig met het ophalen van kennis vanuit het langetermijngeheugen.

In het vervolg van het huidige hoofdstuk zullen de belangrijkste resultaten van de hoofdstukken 2 tot en met 6 worden beschreven en bediscussieerd in termen van theoretische verklaringen van het *testing effect* en de praktische toepassingen ervan. Het hoofdstuk zal eindigen met de belangrijkste conclusies.

Samenvatting van de belangrijkste resultaten

De studie beschreven in hoofdstuk 2 is uitgevoerd in een psychologicurriculum waarin de Voortgangstoets (VGT) werd gebruikt als belangrijkste toetsinstrument. De VGT is een toets gericht op het meten van langetermijnretentie van de kennis behorende bij (een groot gedeelte van) een curriculum. De VGT werd in het bestudeerde curriculum drie keer per jaar afgenomen (VGT1, VGT2, VGT3), om studenten te testen op kennis van de inhoud van de 16 basisvakken ('blokken') van dit curriculum. Daarnaast werden de studenten in dit curriculum afzonderlijk getoetst op hun kennis van alle 16 basisvakken middels een formatieve bloktoets aan het einde van ieder blok (i.e. na iedere vijf weken). De scores op deze bloktoetsen kunnen worden beschouwd als een representatie van het niveau van initieel leren aan het einde van een blok. Hoewel er in dit curriculum gunstige effecten te verwachten zijn als gevolg van vele tests die worden afgenomen (bijv. iedere vijf weken een bloktoets), bestonden er grote verschillen tussen studenten met betrekking tot hun langetermijnkennis (zoals gemeten met de VGT; verschilscores VGT3 en VGT1 gaven een indicatie van kennisgroei). Het doel van deze studie was daarom inzicht te verkrijgen in de factoren die samenhangen met langetermijnretentie van de eerstejaars studenten in dit curriculum. Om dit te onderzoeken werden de relaties tussen het niveau van initieel leren (zoals gemeten met de formatieve bloktoets), de voorkennis, de aanwezigheid tijdens onderwijsbijeenkomsten en de individuele studietijd, en de VGT scores (als een maat voor langetermijnretentie van curriculum kennis) geanalyseerd. De data lieten zien dat initiële kennis een belangrijke voorspeller was van de langetermijnretentie van kennis aan het einde van het eerste jaar

van het curriculum. Studenten met hoge scores op de formatieve bloktoetsen, hadden een uitgebreidere kennisbasis aan het einde van het eerste studiejaar, dan studenten met lagere niveaus van initiële kennis. Echter, voorkennis, aanwezigheid en individuele studietijd hadden geen significant voorspellende waarde voor kennisgroei. De resultaten van deze studie met betrekking tot het niveau van initiële kennis zijn in lijn met eerder onderzoek naar langetermijnretentie van kennis. Zowel Bahrick en Hall (1991), Conway et al. (1991) als Semb et al. (1993) vonden betere langetermijnretentie scores voor studenten met hogere niveaus van initieel leren. Bovendien kunnen deze resultaten geïnterpreteerd worden in het licht van *testing efffect* studies die laten zien dat het initieel ophalen van informatie uit het geheugen succesvol moet zijn, wil het een gunstig effect hebben op het langetermijngeheugen voor die informatie (bijv. Pyc & Rawson, 2009). Met andere woorden, studenten die hogere niveaus van initiële kennis op de bloktoets lieten zien (beter in staat waren tot het initieel ophalen van informatie), waren meer succesvol op de VGT dan studenten die lagere niveaus van initieel leren lieten zien. Het doel van de studie in hoofdstuk 3 was het onderzoeken van de afzonderlijke effecten van twee soorten toetsvormen, te weten meerkeuze vragen en meerkeuze verantwoordingsvragen, op *memory awareness* (besef van herinneren) en langetermijnretentie van kennis. Meerkeuze verantwoordingsvragen zijn vragen die dienen te worden beantwoord door eerst een van de antwoordalternatieven te kiezen en vervolgens uit te leggen waarom dat antwoordalternatief het juiste is. *Memory awareness* werd gemeten met de zogenaamde *remember-know* procedure van Tulving (1985). In deze procedure wordt een '*remember*' respons tijdens het ophalen uit het geheugen beschouwd als een indicatie voor het ophalen van informatie uit het episodische geheugen en een '*know*' respons als een indicatie voor ophalen van informatie uit het semantische geheugen. De verschuiving van *remember* responsen naar *know* responsen over tijd wordt beschouwd als een indicatie van kennisschematisatie (bijv. Herbert & Burt, 2004). De verwachting was dat meerkeuze verantwoordingsvragen een gunstiger effect zouden hebben op kennisschematisatie en derhalve langetermijnretentie, dan reguliere meerkeuzevragen als gevolg van een meer elaboratieve verwerking bij de meerkeuze verantwoordingsvragen. Bij de deelnemers in deze studie werden vier achtereenvolgende toetsen afgenomen. De toetsen hadden betrekking op leermateriaal dat de deelnemers hadden bestudeerd in het kader van hun studie nog voordat zij deelnamen aan de studie. Op de eerste en vierde toets rapporteerden de deelnemers hun *memory awareness* die samenging met het ophalen van hun kennis voor iedere afzonderlijke toetsvraag. Op de tweede en derde (tussenliggende) tests werd de toetsvorm gemanipuleerd (meerkeuze of meerkeuze verantwoordingsvragen) en rapporteerden de deelnemers geen *memory awareness*. Hoewel er tussen de eerste en de vierde toets een verbetering in toetsscores werd waargenomen, verschilde dit niet voor toetsvorm. Daarnaast lieten de resultaten

ook geen verschuiving in *memory awareness* zien, hoewel die in vergelijkbare studies wel werd vastgesteld (bijv. Conway et al. 1997; Herbert & Burt, 2001, 2003).

De studie in hoofdstuk 4 was ontworpen om de *elaborative processing* hypothese van het testing effect te toetsen. Een voorspelling die volgt uit deze hypothese luidt dat elaboratief herbestuderen, net als testen, zou moeten resulteren in betere latere geheugenprestaties dan eenvoudig herbestuderen. De deelnemers in deze studie leerden Swahili-Nederlandse woordparen en werden vervolgens willekeuring toegewezen aan een van drie condities: de zelf-testconditie, de elaboratieve herstudieconditie, of de gewone herstudieconditie. In de elaboratieve herstudieconditie bestudeerden de deelnemers opnieuw de eerder geleerde woordparen met behulp van een ezelsbrug. Bijvoorbeeld het woordpaar *ardhi – grond* had als ezelsbrug: ardhi klinkt als aarde en aarde is een ander woord voor grond. Het experiment eindigde met een *cued-recall* test na een retentie-interval van één week. In deze studie werd een algemeen positief effect van testen gevonden. Op de eindtest was de gemiddelde score in de zelf-testconditie beter dan in de elaboratieve herstudieconditie, maar gelijk aan de gewone herstudieconditie. Dit impliceerde dat elaboratief herbestuderen niet resulteerde in een vergelijkbare geheugenprestatie als zelf-testen noch in een betere geheugenprestatie dan gewoon herbestuderen. De *elaborative processing* hypothese kon derhalve niet bevestigd worden.

Het onderzoek in hoofdstuk 5 richtte zich ook op het toetsen van de *elaborative processing* hypothese. Hoewel de resultaten van meerdere studies deze hypothese ondersteunen (bijv. Carpenter, 2009; Pyc & Rawson, 2009), zijn er recent studies verschenen die aantonen dat elaboratie waarschijnlijk niet de factor is die het *testing effect* verklaart (bijv. Karpicke & Smith, 2012, en zie ook hoofdstuk vier van dit proefschrift).

Uit een studie van Pyc en Rawson (2010) kwam naar voren dat elaboratie in de vorm van een *mediator* bevorderlijk kan zijn voor leren, maar dat het moeilijk is voor deelnemers om een effectieve *mediator* te bedenken. Een *mediator* kan bijvoorbeeld een woord, zinsdeel of concept zijn dat de twee delen van een woordpaar met elkaar verbindt en is vergelijkbaar met de ezelsburggen uit hoofdstuk 4 van dit proefschrift. Een voorbeeld voor een *mediator* bij het woordpaar *ardhi – grond* is 'ardhi klinkt als aarde en aarde is een ander woord voor grond'. In de studie van Pyc en Rawson (2010) moesten deelnemers zelf *mediators* verzinnen. Het feit dat deelnemers in maximaal 51% van de gevallen een *mediator* wisten te herinneren kan worden gezien als een indicatie dat het zelf verzinnen van een *mediator* een lastige opgave is. Als deelnemers in een elaboratieve herstudieconditie niet goed in staat zijn om elaboratief te studeren, omdat ze geen goede *mediator* weten te bedenken, dan rijst de vraag hoe elaboratief deze herstudieconditie daadwerkelijk is. Om dit te onderzoeken, is in de studie in hoofdstuk 5 de vergelijking gemaakt tussen een elaboratieve herstudieconditie waarin deelnemers zelf een *mediator* moesten verzinnen en een elaboratieve herstudieconditie

waarin deelnemers een effectieve *mediator* kregen aangereikt om hen te helpen bij het bestuderen van de woordparen. Beide herstudiecondities werden vergeleken met een testconditie op een *cued recall* toets die werd afgenomen na een retentie-interval van een week. Als de vergelijking tussen een testconditie en een elaboratieve herstudieconditie wordt beïnvloed door het feit dat herbestuderen met een zelf gegenereerde *mediator* te moeilijk is, zou dit de resultaten kunnen beïnvloeden. In dat geval zal er een kleiner verschil in scores zijn wanneer een elaboratieve herstudieconditie met aangereikte *mediators* en een testconditie worden vergeleken dan wanneer een elaboratieve herstudieconditie met zelfgegenereerde *mediators* en een testconditie worden vergeleken.

De resultaten van de studie beschreven in hoofdstuk 5 lieten zien dat de testconditie resulteerde in betere gemiddelde scores op de eindtest dan beide elaboratieve herstudiecondities. Hoewel deelnemers in de aangereikte *mediator*conditie over het algemeen meer *mediators* wisten op te halen uit hun geheugen dan de deelnemers in de zelf-gegenereerde *mediator*conditie, waren de scores op de eindtest in beide herstudiecondities vergelijkbaar. Het lijkt er dus op dat de aangereikte *mediator*conditie wel meer elaboratief was dan de zelf-gegenereerde *mediator*conditie (er werden immers meer *mediators* opgehaald in de aangereikte *mediator*conditie), maar beide herstudiecondities scoorden slechter op de eindtest dan de testconditie. Dit resultaat is in lijn met onder andere de studie van Karpicke en Smith (2012) waarin werd verondersteld dat het *testing effect* niet kan worden verklaard met behulp van de *elaborative processing* hypothese.

In hoofdstuk 6 werd een studie beschreven die de houdbaarheid van de *retrieval effort* hypothese van het testing effect onderzocht. In deze studie werd niet de toetsvorm, maar de strategie gevarieerd van het maken van de toets door de proefpersoon. De *retrieval effort* hypothese stelt dat des te meer inspanning het ophalen van informatie uit het geheugen kost, des te groter het voordelige effect van testen op de latere geheugenprestatie zal zijn. Deelnemers in deze studie bestudeerden een tekst en na het bestuderen van de tekst werden ze willekeurig verdeeld over drie condities: een *direct choice* conditie (MC-conditie), een *generate response* conditie (GR-MC-conditie) of een herstudieconditie. In de GR-MC-conditie moesten deelnemers eerst zelf een antwoord genereren op een vraag alvorens zij uit de antwoordalternatieven het correcte alternatief moesten te kiezen. In de MC-conditie werd deelnemers gevraagd direct de antwoordalternatieven te bekijken en het correcte alternatief te kiezen. De deelnemers in de herstudieconditie werd gevraagd een lijst met geselecteerde kernfeiten uit de tekst te herbestuderen (deze feiten kwamen terug op de eindtest). Deze drie condities werden vergeleken op de toetsscores van een eindtest die na een retentie-interval van een week werd afgenomen. De GR-MC-conditie werd desgevraagd door de deelnemers beoordeeld als de conditie waar de meeste inspanning voor was geleverd. Echter, deze meer inspannende manier van toetsen resulteerde niet in hogere scores op de eindtest. Hoewel de

scores in de twee testcondities beter werden over de tijd in de zin dat de scores op de eindtest hoger waren dan op de tussentest, waren de scores op de eindtest van de herstudieconditie even hoog als die van de twee testcondities.

In het vervolg van deze samenvatting zullen de belangrijkste resultaten worden bediscussieerd in termen van theoretische verklaringen en praktische toepassingen van het *testing effect*.

Algemene discussie

Een van de meest opmerkelijke bevindingen van dit proefschrift is dat de *elaborative processing* hypothese van het *testing effect* niet kon worden ondersteund. Een voorspelling die voortkomt uit deze hypothese, namelijk dat elaboratief herbestuderen vergelijkbare geheugenprestaties tot gevolg zou hebben als testen, is onderzocht in dit proefschrift maar kon niet worden bevestigd. Hoewel een behoorlijk aantal eerdere studies wél ondersteuning vond voor de *elaborative processing* hypothese (bijv. Carpenter, 2009; Carpenter & DeLosh, 2006), werd zij tegengesproken door zeer recente onderzoeken van Karpicke en Smith (2012) en van Karpicke en Blunt (2011). De resultaten van de studies beschreven in hoofdstuk vier en vijf van dit proefschrift zijn in lijn met deze laatstgenoemde studies en suggereren dat de *elaborative processing* hypothese mogelijk niet houdbaar is.

Een andere belangrijke bevinding beschreven in dit proefschrift heeft betrekking op de rol van schematisatie van kennis in verklaringen van het *testing effect*. De resultaten van de studie beschreven in hoofdstuk 3 en 4 laten zien dat kennis schematisatie zoals gemeten met de *remember-know* procedure waarschijnlijk geen rol speelt in de verklaring van het *testing effect*. Schematisatie wordt gedacht het resultaat te zijn van verschillende ervaringen met de nieuw te leren informatie. Verschillende ervaringen met die nieuw te leren informatie wordt soms ook gezien als elaboratie (bijv. Conway et al. 1997). Om die reden werd de relatie tussen de twee toetsvormen, te weten meerkeuze vragen en meerkeuze verantwoordingsvragen, die varieerden in mate van elaboratie en kennisschematisatie ook onderzocht. Op deze manier kon eveneens de *elaborative processing* hypothese worden getoetst. De resultaten lieten echter geen relatie zien tussen de toetsvorm en kennisschematisatie en waren daarom opnieuw niet in overeenstemming met de *elaborative processing* hypothese. Tot slot waren de resultaten van de studie beschreven in hoofdstuk 6 niet in overeenstemming met een aan de *elaborative processing* hypothese gelieerde hypothese, namelijk de *retrieval effort* hypothese.

De resultaten van dit proefschrift, gecombineerd met de eerder genoemde studies van Karpicke en Smith (2012) en Karpicke en Blunt (2011) lijken sterke implicaties te hebben voor het onderzoek naar de verklaringen van het *testing effect*. Mogelijk dient de aandacht te worden verschoven: in plaats van de elaboratieve aard van het ophaalproces in het geheugen te proberen te begrijpen, is het misschien relevanter om de verschillen en overeenkomsten

tussen de encodeerprocessen en processen tijdens het ophalen van informatie te bestuderen. Misschien is het namelijk niet zozeer de elaboratieve aard van het ophaalproces dat verantwoordelijk is voor een verbeterde geheugenprestatie na testen, maar de versterking van het geheugenspoor tussen twee elementen (de zogenaamde *cue* en *target*) als gevolg van succesvol ophalen van informatie uit het geheugen. Wanneer men een *cue* gepresenteerd krijgt, wordt die *cue* gebruikt om te beslissen welke van een aantal mogelijke *targets* de daadwerkelijke in combinatie met de *cue* bestudeerde *target* was. Wanneer onderscheidende kenmerken van een *target* zijn gekoppeld aan een specifieke *cue*, dan is de zogenaamde diagnostische waarde van de *cue* hoog (Nairne, 2002). Karpicke en Smith (2012) en Karpicke en Blunt (2011) noemen dit eveneens *cue diagnositicity*. Het ophalen van informatie (een *target*) uit het geheugen op basis van een *cue*, verhoogt mogelijk de diagnostische waarde van die *cue*, terwijl dit waarschijnlijk na herbestuderen minder het geval is. In toekomstige pogingen om het *testing effect* te verklaren zouden we ons kunnen richten op een beter begrip van de processen en condities van het ophalen van informatie die leiden tot hogere *cue* diagnosticiteit (bijv. Nairne, 2002).

De andere belangrijke onderzoeksvraag in dit proefschrift had niet zozeer betrekking op de theoretische verklaringen van het *testing effect*, maar betrof de vraag hoe de praktische toepassing van het gebruik van *retrieval practice* (het ophalen van informatie uit het geheugen) om langetermijnretentie te verbeteren en zodoende ook het gebruik ervan in het onderwijs te optimaliseren. Om deze onderzoeksvraag te beantwoorden werd het effect van twee verschillende toetsvormen (meerkeuze en meerkeuze verantwoordingsvragen) op *memory awareness* en op langetermijnretentie van kennis onderzocht. Ook werden de mogelijk verscheidene effecten van twee verschillende toetsafnamestrategieën (de *generate response* meerkeuze testconditie en *direct choice* meerkeuze testconditie) op langetermijnretentie van kennis bekeken. Uit de studies beschreven in hoofdstuk 3 en 6 kan worden geconcludeerd dat de verschillende toetsvormen of toetsafnamestrategieën die zijn onderzocht geen verschillend effect hebben op langetermijnretentie van geheugen. Daarnaast lijken de resultaten beschreven in hoofdstuk 6 aan te sluiten bij eerder onderzoek dat suggereert dat het ophalen van informatie niet alleen moeite moet kosten, maar ook succesvol moet zijn, wil het gunstig zijn voor de geheugenprestatie (bijv. Carpenter, 2009). Op basis van dit onderzoek, kan mogelijk worden gesuggereerd dat de genereer meerkeuze testconditie waarschijnlijk enkel bevorderlijk is voor beter presterende studenten en niet voor minder goed presterende studenten, omdat beter presterende studenten waarschijnlijk beter in staat zijn goed te presteren op initiële testen (m.a.w. meer succesvol zijn in het initieel ophalen van kennis uit het geheugen). Docenten zullen daarom voorzichtig moeten zijn met het adviseren van studenten om zichzelf te testen op geleerde informatie. Voor sommige studenten kan dit zeer voordelig zijn, maar als het niveau van initiële kennis niet toereikend is (wat het onderzoek in hoofdstuk 2 ook

suggereert), zijn er ook geen voordelige effecten van testen te verwachten indien er geen feedback wordt gegeven. Erger nog, studenten kunnen zelfs negatieve effecten ondervinden in termen van verminderde studiemotivatie of prestatie. Het is van belang dat studenten de informatie allereerst goed bestuderen zodat ze het succesvol op kunnen halen uit hun geheugen, voordat hun geheugen voor deze informatie (en daarmee hun studieprestatie) kan worden verbeterd als gevolg van testen.

Curriculum Vitae

CURRICULUM VITAE

Lydia Schaap was born in Leersum, the Netherlands, on October 14th 1980. After completing secondary education (VWO) at Griftland College in Soest, she started studying Psychology at Maastricht University. In 2003, she obtained her Master's degree in Cognitive Psychology. In January 2005, she started to work at the Institute of Psychology of the Erasmus University in Rotterdam, coordinating the progress test and teaching several first year bachelor courses and practicals. In September 2007, she started to combine her educational tasks and a membership of the examination board with a PhD project which resulted in the current dissertation on the cognitive processes involved in the testing effect. Since December 1st 2012, Lydia started to work as an assistant professor at the Institute of Psychology of the Erasmus University in Rotterdam.

PUBLICATIONS

International publications

Schaap, L., Schmidt, H. G., & Verkoeijen, P. P. J. L. (2012). Assessing knowledge growth in a psychology curriculum. *Assessment and Evaluation in Higher Education, 37*(7), 875-887. doi: 10.1080/02602938.2011.581747

Schaap, L., Verkoeijen, P. P. J. L, & Schmidt. H. G. (submitted). Effects of different testing formats on long-term retention and schematisation of knowledge.

Schaap, L., Verkoeijen, P. P. J. L, & Schmidt. H. G. (submitted). Further evidence that the elaborative processing hypothesis cannot account for the testing effect.

Schaap, L., Verkoeijen, P. P. J. L, & Schmidt. H. G. (submitted). Investigating the processes underlying the testing effect: The role of elaborative processing, familiarity and recollection.

Schaap, L., Verkoeijen, P. P. J. L, Coppens, L. C., Nugteren, M., & Schmidt. H. G. (submitted). Test-taking strategies that require more effortful retrieval do not influence the testing effect.

Dutch publications

Vermeulen, L., Scheepers, A., Adriaans, M., Arends, L., Van den Bos, R., Bouwmeester, S., Van der Meer, F., Schaap, L., Smeets, G., Van der Molen, H., & Schmidt, H. (2012). Nominaal studeren in het eerste jaar [Studying nominally in the first year]. *Tijdschrift voor Hoger Onderwijs, 3*, 204-216.

Schaap, L. (2008). Bevorderen van langetermijnkennis door middel van toetsing [Enhancing long-term retention of knowledge through assessment]. In P. W. J. Schramade (Ed.), Handboek Effectief Opleiden [Handbook on Effective Instruction]. The Hague, the Netherlands: Reed Business Information.

Presentations

Schaap, L., Verkoeijen, P. P. J. L., & Schmidt, H. G. (2012, September). *Is elaborative processing the explanatory mechanism of the testing effect?* Poster presented at the Graduate Research Day, Institute of Psychology, Erasmus University Rotterdam, the Netherlands.

Schaap, L., Verkoeijen, P. P. J. L., Coppens, L. C., & Schmidt, H. G. (2011, September). *A new application of the testing-effect in education: Test-taking strategies to improve long-term retention.* Poster presented at the Graduate Research Day, Institute of Psychology, Erasmus University Rotterdam, the Netherlands.

Loyens, S. M. M., Bogaarts, N., Schaap, L., Rikers, R. M. P. J. & Schmidt, H. G. (2011, April). *Effects of poblem-based learning on knowledge retention and comprehension.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Schaap, L., & Loyens, S. M. M. (2011, April). *A new direction in research on learning styles: The relation between processing strategies and episodic and semantic memory.*
Poster presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Loyens, S. M. M., Elzakkers, L. & Schaap, L. (2010, April). *Differences in learning strategies among university students: Which aspects of students' learning put their study success at risk?* Poster presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

Loyens, S. M. M., Bogaarts, N. & Schaap, L. (2010, April). *The role of problem-based learning in fostering high engagement in the conceptual change process.* Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

Schaap, L. & Schmidt, H. G. (2008, August). *Why do some students stop showing progress on progress tests?* Paper presented at the meeting of European Association for Research in Learning and Instruction/Northumbria Assessment Conference, Berlin, Germany.

Van de Wiel, M. W. J., & Schaap, L. (2005, April). *The role of the tutor in problem-based learning: Perspectives of students and tutors.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

In preparation

Loyens, S. M. M., Bogaarts, N., Schaap, L., Schmidt, H. G., & Rikers, R. M. J. P. (in prep.). Effects of problem-based learning on knowledge retention and comprehension: Answering the call for controlled experiments.

Loyens, S. M. M., Hung, W., Pronk, S., & Schaap, L. (in prep.). The concept map as a pedagogical replacement tool for problem-based tutorial meetings. Manuscript in preparation.

Loyens, S. M. M., Rosito, A. C., & Schaap, L. (in prep.). Tutor characteristics in Problem-Based Learning (PBL): Effects on student achievement and their importance in different PBL phases.

Loyens, S. M. M., Elzakkers, L., & Schaap, L. (in prep.). Differences in learning strategies among university students: Which aspects of students' learning put their study success at risk?

Loyens, S. M. M., Elzakkers, L., Evora, L., & Schaap, L. (in prep.). Students' learning styles: The influence of age, gender, and year of program.

Schaap, L., & Verkoeijen, P. P. J. L. (in prep.). The testing effect in Problem-based learning: Tutorial group meetings as retrieval practice.

Dankwoord

Het nieuws dat mijn proefschrift was goedgekeurd toverde niet alleen bij mijzelf, maar ook bij een heleboel andere mensen een grote grijns op het gezicht. Die grote hoeveelheid positieve reacties was indicatief voor het feit dat ik tijdens het schrijven van mijn proefschrift er geen moment alleen voor heb gestaan. Graag wil ik daarom middels dit dankwoord, al deze mensen hartelijk bedanken.

Allereerst gaat mijn dank uit naar Professor Henk Schmidt, mijn promotor. Beste Henk, zonder jou had ik überhaupt niet kunnen beginnen aan het proefschrift. Er was eigenlijk geen project, weinig officiële tijd en eigenlijk ook geen geld en toch gaf jij toestemming om onder jouw begeleiding aan een duaal promotietraject te beginnen. Heel erg bedankt voor deze mogelijkheid! Ook bedankt voor de ontzettend inspirerende gesprekken. Ik ken niemand die zoveel dwarsverbanden en nieuwe invalshoeken ziet als jij.

Ook Dr. Peter Verkoeijen, mijn co-promotor en dagelijks begeleider verdient een groot woord van dank. Beste Peter, zonder Henk was ik nooit begonnen aan dit proefschrift, maar zonder jou had ik het denk ik nooit afgemaakt. Vanaf het moment dat jij bij het project betrokken raakte, kwam het ritme er beter in en werd het project veel concreter. Ik heb onzettend veel bewondering voor hoe jij direct de kern uit een verhaal te pakken hebt en dat vervolgens superduidelijk kunt uitleggen. Daarnaast bewonder ik ook je werklust en het feit dat jij nooit gestrest over komt. Hoe druk je het ook had, ik heb nooit het gevoel gehad dat je geen tijd voor me had. Tot slot was het een verademing dat jij zo'n humor hebt. De fietstochtjes van het station naar het werk waren vaak al hilarisch en dan moest de dag nog beginnen. Als teken van dank heb ik je daarom geciteerd in stelling 11 ;-)!

Professoren Henk van der Molen en Marise Born, ook jullie bedankt. Beste Henk en Marise, door de constructie die Henk Schmidt had bedacht, werden jullie als instituutsvoorzitters 'opgescheept' met mij. Jullie hebben steeds het vertrouwen gehad dat het project succesvol zou worden afgerond, ondanks de hoge onderwijslast. Bedankt voor dat vertrouwen en de mogelijkheden die jullie me gaven om het proefschrift ook daadwerkelijk af te kunnen maken.

Geachte Leescommissie, bestaande uit professor Remy Rikers, professor Liesbeth Kester, en dr. Katinka Dijkstra. Bedankt dat jullie de tijd en moeite hebben genomen om mijn proefschrift te lezen en te beoordelen. Overige Commissieleden, Professoren Tamara van Gog, Fred Paas en Sabine Severiens, ook jullie bedankt dat jullie plaats willen nemen in de commissie en met mij willen discussiëren over de inhoud van mijn proefschrift.

Een proefschrift kan niet tot stand komen zonder proefpersonen die hun tijd vrij maken om aan experimenten deel te nemen. Hartelijk dank aan alle studenten van het Instituut voor Psychologie van de EUR die hebben deelgenomen aan mijn studies. Daarnaast zijn er nog veel collega's, vrienden en kennissen die hebben deelgenomen aan de laatste twee studies. Zonder al deze mensen was het niet gelukt om de experimenten te draaien. Graag wil ik dan ook Anja, Arlette, Bas, Cecile, Charlotte, Daantje, Desiree, Eduard, Elsbeth, Emmeke, Estella, Floor, Floortje, Giel, Hanneke, Heleen, Inge, Janna, Janneke, Janneke, Johan, Jolien, Jolijn, Josse, Karin, Kim, Kim, Kyung, Laurine, Lein, Linda, Marian, Mariska, Marjolein, Mark, Martine, Martine, Michiel, Milou, Mirthe, Natascha, Quinta, Ralph, Robert Jan, Ronald, Sanne, Saskia, Steef, Stephanie, Suzanne, Sylvia, Theo, Thomas, Tirza, Twan, Vanessa, Vincent, en Yvonne hartelijk bedanken voor de tijd en moeite! Hannie en Mir, jullie extra bedankt voor het benaderen (cq subtiel verplichten in het geval van Hannie ;-)) van mensen uit jullie sociale netwerk om mee te doen met mijn onderzoek. Daarnaast ben ik veel dank verschuldigd aan Noortje voor het programmeren van de "Mount Pinatubo" studie en Michelle voor het draaien van de "Mount Pinatubo" pilot studie in het kader van haar bachelor thesis.

Een proefschrift komt ook niet tot stand zonder collega's. Alle collega's van het Instituut voor Psychologie wil ik dan ook hartelijk danken voor de collegialiteit en gezelligheid. In het bijzonder wil ik de O&O-collega's (Remy, Fred, Tamara, Sofie, Huib, Peter, Martine, Daniel, Noortje, Gerdien, Mario, Kim, Wim, Lisette, Jacqueline, Jan, Nicole en Lysanne) bedanken: wat is het fijn om met jullie samen te werken! Er zijn leuke initiatieven, zoals de writing week, en naar de 'pubgroupmeetings with double meaning' kijk ik om wisselende redenen steeds uit ;-). Congressen en researchmeetings zijn informatief, leerzaam en er is altijd ruimte voor humor. Nog een extra bedankje gaat uit naar Tamara. Lieve T, dank voor jouw immer luisterend oor, humor, wijze raad en voor de supergoede feedback die ik van je heb gekregen om de puntjes op de i van mijn proefschrift te zetten. Eigenlijk moet ik Bas daarvoor ook nog bedanken, want hij heeft je daardoor minimaal een weekend moeten missen ;-).

Wat voor werk je ook doet, zonder secretariaat ben je nergens. Hanny, Mirella en Iris, jullie waren en/of zijn het hart van het Instituut. Zonder jullie loopt alles in de soep!

Het EBL heb ik niet voor alle studies nodig gehad, maar beste mannen van het lab, jullie waren er wanneer nodig. Geen vraag is jullie te gek en oplossingsgericht denken hebben jullie volgens mij uitgevonden!

Naast het schrijven van mijn proefschrift waren er ook nog een heleboel onderwijstaken die vervuld moesten worden. Lieve meiden van het onderwijsburo en het PsyWeb-team, door jullie ondersteuning werden die taken een stuk gemakkelijker uitvoerbaar. Jullie dachten soms al vóór mij. Dan kwam ik vragen of jullie iets wilden regelen en dan was het al geregeld!

Als welkome afleiding op mijn onderzoek mocht ik heel wat uurtjes in de vergaderzaal doorbrengen met de Examencommissie en de studieadviseurs. Lieve (oud)excie-leden en studieadviseurs: Guus, Lidia (neem je je dwarsfluit mee naar mijn feestje? ;-)), Marja, Martine, Laurine en Nathanja, hartelijk dank voor de super fijne samenwerking en de leerzame tijd. Ik zal de urenlange BSA-vergaderingen in de zomer nooit vergeten (inclusief de 'bullshitknop' ;-)). En bij het lezen van de afkorting P.O. begin ik nog steeds spontaan te lachen!

Lieve Martine en Laurine, jullie nog extra bedankt voor jullie vriendschap die is ontstaan in deze periode. Vergaderen doen we allang niet meer samen, maar lachen (en ok, af en toe een traantje laten) doen we nog steeds!

Als je aan een proefschrift schrijft, is het hebben van een fijne kamergenoot heel belangrijk. Zo'n geluk als ik had met mijn roomie Sofie, dat was wel heel bijzonder. Later in dit dankwoord kom ik daar nog op terug. Soms waren er echter goede redenen voor mijn roomie om een paar maanden niet op het werk te zijn. Dat zou heel eenzaam zijn geweest, als ik niet twee geweldige stand-in roomies zou hebben gehad. Lieve Steef en Jolijn, heel erg bedankt daarvoor! Jullie hebben er die maanden voor gezorgd dat de gezelligheid op T13-44 bleef!

Werk en privé lopen eigenlijk heel erg door elkaar heen, omdat collega's ook vrienden kunnen worden, vrienden soms tijdelijk ook collega's worden (Lein ;-)), omdat werk soms moet wijken voor vrienden en vrienden soms voor werk.

Liefste vriendinnetjes (Arlette, Cecile, Charlotte, Dees, Emmeke, Floortje, Hannie, Ilja, Jacqueline, Laurine, Lein, Lydia, Martine, Mir, Natascha, Sas, Sofie, Steef, Tamara, Ta'Sangka, Vanes) en jullie gezellige wederhelften, allemaal enorm bedankt voor de vele kaartjes, mailtjes, berichtjes, telefoontjes en aandacht omtrent mijn proefschrift. Maar vooral bedankt dat jullie er altijd voor mij zijn en snappen (of in elk geval doen alsof ;-)) dat het werk soms voor gaat. Ik hoop echter dat jullie ook weten dat het werk volkomen onbelangrijk is in vergelijking met jullie! Ik heb permanente tijdnood om jullie regelmatig te kunnen zien, maar in mijn gedachten zijn jullie altijd bij mij. Een paar wil ik er in het bijzonder noemen:

Lieve Arlet, jij bent mijn 'oudste' vriendinnetje en ondanks dat je aan het begin van onze vriendschap een trauma opliep omdat je als vierjarige opgesloten raakte op het toilet, heeft je dat niet afgeschrikt om de rest van je leven met mij bevriend te blijven ;-). Het is super om

te weten dat wij altijd contact zullen blijven houden. Ook al zien we elkaar soms maanden niet, we kletsen verder alsof we elkaar de dag ervoor nog zagen.

Lieve Il, (alias Illie-B of chicka!), jij bent mijn 'verste' vriendinnetje, omdat je helemaal aan de andere kant van de wereld woont. Jij bewijst dat 'uit het oog, niet uit het hart' hoeft te zijn! Bedankt, bedankt, bedankt dat je er op de dag van mijn verdediging bij bent! Wat een geweldig cadeau!

Lieve Sas en Vanes, na het VWO gingen we ieder onze eigen weg qua stad om te studeren, maar de vriendschap bleef! Bedankt dat jullie er altijd voor me zijn en me steunen in wat ik doe. Hoogte- en dieptepunten hebben we gedeeld en zullen we blijven delen. Sas, bedankt dat je blijft bellen en je niet laat ontmoedigen door 'wéér die voicemail' als onze werktijden weer eens niet op elkaar afgestemd zijn. Vanes, ik kijk uit naar alle etentjes in Rotterdam die we nog gaan krijgen. Die avondjes zijn zo waardevol!

Lieve Natas, jij bent ongekend attent en betrokken! Als ik op onmogelijke tijden nog zat te werken, kwam er steevast een smsje van jou langs om me succes te wensen. Dank daarvoor én voor het feit dat je ieder weekend klaar staat om mij alle hoeken van de squashbaan te laten zien ;-).

Lieve Mir, Lein en Dees, tijdens de studie Psychologie zijn we bevriend geraakt. Mir als connecting factor van de 'redheads' ;-). Wat hebben we een geweldige studietijd gehad samen. Zeg: Picknick aan de Maas, Sinterklaas, Luuuudia....en ik grijns! En nog steeds zijn jullie me superdierbaar! Bedankt voor de diepe gesprekken, de etentjes, de gezelligheid en jullie trouw. Lieve Lein, ik vergeet nooit dat je, op een voor mij donkere dag, je vakantie in Vlissingen onderbrak en met je mooie, dikke buik een hoopje ellende kwam troosten! Lieve Dees, een van jouw mooiste woorden aan mij waren geschreven voorin jouw boek dat je een paar jaar geleden hebt uitgebracht. Nu is het mijn beurt in mijn boekje: bedankt dat je nog steeds mijn vriendinnetje bent en mijn leven verrijkt door mij op een andere manier naar de wereld te laten kijken dan dat ik zelf in eerste instantie zou doen. Lieve Mir, jij verdient een apart stukje in het dankwoord (zie verderop).

Lieve Hannie & Ly, bedankt voor alles! Jullie zijn in alle opzichten het meest bijzondere stel dat ik ken. Hannie, ik denk dat ik tijdens het schrijven van mijn proefschrift met niemand zoveel gemaild heb als met jou. Eén-zinnige, onzinnige en dubbelzinnige mails vlogen me soms om de oren. Maar met die mails hield je me heel goed in de gaten. Als er naar jouw mening teveel tijd tussen het sturen en beantwoorden van een mail zat (zeg een uur ;-)) dan kreeg ik gelijk een mail of sms met de vraag of alles wel goed was en of ik niet te hard aan het werk was. Na een geruststellende mail van mijn kant, volgde dan vaak een mail met een voorstel voor een rummikub- of eetdate. Bedankt voor alle welkome afleiding van jullie beiden naast het schrijven van mijn proefschrift en voor het feit dat ik altijd welkom ben bij jullie!

Beste Frank, toen ik jou vroeg of je me wilde helpen met de omslag van mijn proefschrift, was je zeer vereerd. Ik ben op mijn beurt zeer vereerd dat jouw mooie werk nu prijkt op de omslag van mijn proefschrift! Heel erg bedankt!

Sommige mensen hebben het geluk heel lieve ouders te hebben. Anderen hebben daarnaast ook nog heel leuke schoonouders. Ik bevind me in de begenadigde positie daar bovenop óók nog geweldige surrogaatouders te hebben, bij wie er bijna standaard een bord extra op tafel wordt gezet of een bed wordt opgemaakt als ik in een straal van 20 km in de buurt van Maastricht/Vlijtingen kom. Voor mijn ouders komt straks een apart stukje, maar lieve Frank & Ditte, Mieke & Bob, Wim & Resi (alias de Pieten), en Majella & Roger, bedankt dat jullie allemaal op een bepaalde manier een soort ouders voor me zijn en dat de deur altijd open staat!

Lieve grote broer, broartien en sis, tijdens het schrijven van dit dankwoord moet ik terugdenken aan de vele anekdotes uit onze jeugd die nog regelmatig in geuren en kleuren de revue passeren (denk: lego & SRV, 'het deurtje klemt', BZN, gehaakte poedels, kauwen op melk, tumtummetjes bij het zwembad (sis: hoe heette hij, Wilfred?), tankpistool in zak van regenjas en natuurlijk Notentaart!). Ze zijn tekenend voor het 'pesten' dat we onderling nog steeds heel veel doen en waar we erg veel lol om kunnen hebben. Dit is heel speciaal voor mij en het heeft me ook zeker geholpen in het omgaan met tegenslagen bij het schrijven van mijn proefschrift. Commentaar van reviewers is echt níks vergeleken bij wat wij elkaar voor de voeten gooien! Ik ben dankbaar dat jullie er (hopelijk alledrie) bij zijn om me te steunen op 7 juni en ik hoop dat ik tijdens de verdediging geen nieuwe input voor het familie blunder-repertoire lever ;-).

Lieve Johan, voor jou nog een extra woord van dank omdat jij, naast super trotse grote broer, ook nog eens mijn paranimf wil zijn. Ik vind het een enorme steun dat jij achter me staat. Jij hebt de capaciteiten om wel 10 proefschriften te schrijven, maar je blijft het denk ik houden bij dankwoordvermeldingen, of niet ;-)?

Lieve Pap en Mam, dit wordt denk ik het lastigste stukje van het dankwoord, want waar te beginnen? Jullie hebben me op de wereld gezet en geprobeerd me alles mee te geven om een beetje een goed mens te worden. Bedankt voor alles wat ik van jullie geleerd heb: dat je als je vrijheid krijgt, die alleen kunt behouden als je ook je verantwoordelijkheid neemt. Dat eerlijk het langst duurt en dat je van hard werken niet slechter wordt. Wijze lessen die me hebben geholpen bij het afronden van mijn proefschrift. Bedankt dat jullie mijn onafhanke-

lijkheidsdrang niet inperkten, maar me juist het gevoel gaven vertrouwen te hebben in mijn keuzes. Boven alles bedankt dat jullie zulke lieve ouders en mensen zijn!

Lieve roomie, jij hebt het hele proces van het schrijven van dit proefschrift van het meest dichtbij meegemaakt en je kreeg daardoor alle ups and downs als eerste mee. Wat had ik zonder jou moeten beginnen? Wijze raad, een brede schouder, een valse grap, jij wist precies wat ik nodig had als het even tegenzat. Enorm dankbaar ben ik dan ook voor jouw vriendschap die bestaat uit attentheid, ultieme trouw en onwaarschijnlijk veel humor. De briefjes op mijn bureau, smsjes en kaartjes zijn ontelbaar en met de hoeveelheid tranen van het lachen die er geplengd zijn op onze kamer, kun je een badkuip vullen! Met jou is iedere (werk)dag een feestje en ik ben daarom zoooo blij dat jij mijn Chef de Party wil zijn op mijn promotie!

Lieve Mir (alias Jut), ik hoefde jou niet eens als paranimf te vragen, want dat was allang besloten en zo vanzelfsprekend! Het kon niet anders dan dat dit een volgende aflevering zou worden van Jut & Jul on tour ;-). Ik denk dat ik met niemand zoveel gelachen én gehuild heb als met jou. Op de allereerste studiedag in Maastricht zagen we elkaar, en het klikte zo goed dat we nooit meer van elkaars zijde zijn geweken. We hebben wel eens gekscherend gezegd dat we elkaars platonische wederhelften zijn, en ik denk dat dat de lading het beste dekt ;-). Ontzettend bedankt dat je er echt altijd voor me bent en achter me staat en straks bij de verdediging ook nog eens letterlijk!

Lieve Ralph, ze zeggen dat het venijn in de staart zit, maar niets bleek minder waar, want in de staart van mijn proefschrift ontmoette ik jou! Ondanks de enorme drukte, vonden we gelukkig genoeg tijd om vol van elkaar te genieten en elkaar te leren kennen. Bedankt dat je me zo enorm steunt, trots op me bent, een heel klein beetje 'rust in mijn donder brengt' en mijn allerliefste lief bent! Ik hoop dat we nog heel lang samen en met volle teugen van ons leventje gaan genieten!

Lydia
Breda, april 2013.