Brain-based teaching

Behavioral and neuro-cognitive evidence for the power of test-enhanced learning

Carola Wiklund-Hörnqvist

Department of psychology

Umeå 2014

ISBN: 978-91-7601-171-3
Cover design by Lise-Lott Frössberg
Electronic version available at http://umu.diva-portal.org/
Printed by: Print & Media
Umeå, Sweden 2014

"*If you read a piece of text through twenty times, you will not learn it by heart so easily as if you read it ten times while attempting to recite from time to time and consulting the text when your memory fails*"

- Francis Bacon (1620/2000, p. 143)

Table of Contents

Acknowledgements

During the past years, I have had the opportunity to meet and interact with a lot of wonderful people and also been given the opportunity to be part of an interesting and challenging setting. How can I thank you all? The answer is: I cannot, because then I have to write an additional thesis only to express my gratefulness to all of you. So, based on that, thank you all, you know who you are!

First and foremost, I would like to thank my supervisors. Having the opportunity to be supervised by three lovely people with a lot of expertise cannot other than be stimulating! Firstly, to my main supervisor Bert Jonsson, you always look at things from the bright side and I have always felt that you have trusted me when I failed to do so! Always a smile on your face, and a positive attitude, that have helped me keep my nose above the water when sometimes feeling like drowning. Secondly, "Mr. Accessible/Available" Lars Nyberg, I cannot understand how you manage to keep up with all our data and studies (don't think you do either☺). Having had the possibility to communicate with you, always present both in mind and mostly physically, have not only solved burning issues but has also been so inspiring. I have always had a smile on my face and in my mind after discussing with you. Finally, Linnea Karlsson, or "Mrs. Control analysis", you are so competent. During all our meetings you have challenged and guided me which has kept me "on the road". In sum, thank you all for just being you!

I have also had the possibility to meet so many wonderful colleagues, related to different aspects of university life. Of course, some of my colleagues are not only colleagues, but also PhD students, sharing a lot of common "ups and downs". You are all great people. Although, Anna-Maria, my lovely friend and angel, we started our trip together already as students at the department and thereafter as PhD students. How could I have managed this trip without you? The answer is; I could not have. Daniel, my friend the crazy football fan. Initially, we collaborated as memory testers within the Betula study, and then became roommates as PhD students combined with all the UPC/UPL courses we took together. In that sense we have shared a lot of fun discussions, although, not always "higher thinking", during all these years. I really appreciate your friendship! And to the lovely PhD girls in the corridor, Petra and Elisabeth, what can I say? I don't think I need to say anything, because you both know how much I admire you. Thank you all, both named and not, for always being there. I am also grateful to Mikaela, thank you for everything! I admire all the creativeness you are a symbol of. And finally I will mention another colleague, although not PhD student, but a wise man

that with a few words can nail complex issues; Bo Molander. Bosse, with all your wise comments and a big heart you have really fascinated me with all your knowledge that you so kindly share.

I have also been given the possibility to be part of a wonderful, creative and multifaceted lab, the UFBI. Michael, “Mr DataZ”, you are outstanding and always positive to our (stupid) ideas. Johan, a cool, smart researcher with a lot of patience, although as Peter, not a fan of shopping (as evident from conference trips) but Johan, you always contribute with thoughtful comments. Peter, I appreciate all the reflections and perfectionism that you are a symbol of. Lenita, promise me to never give up any of your brilliant ideas. To all other members within UFBI, thank you.

I am not only a PhD student, I am also a wife and a mother with a wonderful family. Although, I am quite sure that you have not always thought of me as a lovely and cheerful mum and wife… but deep in my heart I know that you feel so anyway. Rebecka and William, my wonderful kids, I appreciate and admire you for the lovely ones you are. I am sure that you will succeed in whatever you choose to do in the future. You are the best, and I love you both so much! Fredde, my love, despite 25 years together, I am still in love with you, and you are always open minded to all new things. I admire and respect you for your patience and always positive attitude, and in contrast to me, you are calm☺. And my lovely parents – love you!

Lotta, I do not need to tell you how much you mean for me. You know it! Cannot count how many times you have saved me with your skills (☺) and personality! Given all our fun partynights, do you want to share a bottle of wine with me? I want to share (at least) one with you!

Katta, my guardian angel. Deep into my heart, I would have done everything to have you here by my side today. I might be naïve, I might be stupid, but I have always felt the presence of you when needed. There are so many times that I have been speaking to you, totally convinced that you are somewhere in the room. You were the one that trusted me when I failed to do so. You were the one that always encouraged me to “take the fight” instead of running. You were the one that always had time, love, and patience to listen. You are not just a “sister” to me, you are the soul mate in my life, and you are the one that deserves to be honoured in my thesis. Love you so much and wish you would have had the possibility to be here, but I am sure, that you are right now sitting on a cloud with a big smile on your face, just enjoying the situation. I miss you!

Thank you all. Carpe diem!

Abstract

A primary goal of education is the acquisition of durable knowledge which challenges the use of efficient pedagogical methods of how to best facilitate learning. Research in cognitive psychology has demonstrated that repeated testing during the learning phase improves performance on later retention tests compared to restudy of material. This empirical phenomenon is called the testing effect. The testing effect has shown to be robust across different kinds of material and when compared to different pedagogical methods. Despite the extensive number of published papers on the testing effect, the majority of the studies have been conducted in the laboratory. More specific, few studies have examined the testing effect in authentic settings when using course material during the progress of a course. Further, few studies have investigated the beneficial effects with test-enhanced learning by the use of neuroimaging methods (e.g. fMRI). The aim with the thesis was to investigate the effects of test-enhanced learning in an authentic educational context and how this is related to individual differences in working memory capacity (Study I and II) as well as changes in brain activity involved in successful repeated testing and long term retention (Study III).

In study I, we examined whether repeated testing with feedback benefitted learning compared to rereading of introductory psychology key concepts in a sample of undergraduate students. The results revealed that repeated testing with feedback was superior compared to rereading both immediate after practice and at longer delays. The effect of repeated testing was beneficial for students irrespectively of WMC. In Study II, we investigated test-enhanced learning in relation to the encoding variability hypothesis for the learning of mathematics in a sample of fifth-grade children. Learning was examined in relation to both practiced and transfer tasks. No differences were found for the practiced tasks. Regarding the transfer tasks, the results gave support for the encoding variability hypothesis, but only at the immediate test. In contrast, when we followed up the durability of learning across time, the results showed that taking the same questions over and over again during the intervention resulted in better performance across time compared to variable encoding. Individual differences in WMC predicted performance on the transfer tasks, but only at the immediate test, regardless of group.

Together, the results from Study I and Study II clearly indicate that test-enhanced learning is effective in authentic settings, across age-groups and also produces transfer. Integrate current findings from cognitive science, in terms of test-enhanced learning, by the use of authentic materials and assessments relevant for educational goals can be rather easily done with

computer based tasks. The observed influence of individual differences in WMC between the studies warrant further study of its specific contribution to be able to optimize the learning procedure.

In Study III, we tested the complementary hypothesis regarding the mechanisms behind memory retrieval. Recurrent retrieval may be efficient because it induces representational consistency or, alternatively, because it induces representational variability - the altering or adding of underlying representations as a function of successful repeated retrieval. A cluster in right superior parietal cortex was identified as important for items successfully repeatedly retrieved Day 1, and also correctly remembered Day 7, compared to those successfully repeatedly retrieved Day 1 but forgotten Day 7. Representational similarity analysis in this region gave support for the theoretical explanations that emphasis semantic elaboration.

Sammanfattning på svenska

Inom ramen för utbildningsväsendet används prov främst för att erhålla en uppfattning om elevens nuvarande kunskapsnivå inom respektive ämne och inte som ett pedagogiskt verktyg för att stärka lärandet. Ett flertal studier inom kognitiv psykologi har påvisat att det är gynnsamt att testa sig initialt vid inlärning för bibehållandet av kunskap över tid, vilket är ett av huvudmålen inom utbildningsväsendet. Detta fenomen kallas för *testeffekten* (Roediger & Karpicke, 2006a; 2006b). Testeffekten har visat sig vara ett robust fenomen som är fördelaktigt i kunskapsförvärvandet; inom olika ämnesområden, jämfört med andra pedagogiska metoder, samt genererar i transfer av kunskap. Trots ett flertal vetenskapligt publicerade beteendestudier på testeffekten så föreligger det fortfarande en brist på studier från autentiska klassrumskontexter där man använder faktiskt kursmaterial under pågående kursmoment. Betydelsen av att studera detta i autentiska kontexter är av relevans om vi ska kunna transformera kunskapen om testeffekten till pedagogisk praxis och vidare pedagogiska rekommendationer.

En annan informativ och kompletterande metod är att med hjälp av moderna hjärnavbildningsmetoder, som funktionell hjärnavbildning (fMRI), studera vad som händer i hjärnan vid testbaserad inlärning. FMRI möjliggör inte bara identifikationen av underliggande kognitiva processer och hur det relaterar till vårt minne, utan bidrar även till att öka vår förståelse och kunskap om *varför* testbaserad inlärning är gynnsamt.

Syftet med denna avhandling var att studera effekter av testbaserat lärande i autentiska pedagogiska kontexter, och även undersöka hur dessa effekter relaterar till individuella skillnader i arbetsminneskapacitet (Studie I och II). Vidare var syftet att med hjälp av fMRI studera de underliggande neurokognitiva processerna involverade i testbaserad inlärning samt hur dessa är relaterade till långtidsretention (kvarhållandet av information över tid; Studie III).

I Studie I undersökte vi huruvida upprepat testande med feedback är mer gynnsamt för lärande av nyckelbegrepp och dess definitioner jämfört med om man återstuderar materialet. Deltagarna var universitetsstudenter och materialet vi använde var framtaget från aktuell kurslitteratur under pågående kurs. Resultatet visade att de som upprepade gånger testade sig initialt vid inlärning också presterade signifikant bättre, både direkt efter träning och fem veckor senare, jämfört med den grupp som återstuderade materialet. Individuella skillnader i arbetsminneskapacitet påverkade inte

resultatet. I Studie II fokuserade vi på hur olika typer av upprepat testande påverkar inlärning av matematik för elever i årskurs 5. Den ena gruppens inlärning bestod i att upprepat träna på samma uppgifter, medan den andra gruppen tränade på samma uppgifter men med olika formuleringar/siffror på frågorna (variabel inkodning). Grad av inlärning undersöktes i relation till både tränade och otränade (transfer) uppgifter vid tre olika tillfällen; direkt efter träning, 3 dagar senare och 5 veckor senare. Resultatet visade att för tränade uppgifter så hade typ av initial inlärning ingen betydelse. För transfer uppgifterna gav resultatet stöd för att gruppen som tränade på variabel inkodning var signifikant bättre direkt efter träning, men över tid så glömde denna grupp mer, medan den grupp som tränat på samma uppgifter bibehöll kunskapen bättre över tid. Individuella skillnader i arbetsminneskapacitet var statistiskt signifikant direkt efter träning, men detta var endast för transfer uppgifterna, inte för de uppgifter båda grupperna hade tränat på. Framtida studier bör vidare explicit undersöka effekter av individuella skillnader i arbetsminneskapacitet för att på bästa sätt optimera lärandet och bättre utforma rekommendationer.

I Studie III använde vi fMRI för att undersöka underliggande neurokognitiva processer samt för identifikation av relevanta hjärnregioner involverade i testbaserad inlärning. Studien avsåg två sessioner med magnetkamera, Dag 1 och Dag 7. Dag 1 testades försöksdeltagarna på Swahili-Svenska ordpar tre upprepade gånger, Dag 7 genomgick försöksdeltagarna ett test. Resultatet visade att upprepat testande initialt är exekutivt belastande men över repetitioner så reduceras denna kognitiva belastning parallellt med att semantiska representationer stärks upp. Vi kunde även med hjälp av analyser av mönster i hjärnaktivitet finna stöd för att testbaserad inlärning som främjar långtidsretention karaktäriseras av semantisk elaborering initialt vid inlärning. Denna elaborering resulterar i en förstärkning av den semantiska representationen i hjärnan vilket är nödvändigt för lyckad framplockning en vecka senare.

Sammantaget så indikerar resultaten i Studie I och II på att testbaserat lärande är gynnsamt i autentiska miljöer då man använder aktuell kurslitteratur oavsett åldersgrupp och material. Detta ger således vetenskaplig evidens för applicerbarheten av testbaserat lärande som en gynnsam pedagogisk metod. Med hjälp av fMRI kan vi också bättre förklara *varför* testbaserad inlärning visat sig vara gynnsam för långtidsretentionen. Resultaten i Studie III korresponderade väl med resultatet i Studie II genom att påvisa evidens för att upprepade gånger testa sig på samma material möjliggör att hjärnan kan arbeta elaborativt.

Abbreviations

aPFC = Anterior prefrontal cortex

ACC = Anterior cingulate cortex

ACG = Anterior cingulate gyrus

AG = Angular gyrus

dlPFC = Dorsolaterala prefrontal cortex

HC = Hippocampus

IFG = Inferior frontal gyrus

IPL = Inferior parietal lobe

MCC = Middle cingulate cortex

MFG = Middle frontal gyrus

MTG = Middle temporal gyrus

PCC = Posterior cingulate cortex

PFC = Prefrontal cortex

PIPL = Posterior inferior parietal lobe

PPC = Posterior parietal cortex

ROI = Region of interest

RSA = Representational Similarity Analysis

STG = Superior temporal gyrus

VLPFC = Ventrolateral prefrontal cortex

WMC = Working memory capacity

TPE = Test-potentiated encoding

List of papers

I. Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology*, *55*, 10-16. doi: 10.1111/sjop.12093

II. Wiklund-Hörnqvist, C., Jonsson, B., & Nyroos, M. Transfer in Mathematical Learning: A Comparison Study of Elementary School Children in an Educational Context. (Manuscript submitted for publication)

III. Karlsson, L., Wiklund-Hörnqvist, C., Eriksson, J., Andersson, M., Jonsson, B. & Nyberg, L. Lesser Neural Pattern Similarity across Repeated Tests is Associated with Better Long-Term Memory Retention (Under revision)

Paper I have been reprinted with the permission of the copyright holder.

Introduction

One topic that is currently of wide interest in Sweden, as well as globally, is the identification of effective ways to achieve supportive learning, in school contexts and elsewhere, but this has proven to be a challenging task. Central to the educational context is the acquisition of long-lasting memories (i.e. durable learning) which challenges the use of efficient pedagogical methods when determining how to best facilitate learning. Learning can broadly be defined as behavioral changes following experience or practice. Research in cognitive psychology has demonstrated that repeated testing during the learningphase improves performance on later retention tests (i.e. durable learning) when compared to the restudy of material. This empirical phenomenon is called the testing effect. The aim of the current thesis was to investigate the (neuro-) cognitive processes related to the testing effect. One part of this thesis focuses on behavioral data in an authentic educational setting (Study I & II). In another part, a modern brain imaging technique is used to identify the underlying neural correlates involved in test-enhanced learning and long-term retention (Study III).

Memory and Learning

Memory is a fundamental cognitive ability that enables humans to learn, remember, and organize their experiences and events into meaningful contexts. Memory is important for us all; however, there is no time during which memory is more essential than during the educational years. Students are presented with a constant flow of new information that is going to be converted to, or will at least lay the foundation for, lifelong knowledge. During this time, there are high demands on a well-functioning memory which is necessary when performing well in testing situations; one's performance on tests during the school years dictates his or her life to a great extent. Therefore, it is important for researchers to identify learning activities that can be applied in the classroom, and which can enhance the likelihood that students form durable memories (Thorndike, 1906). This highlights the importance of integrating educational research and cognitive (neuroscience) research as they are not mutually exclusive research domains, especially since both are concerned with the complexities associated with successful learning (Goswami, 2004).

Learning is a rather complex term that incorporates several different perspectives, including biological, philosophical, and psychological approaches. Within the present thesis I will focus on learning from a cognitive psychology perspective. According to Gluck, Mercado and Myers

(2008) “learning is the process by which changes in behaviour arise from experience through interaction with the world; memory is the record of past experiences acquired through learning. Neither learning nor memory is a single cohesive process; there are many kinds of memory and many ways to learn” (Gluck, Mercado & Myers, 2008, p. 39). According to Gluck et al’s (2008) definition, memory will act as the crucial component for learning as it acts like the recorder. This definition fits well with the basic processes proposed in memory function: encoding, storage and retrieval. With respect to memory, it should be obvious that the human memory is rather complex, and it includes several different memory systems that more or less depends upon the basic processes mentioned. Learning, on the other hand, is also a process, and the methods that are used to acquire knowledge will have an effect on how well these parcels of information are ‘recorded’ (Gluck, Mercado & Myers, 2008; Thorndike, 1906).

Memory is an umbrella term that holds many different views in terms of how it can be defined. Two of the most common perspectives of memory are in terms of memory systems and memory processes. Both memory systems and memory processes are to some degree defined based on the temporal interval at which information is processed and retained. In the first part of the 21st century several studies have focused on investigations into human memory in order to describe and explain its complexity. It is now well established that human memory is not a unitary system, but rather a system involving multiple structures and processes that subserve each other (Cabeza & Nyberg, 2000a; 2000b; Squire, 2004; Tulving, 1989). The single term memory does not do justice to the concepts it represents, as it includes different memory systems which involve a diversity of continuous cognitive processes. The two different perspectives on memory; systems and processes, will briefly be described and a simplified presentation of the related key regions in the brain will also be provided.

The term working memory refers to a system with limited capacity that is responsible for temporarily storing and manipulating information required for the execution of complex cognitive tasks, such as learning, comprehension and reasoning (Baddeley, 2000; St. Clair-Thompson & Gathercole, 2006). Working memory is commonly associated with increased activity in the prefrontal cortex (PFC), the parietal regions and is predominantly left-lateralized when the material is verbal (Cabeza & Nyberg, 2000a; Cabeza & Nyberg, 2000b; D’Esposito, 2007). The different subregions within the PFC are, in turn, related to the specific cognitive demands related to the task (Badre & Wagner, 2007).

Within the long-term memory classification system, a distinction is made between declarative and non-declarative memory. Declarative memory refers to the conscious recollection of facts and events (i.e. intentional action) whereas non-declarative memory concerns priming, procedural memory, classical conditioning, and non-associative learning (Squire, 2004). Declarative memory can be further divided into semantic memory and episodic memory. Semantic memory concerns common knowledge while episodic memory refers to personal experiences and events that are associated with a particular time and space (Tulving, 1972; 1989). Episodic memory and semantic memory may be regarded as different memory systems, but at the functional level, they are closely interconnected, as new incoming information into episodic memory can be supported by existing knowledge from semantic memory (Prince, Tsukiura & Cabeza, 2007; Ryan, Cox, Hayes & Nadel, 2008) partly under the supervision of working memory. For example, access to semantic representations is supported by the left inferior frontal gyrus (IFG) in the PFC (Habib & Nyberg, 2008; Martin & Chao, 2001), and both semantic and episodic memory have been found to be associated with increased activity in the temporal, parietal, and some prefrontal regions (Cabeza & Nyberg, 2000a, 2000b; Spaniol, Davidson, Kim, Han, Moscovitch, Grady, 2009; Binder, Desai, Graves & Conant, 2009).

At a basic level, memory can be divided into three basic processes consisting of encoding, storage, and retrieval. Encoding refers to the various processes by which information is transformed into a memory representation. Storage refers to the retention of that information - either temporarily within the working memory system or within a more permanent, long-term memory system. Retrieval is the process that enable access and use of the encoded and stored information. These three functions are not mutually exclusive as they, to some degree, subserve each other. Learning incorporates all of these functions; a new stimulus is perceived that will be encoded into a representation and processed in working memory and integrated into a more persistent long-term memory. Thus, to ensure that the stimulus will be available for retrieval later on, it assumes that this stimulus has been encoded and stored within the long-term memory system. It is well established that these cognitive processes engage large-scaled brain networks (Cabeza & Nyberg, 2000b); however, one key region that is primarily related to declarative memory formation is the medial temporal lobe, especially the hippocampus (HC). The HC is not specifically involved in the (permanent) storage of memories; rather, it plays an important role in binding disparate pieces of information into coherent representations during memory encoding (Eichenbaum, Yonelinas & Ranganath, 2007; Hannula & Ranganath, 2008).

With regard to storage, it is important to understand the underlying processes that make new memories possible to later on be retrieved. One important underlying process is consolidation. Consolidation can be regarded as the interface between the (initial) encoding and (further) storage of that information, and this process can be defined as "the progressive post acquisition stabilization of long-term memory" (Dudai, 2004, p. 52). Consolidation occurs at two levels: synaptic consolidation and system consolidation. Synaptic consolidation takes place within the first few minutes following the initial encoding of a stimulus; this process communicates at a cellular and molecular level. System consolidation, the process underlying the connections made to relevant cortical regions, works in parallel with synaptic consolidation but has a longer time interval, ranging from days to months (even years) following the initial encoding (Dudai, 2004). The degree of consolidation processes is also, in some sense, related to how the initial learning is processed (Alberini, 2005; Dudai & Eisenburg, 2004; Lee 2009).

Ways to enhance learning

As mentioned in the previous section, the cognitive processes involved in memory and learning determine the durability of the memory trace. There is a large body of memory research devoted to the different ways to best enhance learning (Craik & Lockhart, 1972; Jacoby, 1978; Morris, Bransford & Franks, 1977; Slamecka & Graf, 1978). For example, Craik and Lockhart (1972) suggested that the quality of how well a memory trace is 'recorded', is related to the different levels of (cognitive) processing that are engaged during encoding (LOP; Craik & Lockhart, 1972). In their influential framework, they suggested that information is better retained if we engage in deep and more semantic processing compared to shallow encoding of the to-be-learned material (Craik & Lockhart, 1972). Subsequently, in a series of experiments Craik and Tulving (1975) tested the LOP by manipulating the depth of encoding and they concluded that encoding depth not only influenced the strength of a memory, it is also partly determined by the degree of semantic elaboration (Craik & Tulving, 1975). A closely related idea, although not entirely similar, is the generation effect (Jacoby, 1978; Slamecka & Graf, 1978) which posits that when individuals are asked to generate information related to the to-be-learned material, memory improves when compared to restudy or simply remembering without elaboration (Jacoby, 1978; Slamecka & Graf, 1978). Taken together, the qualitative differences in the ways through which learning is enhanced are characterized by constructing task demands that require an active engagement of the individual. The following question then arises: how is teaching in school designed in terms of producing a durable memory?

This issue has recently been highlighted in a meta-analysis by Freeman, Eddy, McDonough, et al. (2014). Freeman and colleagues (2014) reviewed 225 studies in relation to learning and course performance based on teaching methods. They concluded that learning activities that required students to be engaged were the most beneficial pedagogical methods when compared to the traditional approach of teaching (i.e. "teaching by telling"; Freeman et al., 2014, p. 8410). In addition, students' failure rate were substantially lower following active learning (21.8%) when compared to traditional lecturing (33.8%). Freeman et al. (2014) concluded that instead of continuing to teach in a traditional way, it is now time to focus on methods that include active learning, and how those methods should be designed to best produce durable learning. How? Freeman et al. (2014) suggested that findings from memory research and cognitive psychology would be pertinent to answer those questions and laying the foundation for further educational practice.

In line with that suggestion, Dunlosky, Rawson, Marsh, Nathan and Willingham (2013) recently highlighted that there are, of course, multiple ways to enhance learning, but the extent to which those techniques support durable learning is less well known. They evaluated ten common learning techniques in terms of their utility. The techniques included elaborative interrogation, self-explanation, summarization, highlighting/underlining, keyword mnemonic, imagery for text, rereading, practice testing (i.e. retrieval practice), distributed practice and finally interleaved practice. Utility was based on evaluating each learning technique related to four factors: learning condition (e.g. the different methods for using the technique); student characteristics (e.g. age, working memory capacity, motivation); materials (from simple facts to complex problem solving) and criterion tasks (different outcome measurements: free recall, problem solving, classroom quizzes). The degree of utility ranged from low to high; and this was also evaluated based on generalizability and issues associated with implementation (e.g. time consuming, easy to adopt). Following an extensive review, the results showed that of all those techniques that were listed, one was rated as having the highest utility. That technique was practice testing, meaning that it holds a great potential for educational purposes: to produce durable learning (Dunlosky et al., 2013).

Is this idea new? The answer is both yes and no. As early as 1890, William James highlighted an important aspect that involves both memory and learning "a curious peculiarity of our memory is that things are impressed better by active than by passive repetition. I mean that in learning (by heart, for example), when we almost know the piece, it pays better to wait and recollect by an effort from within, than to look at the book again. If we recover the words in the former way, we shall probably know them the next

time; if in the latter way, we shall very likely need the book once more". (James, 1890, p. 646). Despite the fact that James recognized this as early as 1890, there is still insufficient knowledge as to why "active repetition" benefits learning within the educational system. Fortunately, and especially during the last decade, there has been a resurgence of interest in the field of cognitive psychology with respect to the phenomenon that active retrieval promotes durable learning, which is the main goal within education (Dunlosky et al., 2013; Freeman et al., 2014; Halpern & Hakel, 2002; Rawson & Dunlosky, 2011; Roediger & Karpicke, 2006a; 2006b).

The testing effect: the consequence of test-enhanced learning

A substantial number of studies in cognitive psychology has demonstrated that taking an initial test during the learning phase improves performance on a later retention test when compared to restudy of that material (i.e. durable learning, Rawson & Dunlosky, 2011; Roediger & Karpicke, 2006a; Tulving, 1967). This phenomenon is called the testing effect (Roediger & Karpicke, 2006a; 2006b) which is defined as "taking a test usually enhances later performance on the material relative to rereading it or to having no re-exposure at all" (Roediger & Butler, 2011, p. 20). In the literature, the testing effect is sometimes referred to as retrieval practice or test-enhanced learning. To clarify this we need to consider the temporal interval: the general learning procedure that precedes the testing effect is test-enhanced learning. Test-enhanced learning improves retention of the learned material when compared to other common strategies as measured by a final test (i.e. the testing effect). This is an important note to highlight as tests in education are commonly viewed as assessments *of* learning (i.e. high-stake tests; summative assessments) and not as a tool that is used to improve learning (i.e. low-stake tests; formative assessment) (Roediger & Karpicke, 2006a; 2006b).

In spite of the fact that this phenomenon was demonstrated approximately 100 years ago (Abott, 1909; Gates, 1917), the interest in retrieval practice to promote learning has received a lot of attention during the last decade within research in psychology. More specifically, Glover's (1989) article entitled "The 'testing' phenomenon: Not gone but nearly forgotten" reinvigorated interest in the phenomenon. Another recent and influential demonstration of the testing effect was provided by Roediger and Karpicke (2006b). In their experiment(s) they used two prose passages that covered scientific topics. Students read each passage once, and they then restudied one passage and took a paper-pencil free recall test of the other. In a free recall test the participant is asked to recall as much information as possible without any

cues or material present. Learning was assessed by a free recall test after 5 minutes, two days or one week. Restudy of the material was beneficial at the immediate test, but the opposite was found at the delayed tests (Roediger & Karpicke, 2006b; Exp 1).

Results across several studies consequently replicate those results: repeated study trials leads to superior performance on the immediate test (i.e. cramming effect), with the reversed pattern following a delay (i.e. testing effect; Roediger & Karpicke, 2006a; 2006b; see Rowland, 2014 for a review). After the studies conducted by Roediger and Karpicke (2006a; 2006b), numerous studies have replicated the testing effect (mostly in the laboratory); showing that the testing effect is a robust phenomenon. Why is repeated testing so beneficial for retention? Testing is not a neutral event. Rather, testing changes memory. The beneficial effect of testing when compared to restudy can best be described in terms of the notion that different memory processes are engaged during learning. Repeated testing forces the subject to actively retrieve (or try to retrieve) information from memory. That is, each time an item is retrieved or, as some argue, is attempted to be retrieved, it strengthens the memory traces that, in turn, increases the probability that the item will be retrieved in the future. In contrast, the restudy of material is more similar to learning by (passive) encoding into memory.

The typical design for the study of test-enhanced learning

Figure 1 is a simplified illustration of how studies examining the testing effect have commonly been designed. Since the testing effect has been studied in relation to several different factors (see the section "Empirical findings of the testing effect"), the different boxes contains different content and/or manipulations.

As can be seen in Figure 1 (below), most laboratory studies use novel material; it is thus common to include a baseline exposure to the material before the intervention (i.e. familiarization of the to-be-learned material). Studies that have been conducted in authentic settings usually define this baseline exposure in terms of lectures and the assigned readings to the lecture. Following the baseline exposure, an intervention is applied; the number of repetitions has varied substantially across studies. Both within- and between-subjects designs are used. It is important to highlight that depending on the research question of interest the intervention (condition A versus B) has usually included several different factors such as comparing different: pedagogical methods, test-formats, material, populations and in different contexts. Learning is commonly measured by means of an

immediate test and/or retention test(s), where some of the studies test half of the items at the immediate test, while the other half of the items are tested at the retention test. Other studies test half of the sample from each condition at the immediate test, and the other half at the retention test; some studies test all subjects and all items both immediately and at the retention test. Memory accuracy (Roediger & Karpicke, 2006a; 2006b) and/or response time (RT; MacLeod & Nelson, 1984; van den Broek, Segers, Takashima & Verhoeven, 2013) for condition A and condition B on the following test(s) (i.e. the immediate and retention tests) is the index of learning. Condition A and B are statistically contrasted against each other related to the performance at the immediate test and the retention test. The typical outcome is that taking a test during the intervention leads to significantly better performance at the retention test, as compared to not taking a test during the intervention.

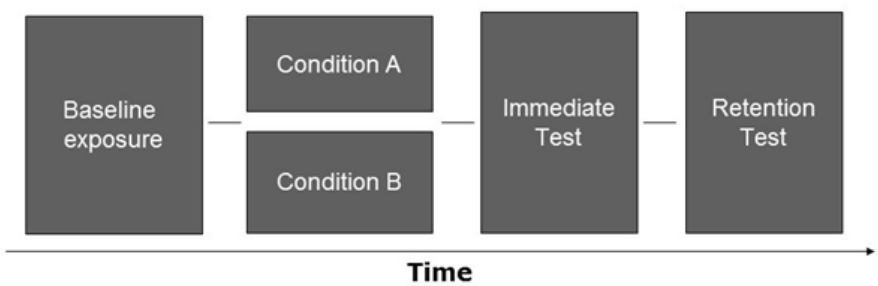

Figure 1. The figure illustrates a typical design that is used in studies of the testing effect.

It is important to distinguish between the direct and the indirect effects of testing. The majority of studies examining the testing effect have focused on the direct effects of testing which simply means that the act of retrieving an item directly from memory enhances later memory. The indirect effect of testing is related to other factors that can be indirectly affected by testing. For example, test-potentiated encoding (TPE) is a mediated effect of retrieval practice. Test-potentiated encoding refers to the idea that prior testing leads to enhanced learning during subsequent encoding/future study. This idea was initially presented by Izawa (1971), which concluded that neither learning nor forgetting occurred during tests, but taking a test facilitates learning during future restudy (Izawa, 1970). Another way of framing the indirect effects of testing is by focusing on the learner. There is some evidence that students do not engage in testing when they study on

their own, but when they do, this also increases the metacognitive knowledge regarding what has been learned and what has not.

Empirical findings of the testing effect from behavioral studies

The robustness of the testing effect has been attested to by several kinds of empirical findings. Below are some of those factors that are of educational significance presented along with examples from empirical studies. It should be stressed that due to a plethora of empirical studies (mostly conducted in the laboratory) that have been published on the testing effect in the last few years, a complete presentation of all the studies is not included in the section below.

Does the testing effect hold for authentic course material outside the laboratory?

Even if most of the studies examining the testing effect have been conducted in the laboratory, there are some studies that have used stimulus materials included in the curricula during the progress of a course. The materials used have included history (Carpenter, Pashler & Cepeda, 2009), statistics (Lyle & Crawford, 2011; Szpunar, Khan & Schacter, 2013) and reading material (Hattikudur & Postle, 2011; Leeming, 2002; McDaniel, Anderson, Derbish, & Morrisette, 2007; Wiklund-Hörnqvist, Jonsson & Nyberg, 2014). More recent studies have also started to examine how test-enhanced learning, when using actual course material, affects summative assessment (i.e. high-stake tests) in college classrooms (McDaniel, Wildman & Anderson, 2012: open-book quizzes on-line) and in middle school classrooms McDaniel, Agarwal, Huelser, McDermott & Roediger, 2011; McDaniel, Thomas, Agarwal, McDermott & Roediger, 2013; McDermott, Agarwal, D'Antonio, Roediger & McDaniel, 2014; Pennebaker, Gosling & Ferrel, 2013; Roediger, Agarwal, McDaniel & McDermott, 2011). For example, across three experiments, McDermott, Agarwal, D'Antonio, et al. (2014) had seventh-grade students learn material from science units included in the curriculum by using short-answer (SA) or multiple-choice (MC) quizzes (with correct-answer feedback). At the end-of-semester exam, they found evidence that prior quizzing enhanced performance at the regular exam, and this was even true compared to material that had been restudied (McDermott, Agarwal, D'Antonio, Roediger & McDaniel, 2014).

As indicated by Dunlosky, Rawson, Marsh, Nathan and Willingham (2013), one critical factor that contributes to the notion that learning activities are of educational significance is that they need to be easy to use – both for the

teachers and for the students. In line with that, different methods have been used in the above-mentioned studies; some studies have used online testing (Hattikudur & Postle, 2011; McDaniel, Anderson, Derbish, & Morrisette, 2007; McDaniel, Wildman & Anderson, 2012), in-class testing (Leeming, 2002; Lyle & Crawford, 2011; Pennebaker, Gosling & Ferrel, 2013; Wiklund-Hörnqvist, Jonsson & Nyberg, 2014), and classroom response systems ('clickers'; McDaniel, Thomas, Agarwal, McDermott & Roediger, 2013; McDermott, Agarwal, D'Antonio, Roediger & McDaniel, 2014). Pennebaker, Gosling, and Ferrel (2013) used an exam-a-day procedure where students (N= 901) started each lecture with an online quiz (which took place in the classroom), and the quiz pertained to the daily reading assignment or to the content from the previous lecture. The results revealed that performance on the regular exam was improved by half a letter grade when compared to a non-quizzed class (N= 935). All of the methods mentioned earlier are relatively easy for teachers to implement, and they have apparently produced robust testing effects beyond the laboratory.

Does the testing effect hold for different kinds of topics?

Most studies, both those conducted in the laboratory and those that took place in authentic classroom settings, have demonstrated a testing effect when using verbal material. If test-enhanced learning is going to be recommended as an effective learning method for practical applications, it is also necessary to find empirical evidence that it holds for different topics that touch upon on different materials and skills. Some evidence exists for that the testing effect holds across educational topics, including geography (Rohrer, Taylor & Sholar, 2010), history (Carpenter, Pashler & Cepeda, 2009), statistics (Lyle & Crawford, 2011), resuscitation skill learning (Kromann, Jensen & Ringsted, 2009) and improves the clinical application of knowledge for medicine residents (Larsen, Butler & Roediger, 2009). The findings by Kromman et al. (2009) and Larsen et al. (2009) are important as they highlights the role of test-enhanced learning in skill learning that extends beyond learning just theoretical facts.

Does the testing effect hold when compared to other common pedagogical methods?

Despite the fact that the majority of the testing effect studies have compared testing with a restudy condition, some evidence exists for the idea that testing is superior when compared to other pedagogical methods that are considered to be active learning activities, specifically concept mapping (Karpicke & Blunt, 2011) and group discussions (Stenlund, Jönsson & Jonsson, submitted). Concept mapping and group discussions are commonly rated as popular learning methods by both teachers and students. These are

assumed to be more elaborative learning methods when compared to more traditional teaching and/or study methods.

Although, retrieval practice does not solely refer to taking tests, it can be applied to several different pedagogical methods as long as the retrieval practice requires students to be actively engaged. Support for this idea was provided in a recent study by Karpicke and Blunt (2014). Across two experiments they examined the influence of retrieval practice in terms of testing and concept mapping with the to-be learned material either present or absent. The results showed that retrieval practice with the material absent, independent of test or concept mapping, produced significantly better performance one week later compared to retrieval practice with the material present. The results clearly point to the fact that the beneficial effects of testing is not affected by the writing procedure per se; instead Karpicke and Blunt (2014) concluded that the crucial key to produce meaningful learning (i.e. durable learning) is to engage in active retrieval with the material absent (Karpicke & Blunt, 2014). These results also highlight that several different pedagogical methods can be used to promote learning as long as they include the act of the subjects' engagement while retrieving information from memory.

Does test-enhanced learning generate transfer of learning?

One criticism that should be reflected upon is whether testing only produces rote learning. Testing enhances the transfer of learning. What is the evidence of transfer? According to Barnett and Ceci (2002), "if schooling has positive effects for measures other than those directly taught, this could be construed as evidence of transfer" (Barnett & Ceci, 2002, p. 618). Indeed, several studies have found that testing facilitates transfer to a greater degree than study across different knowledge domains using different kinds of materials; this materials include prose passages (Butler, 2010), geography (Rohrer, Taylor & Sholar, 2010), mathematical function learning (Kang, McDaniel & Pashler, 2011), and the learning of complex spatial locations (Carpenter & Kelly, 2012). Testing also produces transfer to non-quizzed materials, resulting in improved exam scores (McDaniel, Thomas, Agarwal, McDermott & Roediger, 2013). Rohrer, Taylor, and Sholar (2010) found support for the testing effect regarding transfer tasks in a sample of fourth/fifth-grade students who were learning geography. In two independent experiments the researchers found that students in the testing condition, as compared to those in the study condition, performed significantly better in both the practiced and transfer tests one day after the initial learning phase. The concept of transfer was used differently in those studies. Some of the researchers defined transfer as performance across testformats (Carpenter &

Kelly, 2012; Rohrer, Taylor & Sholar, 2010) and while others defined it across knowledge domains (Kang, McDaniel & Pashler, 2011).

Is test-enhanced learning beneficial across different age cohorts and populations?

Most of the studies within the testing effect literature have been conducted with adults and future studies should further investigate whether the testing effect can be achieved even in young children. For example, Lipko-Speed, Dunlosky and Rawson (2014) had students in the fifth-grade learn key concepts either by taking tests with or without feedback or by restudy science concepts. Approximately one week later, the results showed that testing with feedback was best, but that testing per se did not improve learning more than restudy. Even though they are limited in number, some recently published studies have found support for the testing effect in different age cohorts other than young adults, ranging from children (Carpenter, Pashler & Cepeda, 2009; Goossens, Camp, Verkoeijen & Tabbers, 2014; Karpicke, Blunt, Smith & Karpicke, in press; McDaniel, Thomas, Agarwal, McDermott & Roediger, 2013; McDermott, Agarwal, D'Antonio, Roediger & McDaniel, 2014; Rohrer, Taylor & Scholar, 2010) to older adults (Meyer & Logan, 2013). Recently, an equally sized testing effect was demonstrated for individuals suffering from severe traumatic brain injury (TBI) as compared with healthy controls (Pastötter, Weber & Bäuml, 2013, see also Sumowski, Leavitt, Cohen, Paxton, Chiaravalloti, & DeLuca, 2013 for similar findings). This finding is of educational relevance as there is a lot of evidence from memory research that individual variations in working memory capacity is associated with scholastic performance, and this is particularly evident in children (Alloway & Alloway, 2010; Bayliss, Jarrold, Gunn & Baddeley, 2003; Hitch, Towse & Hutton, 2001).

How is response time related to the improvement following test-enhanced learning?

Memory accuracy is commonly used as the index of whether an item has been learned or not. Another index of how well an item has been learned is by examining latency. Decreased response time (i.e. latency) is a factor that has been suggested to be a consequence of repeated retrieval practice (Keresztes, Kaiser, Kovács & Racsmány, 2013; MacLeod & Nelson, 1984; van den Broek, Takashima, Segers & Verhoeven, 2013). MacLeod and Nelson (1984) suggested that accuracy and latency of correct recall can be seen as an index for different aspects of memory. While accuracy reflects whether the item has been learned or not, latency reflects how much processing is needed for an item to be retrieved. From their view, a reduced response time is indicative that the item is less cognitively effortful to retrieve (MacLeod &

Nelson, 1984). More recent research focusing on the testing effect has suggested that reduced response time following testing reflects a reduced search set from memory (Keresztez et al., 2013) and/or a strengthening of cue-response associations (van den Broek, Takashima, Segers & Verhoven, 2013).

Are there any indirect effects of test-enhanced learning?

Survey studies confirm that students do not engage in self-testing while studying on their own, probably due to a lack of knowledge of its pedagogical benefits (Karpicke, Butler & Roediger, 2009; Kornell & Bjork, 2007; Yan, Thai & Bjork, 2014). Leeming (2002) and Pennebaker, Gosling and Ferrel, (2013) used an in-class exam-a-day procedure where each lecture started with a quizzing procedure that was related to either the assigned reading or to content from a prior lecture. Leeming (2002) surveyed students' attitudes toward the quizzing procedure both before and after it was completed. The results revealed that the students were initially skeptical when informed about the procedure, but after having completed the course, the majority of students reported that the quizzing procedure had helped them monitor their learning, improved their attendance and they reported they would prefer to adopt this procedure in the future (Leeming, 2002, see also Bangert-Drowns, Kulik & Kulik, 1991 for similar findings). One reason for this finding might be related to the fact that retrieval practice has been shown to increase one's metacognitive awareness of what has been learned or not (Kornell & Son, 2009).

Testing increases attention and attendance at lectures (Szpunar, Khan & Schacter, 2013; Pennebaker. Gosling & Ferrel, 2013). Intermixing tests during the lecture prevents mind wandering and reduces test anxiety (Agarwal, D'Antonio, Roediger, McDermott & McDaniel, 2014; Szpunar, Khan & Schacter, 2013). Including short, but several tests, during the lecture also means that each lecture will be broken down in sequences (Szpunar, Khan & Schacter, 2013), which might be closely linked to the concept of "chunking" in the memory literature. Testing can be considered as a type of formative assessment, both for the teacher and for the student (Roediger & Karpicke, 2006b). All of these results confirm that testing is not only about direct effects on memory; rather, testing potentiates future study activities (cf. Izawa, 1970; Thompson, Wenger & Bartling, 1978).

Unresolved issues

Despite the extensive number of published papers on the testing effect, few studies have directly examined the testing effect using actual course material during the progress of a course. In a recent review, Rawson and Dunlosky

(2011) listed 168 published experiments during the period from 2000 – 2010, but few of those used educational material integrated in the on-going course. To my knowledge, only four of those 168 experiments used authentic course material during the progress of a course, and none of them examined the immediate effects of testing as compared to study (Carpenter, Pashler & Cepeda., 2009; Kromman, Jensen & Ringsted, 2009; Leeming, 2002; McDaniel, Anderson, Derbish & Morisette, 2007). Even if those studies found support for a testing effect after a longer retention interval, one possibility could be that students in the testing group continued to practice testing while regulating their own study methods, or they devoted more study time to that course, as indicated by Leeming (2002). In Study I, we examined the testing effect immediately after the practice session as well as across time.

Even if some studies have confirmed the testing effect in children, they require replications to further identify the benefits and boundaries of these findings. Before arguing for a large-scale implementation of test-enhanced learning across schools, additional studies are required. For example, to the best of my knowledge, only two studies have been conducted with children in the fifth-grade (Lipko-Speed, Dunlosky & Rawson, 2014; Rohrer, Taylor & Sholar, 2010). Both of these studies found support for testing with feedback as superior compared to study, but Lipko-Speed et al. (2014) found that testing without feedback was comparable to restudy. Rohrer et al. (2010) found a testing effect for both practiced and transfer tasks as measured one day after practice, but they did not follow up regarding the durability of learning across time. Both studies assessed the final test at a rather short interval; one day and approximately one week after the initial learning session. As suggested by Dunslosky, Rawson, Mars et al. (2013), the durability of the learning intervention is crucial for the goal of education, to produce durable memories. No studies have examined the effects of testing among children by using mathematics. Mathematics is one of the standard topics covered during all school years. In the fifth-grade, the educational material becomes more complex and the students work to develop a deeper understanding of mathematics; they also engage in activities that require complex thought extending beyond the memorization of rules. Study II investigates the effect of test-enhanced learning for mathematical achievement in a sample of fifth graders.

Future studies should continue to explore test-enhanced learning in the classroom related to individual differences in working memory capacity (Dunlosky, Rawson, Marsh, Nathan & Willingham, 2013). Is test-enhanced learning an effective method for those suffering from low working memory? This is an important issue, as young children in particular, learn at different

degrees, and working memory capacity has been shown to be one important underlying factor for learning (Alloway & Alloway, 2010; Bayliss, Jarrold, Gunn & Baddeley, 2003; Hitch, Towse & Hutton, 2001; Jonsson, Wiklund-Hörnqvist, Nyroos & Börjesson, 2014) and a factor that influences scholastic performance as measured by the national curriculum assessment (Nyroos & Wiklund-Hörnqvist, 2012). As far as I know, this issue has only been investigated in two published studies (Brewer & Unsworth, 2012; Spitzer, 1939). We know from the literature that different learning methods can be more or less effective for individuals suffering from working memory impairment, and the teacher is important for providing correct guidance (Gathercole & Alloway, 2008). In Study I and II the influence of WMC in relation to performance is addressed.

As a rule of thumb, it is important that the findings from laboratory studies on the testing effect are replicated in authentic settings, especially if we want to transform the science of learning into educational practice, and if we want to develop recommendations for practice. This simply means that using educational material included in a course also means that some experimental control is lost. The other way around, given that this is considered as akin to an authentic educational situation, finding support for that test-enhanced learning promotes learning more than study also give rise to potential practical recommendations for teachers (Daniel, 2012).

Factors that influence the magnitude of the testing effect

The empirical findings mentioned above seem quite clear and straightforward but what are the critical factors involved? Identifying the potential boundary conditions that influence test-enhanced learning is not only a question of scientific interest, but it is also of practical relevance for the educational community, as it can lead to the development of more specific instructional guidelines. Based on previous research, mainly from the laboratory environment, some key factors have been identified. Those are as follows: test format, feedback, number of repetitions and retrieval success.

Test-format – does test-format matter?

Two different types of tests have commonly been used within the testing effect literature: production and recognition tests. Production tests require the subject to generate a response from memory without any available cues (i.e. short-answer test, fill-in-blank and essay questions). Recognition tests are those that require the subject to select a response from the information that is provided (i.e. multiple-choice questions and yes/no tests). Most

(laboratory) studies suggest that short-answer tests are more beneficial for retention when compared to multiple-choice tests, as retrieval requires more cognitive effort in the former modality (Butler & Roediger, 2007; Carpenter & DeLosh, 2006; Kang, McDermott & Roediger, 2007; McDaniel, Anderson, Derbish & Morrisette, 2007); however, recent research has started to question that argument by presenting some evidence that both short-answer and multiple-choice tests can be equally as effective for retention (McDermott, Agarwal, D'Antonio, Roediger & McDaniel, 2014; Smith & Karpicke, 2014).

In a recent study by McDermott, Agarwal, D'Antonio, Roediger, and McDaniel (2014), the results showed that multiple-choice quizzing can be equally as effective as short-answer tests to enhance students' performance at a later regularin-class unit exam (covering both short-answer and multiple-choice questions). Another recent study revealed the same outcome when the retention test was multiple-choice, but there was a significant advantage for the use of an intervening test that was designed with short-answers when compared to that designed with multiple-choice questions when the delay test was also in a short-answer format (Stenlund, Sundström & Jonsson, 2014). It should be noted that both studies included feedback, independent of test format, which could have influenced the results to some degree. This highlights an important issue for practical reasons in real educational settings, as multiple-choice questions are less time-consuming for teachers in terms of grading and administering as compared to short-answer tests. Although, given that feedback is presented, having short-answer questions with feedback is also relatively easy to apply without having the teacher grading as the student can correct their answers by themselves.

Feedback – why?

Many past studies have shown that a test without feedback is significantly more beneficial than restudy (Carpenter & DeLosh, 2006; Carpenter, 2009; Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a); however, providing feedback seems to further boost learning, which is of educational relevance (Agarwal, Karpicke, Kang, Roediger & McDermott, 2008; Brosvic, Epstein, Cook & Dihoff, 2005; Brosvic & Epstein, 2007; Butler, Karpicke & Roediger, 2008; Kang, McDermott & Roediger, 2007; Pashler, Cepeda, Wixted & Rohrer, 2005; Roediger & Butler, 2011). For example, taking a test without knowing the answer, and without any possibility to relearn it, seems counterproductive if the purpose of the test is to promote learning (see Fig.1 in Karpicke & Roediger, 2007 for an illustration).

Karpicke and Roediger (2010) found that when repeated testing was combined with feedback, the performance level (during learning as well as one week later) increased by approximately 29% when compared to testing without feedback. Feedback may serve two important functions. First, feedback may prevent retrieval failures from being repeated and improve knowledge. This is particularly important if tests are designed as short-answer questions, as feedback serves to ameliorate poor initial test performance. Second, including feedback in the form of correct answers prevents erroneous learning from occurring (Kulhavy, 1977; Roediger & Marsh, 2005). This is important when the test is designed as a multiple-choice test as some correct responses might be selected with low levels of confidence (Butler, Karpicke & Roediger, 2008) and rely more on guessing than on knowledge (Butler & Roediger, 2008).

Fazio, Huelser, Johnson and Marsh (2010) examined how different kinds of feedback influenced learning of non-fiction passages. Over the course of three experiments they had participants read passages followed by either no feedback, right/wrong feedback, or correct-answer feedback. The results revealed that providing feedback in the form of a correct answer (as compared to "right" or "wrong") was most beneficial (Fazio et al., 2010). This is in line with previous research that correct-answer feedback both facilitated learning and improved the retention of word-pairs one week later (Pashler, Cepeda, Wixted & Rohrer, 2005).

Number of repetitions – how much is enough?

As a rule of thumb "more is always better". Taking one single test improves performance at a later retention test relative to restudy (Roediger & Karpicke, 2006a), but taking two or more consecutive tests improves performance when compared to taking one single test (Carpenter, Pashler, Wixted & Vul, 2008; Hashimoto, Usui, Taira & Kojima, 2011; Rosburg, Johansson, Weigl & Mecklinger, 2014; Wheeler & Roediger, 1992). However, one pertinent question to ask when determining how test-enhanced learning should be optimized is essentially, how much is enough? The answer to that question is "it depends". Schedules of how to optimize students learning in terms of durability (i.e. retention) and efficacy (time-consuming) is also a question of design (Rawson & Dunlosky, 2011). Given that the students only take the intervening tests during one single session (e.g. the exam-a-day procedure; weekly quizzes), the number of repetitions becomes important. Given that students take the same intervening test repeatedly spaced across time (i.e. days), the number of re-learning sessions becomes important. The critical aspect here is retrieval success.

Successful (repeated) retrieval

Even if unsuccessful retrievals holds the potential to further enhance subsequent learning (Kornell, Hays & Bjork, 2009; Richland, Kornell & Kao, 2009), retrieval success has been suggested to be one important factor underlying the testing effect (Jang, Wixted, Pecher, Zeelenberg & Huber, 2012; Rawson & Dunlosky, 2011; Rowland & DeLosh, 2014; Vaughn & Rawson, 2011). Increased performance at the retention test has been found to be a function of more successful retrievals during the intervention (Pyc & Rawson, 2009; Karpicke & Roediger, 2007). Recent research has started to investigate this issue by manipulating the number of retrieval successes during the intervening phase (i.e. referred to as a criterion level) in relation to performance at a subsequent retention test (Rawson & Dunlosky, 2011; Vaughn & Rawson, 2011). Vaughn and Rawson (2011) found that students who practiced retrieval until the items were recalled four to five times versus only once, performed significantly better at the final test, as measured one week later. To revisit the question of "how much is enough?", Rawson and Dunlosky (2011) showed that in order to reach a criterionlevel of four correct recalls, this required the administration of approximately six to seven repeated tests with feedback. It should be noted that the factor "retrieval success" has received less attention in past studies, but more recent research has pointed out that retrieval success is a key component (Rowland & DeLosh, 2014; Jang et al. 2012). Rowland and DeLosh (2014) investigated the testing effect based on initial retrieval success (test versus study). Contrary to the typical interaction effect found in the testing effect literature, they also found support for a testing effect at the immediate test and this was replicated across three experiments- but only when including items that were successfully retrieved during the intervention (Rowland & DeLosh, 2014; see also Jang et al., 2012 for similar findings).

Theoretical explanations of the testing effect

While there are a substantial number of empirical studies supporting the testing effect, the theoretical explanations have lagged behind. Different explanations have been put forward to explain the testing effect. The benefits associated with retrieval practice have mainly been explained in terms of the mechanisms of retrieval (Morris, Bransford & Franks, 1977); however due to the wealth of studies examining the testing effect during the last few years, more specific explanations have emerged. The most prominent accounts that researchers adhere to can broadly be divided into two categories: one that suggests that the mnemonic benefits of testing are attributed to memory strength (Bjork & Bjork, 1992; Kornell, Bjork & Garcia, 2011; Pyc & Rawson, 2009) and another view holds that testing enhances semantic elaboration in

memory (Carpenter, 2009; 2011, Pyc & Rawson, 2010). Given that the explanations cannot be regarded as entirely distinct from one another, they are presented individually in the next sections.

The transfer-appropriate processing hypothesis (TAP)

One theoretical explanation for the testing effect is the transfer-appropriate processing hypothesis (TAP; Morris, Bransford & Franks, 1977), which posits that memory performance is enhanced because the cognitive processes involved during learning/encoding match the processes required during retrieval (i.e. when learning is evaluated). Within the testing effect literature, empirical support for this idea mainly comes from studies comparing test (retrieval) versus study (encoding) conditions, as those learning activities rely on different memory processes. Support also comes from studies examining how initial test format affects later performance, depending on the similarity between test format at the intervening and the final test. Evidence for the latter study design example is somewhat mixed. In one study, Duchastel and Nungester (1982) found that performance was higher at the final test when the test format matched the intervening test, as measured by both short-answer (SA) and multiple-choice (MC) questions. In another study, Kang, McDermott, and Roediger (2007) tested the same hypothesis by using SA and MC questions both during the intervention phase and during the final test. The results showed that taking SA questions during the intervention phase resulted in enhanced performance at a final test for both the MC and SA questions, as compared to taking an initial MC test, disclaiming the TAP hypothesis (Kang, McDermott & Roediger, 2007; see also Carpenter & DeLosh, 2006 for similar findings; but see; Stenlund, Sundström & Jonsson, 2014, for mixed findings).

The retrieval effort hypothesis

The *retrieval effort hypothesis* (Pyc & Rawson, 2009) basically states that more effortful retrievals are better for memory than less effortful retrievals. Empirically, this has been manipulated by varying test formats or item lag (Pyc & Rawson, 2009). Empirical support for this theoretical account has mostly stemmed from studies comparing production tests (e.g. short answer, cued recall, fill-in-blank) with recognition tests (e.g. multiple choice). The typical outcome is that production tests, which are defined as more effortful, also produce better retention than recognition tests, which are less effortful (Carpenter & DeLosh, 2006; Glover, 1989, but see McDermott, Agarwal, D'Antonio, Roediger & McDaniel, 2014, exp. 1a & 1b for contradictory findings). Another way of thinking about retrieval difficulty is to vary the lag between items during the intervening test. Pyc and Rawson (2009) had subjects learn a set of Swahili–English word-pairs. The level of difficulty was

manipulated by varying the interstimulus interval (ISI) between the items during the learning session. A long ISI was considered to be more difficult than a short ISI. Across two experiments, the researchers found support for that successful but difficult retrievals enhanced retention more than successful and easy retrievals (Pyc & Rawson, 2009). It might be valuable to mention that the retrieval effort hypothesis is a modified hypothesis based on the theoretical account introduced by Bjork (1994) – *the desirable difficulty framework* – which suggests that successful but difficult retrieval processing during encoding produces better retention than successful retrieval that is less difficult. This framework has recently been modified and a new model has been introduced: the bifurcation model (Kornell, Bjork & Garcia, 2011: see below).

Memory strength

According to Bjork and Bjork's (1992) "new theory of disuse", a memory representation is characterized by two different strengths: storage strength and retrieval strength. The critical aspect for long-term retention is storage strength. When an item is stored, it has some probability to be retrieved later. Retrieval strength is the temporary accessibility of information in the short-term, with less emphasis on long-term retention (Bjork & Bjork, 1992). The theory assumes that both encoding and retrieval events strengthen storage and retrieval strength, but that retrieval-based learning activities have more potential for strengthening up the memory representation when compared to encoding events (Bjork & Bjork, 1992). In terms of the testing effect literature, this would explain the typical outcome found that the study condition is superior at the immediate test compared to the testing condition, but with the reversed pattern after a delay (Roediger & Karpicke, 2006a; 2006b).

The Bifurcation model

An idea that is closely related to Bjork and Bjorks (1992) theory about differentiating between storage and retrieval strength is the bifurcation model recently proposed by Kornell, Bjork and Garcia (2011). The bifurcation model posits that items will be differentially distributed in terms of memory strength along a continuum (low-high) depending on how they are initially learned (study versus test). According to the model, tested items will be bifurcated in terms of low or high in memory strength depending on retrieval success. Tested items with high memory strength are those that are successfully recalled during practice. Tested items with low memory strength are those that were not retrieved during practice. In contrast, a comparable number of studied items will be moderate in memory strength. The rate of forgetting is the same for both conditions. Across time, a majority of the

studied items will fall below the threshold line, making them unrecallable at the retention test but the tested items that are high in memory strength will still remain above the threshold line, making them recallable (Kornell, Bjork & Garcia, 2011). The predictions made by the bifurcation model fit well when explaining the interaction effect typically viewed in the testing effect literature - specifically, that the study condition is superior at the immediate test, with the reversed pattern observed after a delay (Roediger & Karpicke 2006a; 2006b; Karpicke & Roediger, 2008).

The semantic elaboration view

The semantic elaboration view comprises two hypotheses: *the elaborative retrieval hypothesis* (Carpenter & DeLosh, 2006; Carpenter, 2009) and *the mediator effectiveness hypothesis* (Carpenter, 2011; Pyc & Rawson, 2010); both suggest that retrieval practice promotes semantic elaboration to a greater extent than study. The *elaborative retrieval hypothesis* posits that retrieval during memory search for a target also activates associated candidates related to the target (Carpenter, 2009). From this view, elaboration refers to the process of adding additional information related to a target when presented with a cue. The elaboration process supports the creation of new retrieval routes in memory, which facilitates later retrieval (Carpenter, 2009). The *mediator effectiveness hypothesis* suggests that semantic information is activated during repeated retrieval, which benefits later recall (Carpenter, 2011; Pyc & Rawson, 2010). More specifically, the *mediator effectiveness hypothesis* suggests that tests are more beneficial than study as they encourage the use of more effective mediators during encoding. Mediators are defined as a word, concept, or other semantic element that binds the target to the cue (Pyc & Rawson, 2010). To test this hypothesis, Pyc and Rawson (2010) had subjects learn Swahili–English word-pairs either by repeated studying or through a comparable number of cued recall tests, followed by restudying the word-pair again. During restudy, all subjects were asked to generate a mediator that would help them remember the target later on. The results showed that one week after the initial learning session, testing was superior when compared to study (i.e. the classic testing effect), but the results also revealed that the subjects engaged in testing were better at recalling the mediators they had been asked to generate along with the target word (Pyc & Rawson, 2010; see also Carpenter, 2011 for similar findings).

Encoding variability hypothesis

The encoding variability hypothesis suggests that memory improves as a function of increased encoding variability as compared to encoding constancy (Melton, 1970; Martin, 1968). A memory representation exposed

to a number of different "encodings" is associated with several different retrieval routes in memory, which makes more cues available at subsequent retrieval. The term variability is not clearly defined, and it can be related to several different types of variability such as *context* variability (Melton, 1970) and *referential* variability (Martin, 1968). Context variability (Melton, 1970) refers to the idea that if the encoded items are distributed across time, this will result in the availability of having more contextual cues, thereby increasing the likelihood of successful retrieval. Referential variability (Martin, 1968) is related to how the to-be-learned material is presented. For example, when two different semantic interpretations are presented for the same target (i.e. a concept) this allows the subject to encode the item in many different ways, as compared to re-exposure of the same question. One way to examine the encoding variability hypothesis in terms of referential variability is to re-formulate a question about the same piece of information during the intervening phase, which we did in Study II.

The encoding variability hypothesis would predict that taking a re-formulated question, as compared to taking the same question, would improve retention, as the former process produces additional retrieval routes in memory. Few studies have examined the encoding variability hypothesis within the testing effect literature, and among those that have, the results are mixed (Butler, 2010; McDaniel & Masson, 1985; McDermott, Agarwal, D'Antonio, Roediger & McDaniel, 2014, Exp 3). McDaniel and Masson (1985) had subjects learn concrete nouns either via a semantic or a phonemic encoding task. One group served as the control group, whereas the other group took an immediate cued recall test following the initial encoding session. One day later, all subjects took a cued recall test in which the target items were differentially cued with either semantic or phonemic cues compared to the initial encoding. Performance at the delayed test revealed that the test group was better than the control group when there was a mismatch between the cues in the encoding task and the final test, favoring the encoding variability hypothesis (McDaniel & Masson, 1985, but see Butler, 2010; McDermott, Agarwal, D'Antonio, Roediger & McDaniel, 2014, exp. 3 for contradictory findings).

Linking cognitive processes with the brain: functional imaging

Cognitive neuroscience can be seen as the interdisciplinary field for psychology and neuroscience research. Functional imaging techniques that study changes in brain activity following experimental manipulation have emerged as promising tools for the study of the neural correlates of many different cognitive activities. The most commonly used imaging technique is functional magnetic resonance imaging (fMRI), which is used in Study III.

FMRI offers an advantage, as it makes it possible to investigate how different cognitive processes and theoretical concepts are related to activity within the brain (see Cabeza & Nyberg, 2000b; D'Esposito, 2007; Poldrack, 2008). Consequently, fMRI results have provided a better understanding of the phenomenon under investigation and they have contributed to the further development of theoretical accounts (Cabeza & Nyberg, 2000b; D'Esposito, 2007).

The fMRI technique

FMRI is a non-invasive technique that is used to measure and map brain activity when individuals are engaged in some kind of task; this is performed to get a better understanding of how the brain works in vivo (see Huettel, Song & McCarthy, 2008 for an extensive review). It works by detecting changes in blood oxygenation following an experimental manipulation. The neurons within the human brain are central to the signal acquired by the applied magnetic gradients and electromagnetic pulses. Increased information processing in neurons will involve some metabolic changes, or requirements, and the increased neuronal activity will increase the blood flow. These changes or processes are energy consuming and the region(s) in the brain that are actively engaged in the task require oxygen, which is delivered by the haemoglobin due to the increase in blood flow. Oxygenated and deoxygenated blood differs in terms of their magnetic properties, which act in response to the magnetic field. Highly oxygenated blood gives a strong magnetic resonance (MR) signal, whereas less deoxygenated blood gives a weaker MR signal. The endogenous blood-oxygen-level-dependent (BOLD) signal provides information about the changes of oxygenated to deoxygenated haemoglobin caused by neural activity following an experimental manipulation.

The change in MR signal that is determined by the neural activity is referred to as the hemodynamic response (HDR). The HDR results from the decreased amount of deoxygenated haemoglobin within a voxel. The time course for the HDR in relation to the actual neuronal response is delayed and the first recognizable hemodynamic change is present approximately 1-2 seconds after the initial activity. In general, the overall peak of the HDR occurs around 4-6 seconds after stimulus onset, and it returns to baseline after 15-20 seconds. It is important to emphasize that the responses are always a relative measure of neural activity (Buckner, 1998; Poldrack, Mumford & Nichols, 2011).

Two common experimental designs used for fMRI are blocked design and event-related designs. In a blocked design, two or more conditions are

compared, and each trial is presented in a continuous manner; and each trial is considered as a single block and compared with the other condition. The blocked design offers good detection but it has some limitations with respect to timing and shape. In contrast, event-related design allows for the extracting of different neural processes within each trial and across trials, i.e. item-related processes (Cabeza & Nyberg, 2000b). Within each trial, several events can be of interest (e.g. retrieval versus feedback), and the different events are separated in time by the use of an ISI. The ISI refers to the time differences between the off-set of the first stimulus and the on-set of the next stimulus. Compared to the blocked design, event-related designs make use of the transient changes in activation evoked by stimuli, and the stimuli are often randomized and jittered (≈2-8sec). Event-related designs are considered as rather flexible, as they allow for the post-hoc sorting of events; for instance, make use of the subsequent memory paradigm (SMP), and events can be analyzed across runs (e.g. repetition) as well as over time (e.g. post-hoc sorting). As an example, within the SMP, brain activity during a single encoding session is analyzed according to whether the items are going to be remembered or forgotten on a subsequent memory test (Paller & Wagner, 2002). The higher activity observed for items that are subsequently remembered is commonly referred to as the subsequent memory effect (SME). Using the event-related design, different item categories (remembered versus forgotten) can then be contrasted and analyzed.

The goal of task fMRI studies is to detect changes in the signal that are related to the experimental manipulation, but these changes in the signal also incorporate unwanted noise. To reduce the variance that is not related to the experimental manipulation, the acquired images are commonly preprocessed in several steps before being entered into the statistical analysis (Huettel, Song & McCarthy, 2008; Poldrack, Mumford & Nichols, 2011). The following preprocessing steps are commonly performed: all images are corrected for the time differences due to slice-wise acquisition (slice-timing); then, corrections are made for movement artifacts across and within sessions for each subject (realigned and unwarped). In the next step, all images are transformed into a common template, which improves the validity when considering comparisons between subjects; specifically, this controls for potential individual differences in the brain's anatomy (normalized), and it finally controls for the remaining individual differences – a smoothed filter is applied to all images, increasing the sensitivity in the obtained signal (smoothed).

In Study III, functional data were preprocessed and analyzed using Statistical Parametric Mapping (SPM 8; The Wellcome Department of Cognitive Neurology, London, UK). One important issue in the analysis of

fMRI findings is to set an appropriate statistical threshold to correct for multiple comparisons (i.e. the false discovery problem). The common approach used in fMRI is to control for multiple comparisons by setting statistical thresholds at both the cluster and voxel levels (see Genovese, Lazar & Nichols, 2002; Poldrack, Mumford & Nichols, 2011; Worsley, Andermann, Coulis, MacDonald & Evans, 1999 for details). The appropriate statistical threshold should be considered in relation to the sensitivity of the analyses (Genovese, Lazar & Nichols, 2002).

As always, several statistical methods can be used to answer the question of interest. The "classical" fMRI-analysis approach simply asks: what regions are involved in condition A when compared to B? This is commonly analyzed by T- or F- statistics that compare the average regional differences in the BOLD-signal in regions of voxels. A rather novel way to examine these data is by using the representational similarity analysis approach (RSA; Kriegeskorte, Mur & Bandettini., 2008; Xue Dong, Chen, Lu, Mumford & Poldrack, 2010). With the RSA one can ask: what is the 'representational content' in condition A compared to B? For example, Xue et al. (2010) had subjects intentionally encode (i.e. study) words three consecutive times in a scanning session. Subsequently, all participants performed a memory test outside the scanner. Words were backsorted based on retrieval success and they were further analyzed with the RSA. The results revealed that the pattern of brain activity elicited by repetitions of the same word was more similar for words that were subsequently remembered compared to those that were forgotten (Xue et al., 2010). Those results were interpreted as representational consistency is functionally linked with subsequent memory success (Xue et al., 2010). In contrast to the classical approach, the RSA measure patterns of brain activity which is appropriate for the purpose of Study III. RSA is a multivariate pattern analysis.

In Study III, the RSA is based on regions of interest (ROI:s) which, in the current study, are defined as the brain regions that show overlap in activity at Day 1 and Day 7. For a pre-defined ROI, the activity in each voxel was correlated to activity in the same voxels across repetitions. The outcome offers a potentially more sensitive and fine-graded pattern instead of the average activity. Such information can disappear when averaging the results, as in the classical contrast analysis. As a rule of thumb, these two methods complement each other as they are informative to get a better understanding both to "where" in the brain the manipulation differs, and also in terms of "how" it differs.

As is the case for all methods, there are always some pros and cons. FMRI provides very good spatial resolution while its temporal resolution is less

specific. The fMRI technique has proved good test-retest reliability, both over the short- and long-term (Aron, Gluck & Poldrack, 2006). Moreover, fMRI has the potential to compare psychological theories that predict the same outcomes, but that differ in terms of the hypothesized mechanism explaining the outcomes. In contrast to behavioral studies, fMRI has the potential to better explain why testing is beneficial. Therefore, we used fMRI in Study III.

Empirical findings of the neural correlates of test-enhanced learning

Despite the fairly extensive number of empirical studies conducted on the testing effect at a behavioral level, the underlying neurocognitive mechanisms involved are not so well understood. To date, only seven published studies have explicitly examined aspects related to the testing effect using fMRI (Eriksson, Kalpouzos & Nyberg, 2011; Liu, Liang, Li & Reder, 2014; Keresztes, Kaiser, Kovács & Racsmány, 2013; Nelson, Arnold, Gilmore & McDermott, 2013; van den Broek, Takashima, Segers, Fernández & Verhoeven, 2013; Vestergren & Nyberg, 2014; Wing, Marsh & Cabeza, 2013). Results from those studies are to a great extent diverse, owing to differences primarily related to methodology (see Table 1 below). Three of the studies directly compared brain activity related to items studied or tested (Keresztes et al., 2013; van den Broek et al., 2013; Wing, Marsh & Cabeza, 2013), one study used an intermixed test/study paradigm (Liu et al., 2014) while one focused on the number of repeated successful retrievals (Eriksson, Kalpouzos & Nyberg, 2011) and two of the studies examined brain activity related to test-potentiated learning (TPE; Nelson et al., 2013; Vestergren & Nyberg, 2014).

Common across all studies was that they used verbal material in terms of word-pairs as the to-be-learned material, and they all used cued recall as the test format. As can be seen in Table 1, the scanning procedure and the experimental design differed, to some degree, between the studies. The differences in terms of "when" the scanning procedure was used and "how" learning had been manipulated hold implications for how the results can be interpreted (see Table 1). Learning is a process as well as a product. From this view, it is also important to remember that the testing effect is defined as an effect (i.e. product) produced by test-enhanced learning (i.e. a process). It would be naïve to not consider this in relation to imaging data. The differences found from the studies presented below is informative, as some focus on the use of the neuroimaging method on learning as an outcome, rather than as a process and vice versa. The studies will briefly be described in separate sections based on their focus, as presented above.

Table 1.

A summary of the published fMRI studies related to the testing effect

Study	Paradigm	Baseline	Intervention	Immediate test	Retention test	Analysis	Main finding	Conclusions
Wing et al. (2013)	TE	*** Intentional encoding: Rating semantic relatedness of 192 weakly related English word-pairs**	***One single session: half of the items tested/ half restudied**	No	24 hours	Backsorted items: Contrast analysis (SR vs SF) & Functional connectivity analysis	More activity for T items SR in: aHC, lateral temporal cortices, striatum & mPFC. T enhanced hippocampal connectivity with VLPFC, right IPC & midline regions.	Testing is more beneficial than restudying due to enhanced relational binding & elaboration of related semantic information.
van den Broeck et al. (2013)	TE	Intentional encoding followed by three additional encoding activities (write mnemonic associations, re-exposure by yes/no response, re-exposure for all items)	***Three repetitions: half of the items tested/ half restudied**	No	1 week	General contrast during practice & Backsorted items: Contrast analysis (SR vs SF) No rep. effects	T > S: LIFG, ventral striatum, midbrain areas. Higher activity in LIPL & left middle temporal areas during T but not during S was predictive for better recall on the final test.	Testing leads to semantic elaboration, selective strengthening of associations between cue & target & is more cognitively effortful as compared to study
Keresztes et al. (2013)	TE	Intentional encoding of all word-pairs once	6 cycles (half of the items tested/half restudied) Each cycle included: Test: 30 word-pairs Study: 30 word-pairs Feedback: 60 word-pairs (re-exposure)	***Yes** (half of the participants)	***1 week** (half of the participants)	Exploratory analysis & ROI: Updating network as mask (Areas based on n-back localizer task)	The benefit with T compared to S is due to differential activation patterns in the fronto-parietal regions. Prior S leads to decrease across time, T does not.	Temporally stable activations in brain regions related to cognitive control is the key underlying the testing effect.

Study	Paradigm	Baseline	Intervention	Immediate test	Retention test	Analysis	Main finding	Conclusions
Liu et al. (2014)	T/S	***Intentional encoding of all word pairs once**	***Test/ Restudy of all items. Repeated 2 times** (the intervention was defined as the immediate & retention test)	***Yes**	***≈ 20 min**	Exploratory analysis & ROI: based on SME regions (Kim, 2011)	Right PFC & PPC is predictive for SM, but only based on their activation during retrieval practice & not during initial S. Left striatum involved in re-learning from FB (i.e. re-study opportunity).	The recruitment of right PFC, PPC during retrieval practice (along with well-known SME regions) may be key regions that foster the TE. The striatal contribution indicates the importance of feedback for learning.
Eriksson et al. (2011)	RSR	Intentional encoding	Test/Study cycles intermixed Criterion level: minimum 80% correct	No	***1 day later** 1 week 5 months	Parametric mapping based on RSR during the intervention	RSR leads to higher brain activity in the ACC & decreased activity in fronto-parietal regions. The degree of ACC activity increase correlated with memory performance 5 months later.	RSR may operate at the systems level by enhancing the consolidation of memory representations. Decreased fronto-parietal activity reflects reduced demands of cognitive control.

Study	Paradigm	Baseline	Intervention	Immediate test	Retention test	Analysis	Main finding	Conclusions
Nelson et al. (2013)	TPE	***Intentional encoding of all word pairs once**	Test (42 word-pairs) Study (42 word-pairs) No re-exposure (42 word-pairs) ***TPE: restudy all 126 word-pairs once**	No	1 day later	Exploratory analysis & ROI analysis (based on outcome from the exploratory analysis)	T leads to a decrease in the frontal regions & an increase in parietal regions. Although comparing T with S at the TPE phase, activity in fronto-parietal regions was higher for items prior T.	Testing facilitates subsequent encoding by engagement of retrieval processes during the following study phase (i.e. TPE)
Vestergren & Nyberg (2014)	TPE	Day 1: Intentional encoding of all 126 word pairs x 5 times. Day 2: Intentional encoding of all 126 word pairs once.	Prior scanning: Half of the items tested/ half restudied once ***TPE: restudy all 120 word-pairs once**	Yes	No	Exploratory analysis & ROI analysis	Regardless of RS, items prior T was accompanied with higher activity in the anterior insula, IFG, & HC during subsequent encoding compared to items prior S.	Prior testing potentiates subsequent encoding, possibly due to increased access of semantic representations, which enhances deeper processing

Note. The overview of the experiment refers to Figure 1 (typical design of studies), with main findings and conclusions. The bold asterisks and text refers to when the scanning procedure was conducted. Abbreviations: ACC = anterior cingulate cortex; aHC = anterior hippocampus; FB = feedback; HC = hippocampus; IFG = inferior frontal gyrus; IPC = inferior parietal cortex; LIFG = left inferior frontal gyrus; mPFC = medial prefrontal cortex; PFC = prefrontal cortex; PPC = posterior parietal cortex; RS = retrieval success; RSR = repeated successful retrieval; ROI = regions of interest; S = study; SME = subsequent memory effect; SM = subsequent memory; SR = subsequently recalled; T = test/testing; TE: testing effect; TPE = test-potentiated learning; VLPFC = ventrolateral prefrontal cortex.

Studies focusing on the testing effect

The three studies that directly compared testing with study differed in respect to when the scanning procedure was used. Wing et al. (2013) scanned participants during baseline exposure and during the intervention, whereas both van den Broek et al. (2013) and Keresztes et al. (2013) scanned participants after the intervention phase. In addition, both Wing et al. (2013) and van den Broek et al. (2013) used a within-subjects designs whereas Keresztes et al. (2013) used a between-subjects design.

Wing, Marsh and Cabeza (2013) had participants learn weakly related English word-pairs during scanning. After initial encoding, half of the pairs were practiced through testing and the other half were restudied once (see Roediger & Karpicke, 2006a, Exp 1 for a similar design). The next day, all subjects conducted a final test (cued recall) of all of the items, and items were backsorted as subsequently remembered or forgotten based on how they were initially learned (test versus study). Words initially tested were significantly better retained (63%) compared to the restudied words (51%). An interaction effect akin to the testing effect paradigm showed that higher activity in the bilateral anterior HC, the left middle/inferior temporal gyrus, the left anterior cingulate gyrus and the striatum predicted later memory for the tested items but not for restudy. A follow-up analysis was done to further explore which regions covaried with the HC during successful testing. Increased coupling between the HC and VLPFC (IFG), medial PFC, right supramarginal gyrus, and left middle temporal gyrus was found for the tested items subsequently remembered. Wing, Marsh and Cabeza (2013) concluded that brain activity that fosters the testing effect might be related to relational binding and controlled semantic elaboration with additional support recruited from the HC, which acts as the interface with other regions involved in consolidation.

In contrast to the study by Wing et al. (2013), van den Broek and colleagues (2013) had subjects learn 100 Swahili-German word-pairs outside the scanner. Following baseline exposure, all participants underwent fMRI while being tested on half of the words and restudied the other half three consecutive times. One week later, all subjects returned for a final test outside the scanner and words were categorized based on retrieval success (i.e. remembered/forgotten) related to prior learning activity (test versus study). The behavioral data confirmed a significant testing effect ($p < .001$). General differences in brain activity for the test versus study conditions were found in the bilateral VLPFC (IFG) and the striatum. A main effect for tested items correctly recalled one week later showed significantly higher activity in the left superior medial/frontal gyrus, the left middle cingulate cortex, the

bilateral MTG and the IPL during testing compared to study. An interaction effect showed that the critical brain regions predictive for subsequent memory, but only for the testing condition, were identified as the left IPL (supramarginal/AG) and the MTG. van den Broek et al (2013) suggested that the advantage for testing, as compared to restudy, is due to the processes related to targeted semantic elaboration and to the selective strengthening of associations between retrieval cues and targets. Moreover, they concluded that testing, as compared to study, is mirrored by increased effortful cognitive control and modulation of memory strength, as indicated by the higher level of activity in the IFG and striatum. Unfortunately, they did not report any data related to repetition effects.

Prior to this thesis, Keresztes and colleagues (2013) is the only published study that had used fMRI to examine the neurocognitive processes involved during both the immediate and final retention test. In their study, they had participants learn 60 Swahili-German word pairs. Half of the items were repeatedly tested and half were restudied. Following the intervening phase, half of the participants performed a cued recall task in the MR scanner, and the other half of participants returned one week later for the same scanning procedure. The behavioral data confirmed that the tested items were better retained than the words that were restudied after one week (50 % versus 39%, respectively). Imaging data showed that within an updating network (i.e. sets of brain regions active during an updating task) and across time (immediately versus 1 week later), activity levels differed depending on prior learning. An interaction effect showed that within the ROIs, the pattern of activity differed for the study and test conditions across time. For the study condition, there was a significant decrease from the immediate to the delayed test in the left insula, the bilateral ACC, the left aPFC, the right superior parietal cortex and the right middle orbitofrontal. For the test condition, there were no significant differences in activity level in those regions across time. Moreover, when extending the analysis beyond the identified regions in the updating network in a whole-brain analysis, an interaction between condition and retention interval was found bilaterally over the IFG, indicating higher activity in this region for the testing condition one week later. Keresztes et al. (2013) concluded that the benefits of testing compared to study is characterized by the stabilization of activity levels in brain regions related to cognitive control.

Taken together, only Wing et al. (2013) reported activity in the HC, but that was the only experiment that scanned initially during the intervention, which thus makes sense. The engagement of the VLPFC (IFG) was reported by all three studies, which consequently found that there was higher activity in this region during testing. The IFG has been suggested to be involved in selective

semantic processing possibly reflecting a semantic working memory system (Martin & Chao, 2001). There was increased striatal activity during testing, which may reflect some executive demands and/or modulation of memory strength (Scimeca & Badre, 2012). Activity in left middle temporal cortex was also identified as being predictive of subsequent memory in two of the studies. The parietal contribution reported in all three studies identified this region as important for performance both at the immediate and at the delayed test, and this was only evident for testing and not study. Studies of memory retrieval have consequently reported increased levels of activity in the parietal region during successful retrieval compared to other trials (Cabeza, 2008; Cabeza, Ciarameli, Olson & Moscovitch, 2008; Seghier, 2013; Wagner, Shannon, Kahn & Buckner, 2005). Only one study used the neuroimaging technique at the delayed test one week later (Keresztes et al., 2013). According to their study, the stabilization of activity patterns in several fronto-parietal regions promoted learning from tests. No study reported repetition effects. Even if there are some methodological differences, all of the studies clearly pointed out that the fronto-temporo-parietal regions were more engaged during testing than study. Despite the fact that none of the studies examined repetition effects, the researchers emphasized more or less a semantic elaboration explanation for the beneficial effects of testing (see Table 1).

Studies focusing on test-potentiated encoding

Two studies have explicitly examined brain activity in relation to test-potentiated encoding (TPE; Nelson et al., 2013; Vestergren & Nyberg, 2014). Common among both experiments, participants intentionally encoded word-pairs and continued to learn them either by retrieval practice or by restudy, followed by a subsequent encoding phase; the studies ended with a final cued recall test. The response of interest was brain activity during the subsequent encoding phase, as related to prior learning activity (tested or studied). Simply put, are there any differences in brain activity during subsequent encoding depending on prior learning activity? The main difference in methodology was that Nelson et al. (2013) scanned participants both during the baseline exposure and during the subsequent encoding phase whereas Vestergren and Nyberg (2014) only scanned participants during the subsequent encoding phase (further differences between the studies can be seen in Table 1).

Nelson et al.'s (2013) experiment included four phases. In phase one, during the first scanning session, participants' intentionally encoded 126 weakly related word-pairs. In phase two, the word-pairs were divided into three conditions; 42 of them were tested, 42 were restudied and 42 word-pairs

were not shown again. In phase three, participants were again scanned while restudying all 126 word-pairs (i.e. test-potentiated encoding). In phase four, approximately one day after scanning, all participants performed a cued recall test including all 126 word-pairs. On a behavioral level, no significant differences were found at the final cued recall test between performance for the tested and studied items (p = 0.97), although, for the tested items, the data showed that performance improved from the intervening test (40%) to the final cued recall test (53%).

Nelson et al. (2013) reported imaging data based on five ROIs located in the left lateral PFC, lateral parietal cortex, and medial parietal cortex. First, contrasting brain activity between the baseline exposure and the subsequent encoding phase for the tested items revealed a significant decrease in the left dlPFC and aPFC, and a significant increase in the left pIPL/dorsal AG, precuneus and MCC. These findings were interpreted as frontal regions sensitive for repetition priming/suppression and the parietal regions sensitive for retrieval practice. Second, contrasting brain activity during the subsequent encoding phase related to encoding history (test or study) showed significantly higher activity for prior testing in the left dlPFC, pIPL/dorsal AG, precuneus and MCC as well as marginally higher activity, although not significant, in left aPFC compared to prior study.

In conclusion, testing leads to a decrease in activity in the frontal regions and to an increase in the parietal regions. When comparing testing with study at the subsequent encoding phase, activity in the same regions are higher following prior testing compared to study (Nelson et al., 2013). Moreover, Nelson et al. (2013) found a significant positive correlation for activity in the left pIPL/dorsal AG as a function of the amount of new learning that occurred during the subsequent encoding phase in the testing condition. They concluded that testing potentiates learning due to the engagement of retrieval processes during the subsequent study phase.

In the study by Vestergren and Nyberg (2014), participants intentionally encoded 120 Swahili–Swedish word pairs at two sessions separated by one day. Directly following the second encoding phase (day 2), all participants were tested on half of the items; they then restudied the other half once. Immediately after, all participants underwent a scanning session while they restudied all 120 word pairs once. After the scanning procedure was conducted, all participants performed a cued recall test outside the scanner. At a behavioral level, no significant differences were found at the final cued recall test between performances on the tested and studied items. In addition, for the tested items, data showed that participants' performance improved from the intervening test (43%) to the final cued recall test (54%).

Regarding the imaging data, first, contrasting brain activity during the subsequent encoding phase related to encoding history (test or study) showed that regions within the frontal and medial temporal lobes were generally more active for the tested items compared to the studied items, with the reversed pattern observed in the parietal and middle/lateral temporal lobes. More specifically, they observed higher activity in the bilateral anterior insula, inferior occipital gyrus, left HC, IFG, and right caudate nucleus for items previously tested compared to those studied. In contrast, lower activity for the prior tested items was found in several regions, including the left middle cingulate gyrus, right supramarginal gyrus, bilateral precuneus and bilateral middle temporal gyrus.

Second, for the following imaging analysis, all items were divided into different categories based on encoding history (test or study) and retrieval success (recalled or not recalled), both before scanning and after scanning (items recalled or not recalled). Of primary interest was whether unsuccessful retrievals of items during prior testing would potentiate subsequent encoding differently, as compared to those items that were successfully retrieved. Tested items, independent of prior retrieval success but subsequently recalled at the post-scan test showed significantly higher activity in the left anterior insula and IFG compared to the tested items not subsequently recalled as well as compared to the studied items, independent of retrieval success. In conclusion, testing potentiates subsequent encoding, and this was evident for both previously unsuccessfully and successfully recalled items. Vestergren and Nyberg (2014) concluded that TPE is characterized by processes that are related to controlled access of conceptual representations as indicated by IFG, and language related semantic processing is induced by the engagement of the anterior part of the insula.

To summarize, both studies found significant support for test-potentiated encoding, with common as well as diverse brain regions. Common for both studies was that higher activity in the VLPFC was found following test compared to study. Moreover, the behavioral data showed that none of the studies found significant differences between the tested or restudied words at the final cued recall test. Both studies reported improved performance between test one and two, which provides support for the cognitive accounts regarding indirect effects of testing (cf. Izawa, 1970; Thompson, Wenger & Bartling, 1978).

Most surprising is that despite the fact that some common regions were found, the magnitude of activity related to the learning condition differed between the studies. For example, while Nelson et al. (2013) identified left dlPFC, pIPL/dorsal AG, precuneus and MCC as being significantly more

active following testing during the subsequent encoding phase, Vestergren and Nyberg (2014) found the reversed pattern in some of those regions (e.g. precuneus, middle cingulate gyrus). Given that approximately the same regions appeared in both studies, this is indicative to conclude that those regions are differentially affected by prior learning activity (study or test). It should also be noted that while Nelson et al. (2013) focused on the parietal region, Vestergren and Nyberg (2014) focused on the insula. The differences in activity patterns are also a matter of differences in study design, and the two studies differed with regard to both the experimental design and the analysis. A recent meta-analysis identified increased activity in insula as an index for successful encoding (Kim, 2011). Based on the wealth of studies investigating memory, the idea that the parietal cortex is critical for successful retrieval is not new (Cabeza, 2008; Cabeza, Ciarameli, Olson & Moscovitch, 2008; Seghier, 2013; Nyberg & Cabeza, 2000; Spaniol, et al., 2009; Wagner, Shannon, Kahn & Buckner, 2005).

Studies focusing on an intermixed test/study design

The two studies that used an intermixed study/test design during the intervention cannot be solely compared with each other as they differ a lot with respect to their methodological designs. Common in both experiments was that they used word-pairs as stimuli material and that participants both practiced retrieval and restudied all of the word-pairs. The main differences between the studies were the scanning procedure used, as well as the number of test/study runs performed during the intervention.

Liu et al. (2014) investigated the testing effect by using an intermixed Test/Study paradigm (see Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a; for similar designs). Participants intentionally encoded 45 high-frequency Chinese word-pairs. Immediately after, participants were tested on all items followed by immediate re-exposure to the intact word-pair (i.e feedback). This procedure was done two consecutive times. The time lag used for the same word to appear on test one and two was separated by approximately 20 minutes. Based on previous research (Kim, 2011) Liu et al. (2014) first analyzed data in six pre-defined brain regions known as important for subsequent memory: bilateral PFC, PPC, and HC (SME; Kim, 2011) followed by a whole-brain analysis.

First, the brain regions that were significantly more active during initial study, and that would lead to subsequent retrieval success at test one were found in the bilateral HC, the left PFC and the PPC (but not right PFC or PPC). The brain regions significantly more active during successful retrieval at test one and which were also predictive for correct recall at test two were

the bilateral PFC, the right PPC and the left HC (but not right HC or left PPC). The general degree of activity in those regions (including right HC and left PPC) during successful retrieval at test one was higher when responses on test two were also correct, as compared to when the test two responses were incorrect. Interestingly, both the right PFC and PPC were significantly engaged during retrieval identified as crucial for later memory but not engaged during the initial study. To follow up on their analyses beyond the pre-defined SME regions, Liu et al. (2014) also conducted a whole-brain analysis. During the initial study, both the left middle frontal gyrus and cingulate cortex (BA 32) were significantly more active for words that would be successfully retrieved at test one when compared to those not retrieved. Additional regions beyond the pre-defined SME regions that were significantly involved in retrieval success during test one, and which were predictive for retrieval success at test two, were found in the bilateral IFG and temporal gyrus. The involvement of the temporal gyrus was explained in terms of elaboration of semantic memory representations that facilitates learning. In line with behavioral studies, feedback improved performance between test one (37%) and test two (57%) and this improvement was supported by increased activity in the left caudate and putamen during the feedback event.

The finding by Liu and colleagues (2014), in that the cingulate cortex was more active during the initial study for items subsequently retrieved at test corresponds with an earlier study by Eriksson, Kalpouzos and Nyberg (2011). The latter authors found that repeated successful retrieval during testing was related to increased activity in the right ACC. In contrast, activity in the right mid vlPFC and superior parietal cortex was lower as a function of the number of successful retrievals (Eriksson, Kalpouzos & Nyberg, 2011). Five months later, memory performance across subjects was positively correlated to the increase in ACC activity during the intervention. As an important note, in the Eriksson et al. (2010) study, the participants were scanned during one single test the day after repeated retrieval and not during retrieval practice.

In sum, despite the fact that both studies differed a lot in terms of experimental design, and given that the specific locations of the active regions differed, the contribution of the ACC/cingulate cortex and the right parietal region in those studies seems to be indicative for the same cognitive processes. The ACC has also been identified as being predictive for subsequent memory in two other studies contrasting test versus study (see Keresztez et al., 2013; Wing et al. 2013). Eriksson et al. (2011) found that activity in the ACC increased as a function of increased successful retrieval. In line with that, Liu et al. (2014) found that during the initial study, the cingulate cortex was significantly more active when the item was successfully

retrieved at test one, as compared to what was observed during a retrieval failure. An interesting, although marginally, overlap between the studies, was the parietal contribution. Eriksson et al. (2011) found that activity in the (superior) parietal cortex decreased as a function of the increased number of successful retrievals. Liu et al. (2014) found that the right PPC was engaged during successful retrieval at test one but it was not engaged during the initial study, indicating this region as sensitive, and nevertheless important for, retrieval practice effects (see Nelson et al., 2013 for related findings).

Unresolved issues

Repeated successful retrieval is a key factor that underlies the testing effect, but empirical evidence for "why" and "how" remains to be answered (Jang, et al., 2012; Rawson & Dunlosky, 2011; Rowland & DeLosh, 2014). Two of the psychological explanations that exist center around the idea that the memory representations are altered, but in terms of "how", the explanations diverge to some degree. One family of ideas suggests that the memory trace itself is strengthened. Another holds that elaboration and increased associative networks are what underlie the beneficial changes on memory.

In terms of brain activity, and based on prior imaging studies, this could be reflected either by representational consistency or representational variability. Representational consistency could be reflected by the reactivation of modality-specific cortices during successful retrieval (Danker & Anderson, 2010; Nyberg, Habib, McInstosh & Tulving, 2000). However, retrieval is an active process, and it is not necessarily a simple replay of the neural ensembles engaged during learning; this process may therefore strengthen representations via representational variability (Dudai & Eisenberg, 2004). To date, none of the imaging studies related to test-enhanced learning (as presented above) reported repetition effects, and none of them have explicitly examined the neurocognitive mechanisms involved in successful repeated retrieval practice per se.

Repeated encoding that leads to later memory success has been suggested to be characterized by greater neural pattern similarity across repetitions (Xue et al., 2010). This means that each encoding phase reactivates the same memory traces that were active prior, thus, strengthening the memory representation (Xue et al., 2010). Based on these results, Xue et al. (2010) concluded that "reactivation of the same neural pattern during initial learning, whether during repeated practice, memory consolidation, and/or memory retrieval, can enhance memory" (Xue et al., 2010, p. 100). But does it hold for repeated retrieval as suggested by Xue et al. (2010)?

The suggestion made by Xue et al. (2010) is interesting if we apply it to the testing effect paradigm. From the literature, we know that testing is superior when compared to study for retention. We know that test-enhanced learning relies on encoding by retrieving from memory. Study relies on encoding into memory. In that sense, does the brain process the to-be-learned material by reactivation of the same neural pattern across repetitions (consistency), as suggested by Xue et al. (2010)? Or does it accomplish this with more dissimilarity (variability)? In response to the suggested theoretical explanations presented above, two hypotheses is suggested. If repeated retrieval is characterized by (i) more neural pattern similarity across successful repeated tests (which governs long-term retention), this would emphasize that consistency and memory strength of the trace per se is what underlies the beneficial effects with testing; however, (ii) if less neural pattern similarity across successfully repeated tests governs long-term retention, this would emphasize that variability and elaboration are what underlie the beneficial effects with testing. This question is addressed in Study III.

The main objectives of the thesis

The main objectives of this thesis were to investigate the (neuro-) cognitive processes related to test-enhanced learning. One part of this thesis focuses on behavioral data in authentic educational settings using different materials included in the curriculum and across age-groups (Study I & II). In another part of this thesis, fMRI is used to identify the underlying neural correlates involved in test-enhanced learning and long-term retention (Study III).

Specific aims:

Study I:

To examine whether repeated testing with feedback benefits learning, as compared to restudy of key concepts in introductory psychology in an educational context in a sample of undergraduate students. Of additional interest was to investigate whether WMC could be seen as a predictor of performance and how this was related to learning condition.

Study II:

To examine how repeated testing with feedback benefits learning and transfer of mathematical learning in relation to the encoding variability hypothesis in a sample of fifth grade students. Of further interest was to investigate whether WMC as measured two years earlier could predict mathematical performance.

Study III:

To identify brain regions important for learning due to successful repeated testing and to investigate whether repeated testing is associated with retrieval variability or consistency.

Overview of the empirical studies

A schematic overview of the three different studies is presented in Table 2. Study I and II were behavioral studies conducted using computer-based tasks. Study III used the functional magnetic resonance imaging (fMRI) technique to answer the research question. Common to all studies was that written informed consent was obtained in accordance with the Declaration of Helsinki, and all studies were approved by the Regional Ethics Committee at Umeå University. Participation was voluntary.

Table 2.

A schematic overview of the studies in the thesis

Study	Design	Sample (N)	Material	Intervention	Immediate test	Delayed test
I	Behavioral: between-subjects design	Undergraduate students (83)	Introductory psychology: key concepts	Test with feedback versus Study	Yes	18 days 5 weeks
II	Behavioral: between-subjects design	Fifth-grade students (42)	Mathematics: mental arithmetic & number understanding	Testing with feedback: Same-Test versus Variable-Test	Yes	3 days 5 weeks
III	fMRI and behavioral: within-subjects design	Young adults (23)	Swahili–Swedish word pairs	Repeated testing	No	1 week

Study I

Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology*, *55*, 10-16. doi: 10.1111/sjop.12093

Aim

The purpose of this study was to examine whether repeated testing with feedback benefits learning compared to rereading of introductory psychology key-concepts during the progress of an on-going course. While we know from a wealth of cognitive research that individual differences in WMC are predictive of scholastic performance (Alloway & Alloway, 2010; Bayliss, Jarrold, Gunn, & Baddeley, 2003; Hitch, Towse & Hutton, 2001; Ullman, Almeida & Klingberg, 2014), the investigation of the benefits of testing as a function of individual differences in WMC in the literature has been ignored (Dunlosky, Rawson & Marsh, et al., 2013).

Participants

Eighty-three undergraduate students that were registered in a cognitive psychology course participated in the study. Their ages ranged from 19-44 years (M = 23.8, SD = 3.94). Participation was voluntary.

Materials and procedure

The materials used in Study I consisted of 57 key concepts from three topics in the assigned cognitive psychology curriculum. The participants were randomly assigned to either the repeated testing group with feedback (ST_{fb}, n =43) or the restudy group (SS, n =40). Each of the three learning occasions (one for each topic) included a learning phase followed by an immediate test (five-minute delay). Learning was assessed by means of a test at three different time points: immediate, an average of 18 days later, and at a five-week delay.

During the learning phase, each key concept was repeatedly presented six times. For the SS-group, a key concept was presented, and the instructions only required that the subjects studied it. For the ST_{fb}-group, a key concept with the keyword left out was presented and subjects were requested to type in the correct answer on a blank screen, and they were provided with correct-answer feedback.

In the immediate and delayed tests, participants were presented with a fact and were requested to type in the correct answer. No feedback was presented. The same questions were used as during the intervention. The software used for the experiment was E-prime 2.0 (Schneider, Eschman & Zuccolotto, 2002). All items were randomly presented in the center of the screen, both during the learning phase and during the following tests.

Working memory capacity (WMC) was assessed with the Automated Operation Span (Aospan; for full task details, see Unsworth, Heitz, Schrock, & Engle, 2005), which requires participants to remember a series of letters while performing a concurrent task (i.e. a complex working memory task). The Aospan shows good internal consistency (.78) and test–retest reliability (.83; Unsworth et al., 2005).

Results

In line with prior studies, the results revealed a significant testing effect. Contrary to most of the prior research, we also found an immediate effect of testing with feedback when compared to the restudy of key concepts in the curricula at the immediate test (see Fig 2).

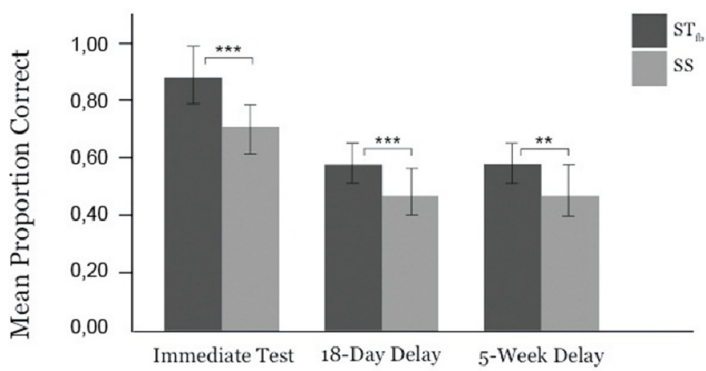

Figure 2. Mean proportion of correct responses for the STfb and SS group for the three time-points. Error bars represent ± 1 standard error of the mean. Reprinted with permission from *Scandinavian Journal of Psychology*.

A repeated one-way analysis of variance (ANOVA) yielded that students in the STfb group significantly improved their learning across the six repetitions (all *p*'s <.01). Regarding WMC, the effect of repeated testing was beneficial for students irrespectively of WMC.

Short discussion

Study I provided evidence for that repeated testing with feedback improves the knowledge of key concepts in introductory psychology during the progress of an on-going course compared to restudy. The testing with feedback group significantly outperformed the study group both over the short- and long term. Most forgetting in both groups occurred between the immediate test and the 18-day delay, and after that, the level of performance was quite stable. No influence of individual differences in WMC was found. In Study I, the intervention took place immediately after the topic lecture took place, which means that the students were also encouraged to have read the assigned chapter before the topic lecture. In spite of that, as indicated by the significant improvement across learning trials, the initial level of knowledge of the key concepts was low. The practical implications, based on the results from Study I, suggest that computer-based quizzes administered after a lecture could be used to improve students learning of important key concepts.

Study II

Wiklund-Hörnqvist, C., Jonsson, B., & Nyroos, M. Transfer in Mathematical Learning: A Comparison Study of Elementary School Children in an Educational Context. (Manuscript submitted for publication)

Aim

The aim of the study was to examine the encoding variability hypothesis by manipulating test-enhanced learning of mathematics. Two mathematical topics – number understanding and mental arithmetic – were identified as being the two mathematical topics in which Swedish fifth graders had the lowest performance levels one year earlier. Prior research has established that WMC is one influential factor that can predict children's scholastic performance (Alloway & Alloway, 2010; Bayliss, Jarrold, Gunn, & Baddeley, 2003; Hitch, Towse & Hutton, 2001). Of further interest was to examine whether WMC, as measured two years earlier, could predict math performance.

Participants

The sample consisted of 42 students in the fifth grade. They were 10–11 years old, and all students were registered at a Swedish public compulsory school. Participation was voluntary. All children were approved for participation, and their parents' informed consent was obtained. The teacher was

consulted to ensure that none of the students was identified as having some type of learning disability.

Materials and procedure

All mathematical tasks were adopted from the Swedish national tests in mathematics. Two different mathematical topics were examined: mental arithmetic and number understanding. The difficulty level for both topics was set in accordance with the knowledge requirements outlined in the mathematical syllabus. Encoding variability was manipulated in terms of re-formulating some of the questions in one group (Variable-Test group), versus keeping them constant in a second group (Same-Test group). Half of the students were assigned to the Variable-Test group (n = 21), and half to the Same-Test group (n= 21). As can be seen in Figure 3, each of the two training occasions (one for each topic) included a learning phase followed by an immediate test (five-minute delay). Learning was assessed by means of a test at three different time-points: immediate, 3-days later and at a five-week delay.

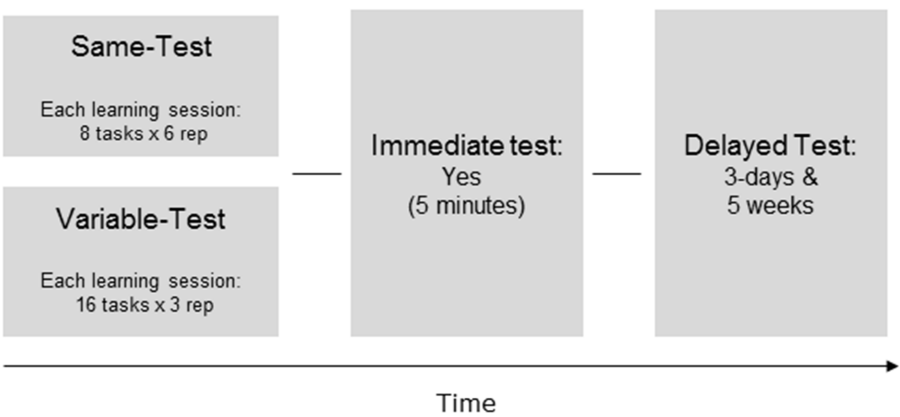

Figure 3. A schematic overview of Study II.

During the learning phase, a question appeared in the middle of the screen. Participants were asked to respond by choosing one answer among four multiple-choice alternatives, followed by immediate correct-answer feedback. The only difference between the groups was the number of times they repeated exactly the same question. For each topic, The Same-Test group practiced eight tasks six times. The Variable-Test group practiced the same eight tasks three times, as well as an additional eight isomorphic tasks three times.

At the immediate test, two types of tasks were administered for each topic: eight practiced and eight transfer tasks. The same procedure was used as during practice but, importantly, with the exception that the feedback slides were removed. At the delayed tests, the same tasks and procedure as at the immediate test were used. As all participants had been exposed to the transfer tasks during the immediate test, they could no longer be labeled as "transfer". Therefore, they were labeled as "prior transfer" at the two follow-up sessions (i.e. 3-days later, and at the 5-week delay). The instrumentation used to collect the data consisted of a personal response system, commonly called "clickers" (Liu & Stengel, 2011; Shapiro & Gordon, 2011). The index for WMC as measured two years earlier was based on three different tasks: block span (WAIS-III NI; Wechsler, 1999), digit span (WISC-IV; Wechsler, 2003), and listening span (Daneman & Carpenter, 1980; D'Amico & Guarnera, 2005).

Results

As can be seen in Fig. 4a (below), a significant interaction between task and group provided support for the encoding variability hypothesis at the immediate test. Children in the Variable-Test group performed better on the transfer tasks when compared to the Same-Test group. To follow-up the durability of learning across time (Fig 4b-c), two separate ANOVAS were conducted for each task and across time (3-days & 5-weeks delay). A significant interaction between the prior transfer tasks and group (see Fig 4c) showed that the Variable-Test group forgot more when compared to the Same-Test group, disclaiming the encoding variability hypothesis across time.

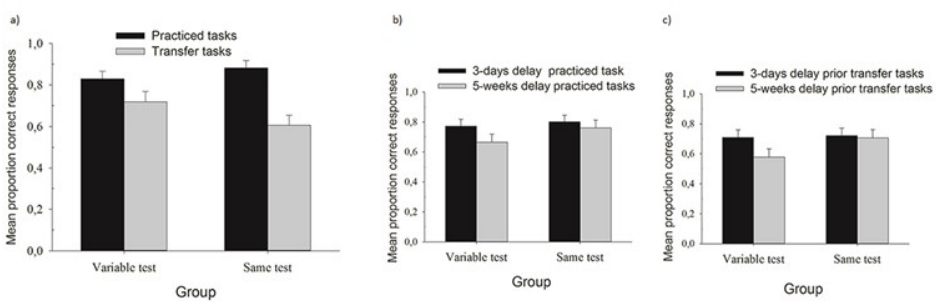

Figure 4. The mean proportion of correct responses for the Variable-Test and the Same-Test groups for each of the tasks. a) Immediate test b) practiced tasks at the follow-up sessions; and c) transfer tasks at the follow-up sessions.

Regarding WMC, the results showed that WMC explained 12% of the variance in the transfer tasks at the immediate test, with listening span as the only significant predictor (β =.54, p = .02). All other predictors were not significant for any of the mathematical tasks or time-points (all p > .05).

Short discussion

The results in Study II provided support for the encoding variability hypothesis at the immediate test, but not across time. The interaction found at the immediate test showed that the Variable-Test group performed better on the transfer tasks at the immediate test, while no such differences were found for the practiced tasks. When the durability of learning was examined, the initial support for the encoding variability hypothesis disappeared. An interaction effect showed that the Same-Test group remained stable in terms of performance for the prior transfer tasks, whereas the Variable-Test group gradually forgot the items over time. One possible factor that might influence the optimal design of a situation that produces durable learning might be related to individual differences in WMC. WMC was a significant predictor of the transfer tasks at the immediate test, indicating that more effortful retrieval might aid future retention. A combination of intensive practice of a few tasks together with effortful retrieval at the immediate test (i.e. transfer tasks) might induce elaboration during the intervention; this combined with retrieval effort might result in storage strength following the immediate test.

The findings from Study II suggest that for children in the fifth grade, lots of practice testing using a few tasks encourages durable learning when compared to the variable encoding of mathematics. The results also highlight the importance of follow up for the durability of learning. In sum, the results from Study II suggest that test-enhanced learning in a sample of children in the fifth grade, does not only improve learning, it also produces transfer to tasks that were never previously practiced.

Study III

Karlsson, L., Wiklund-Hörnqvist, C., Eriksson, J., Andersson, M., Jonsson, B. & Nyberg, L. Lesser Neural Pattern Similarity across Repeated Tests is Associated with Better Long-Term Memory Retention (submitted)

Aim

Based on behavioral studies identifying successful repeated retrieval as a key factor, Study III aimed to investigate the changes in brain activity associated with retrieval success both during repeated retrieval on Day 1 and one week later. The aim of Study III was to identify the brain regions that are

important for learning due to successful repeated testing – and to investigate whether repeated testing is associated with retrieval variability or consistency.

Participants

In response to advertisements at the local university, 23 volunteers applied to participate in the third study; the results of the analyses of two subsamples are reported in the study (n=18 and n=11). The inclusion criteria were as follows: age range 18-35 years old; right-handed; normal or corrected-to-normal vision; neurologically healthy; no reading difficulties, not pregnant; no metal in the body; no prior experience with the Swahili language and having Swedish as the native language. In addition, all participants were asked whether they approved that their imaging data (T1 images) would be examined by a physician.

Materials and procedure

In Study III, fMRI was used to collect brain imaging data. We used an event-related experimental design to extract item-related processes (e.g. retrieval success). Statistical analysis was performed using traditional conditional analysis and RSA (see the full article for details).

For a schematic overview of the study, see Fig 5 (below). Fig. 5a relates to the experimental procedure, while Fig. 5b illustrates the item categorization included in the statistical analysis (for specific details see the article). The material used was 60 Swahili-Swedish word-pairs adopted from Pyc and Rawson (2009) and Nelson and Dunlosky (1994). As can be seen in Fig. 5, the experiment was performed in three phases: at Day 1, a study-phase prior scanning was followed by a repeated test procedure within the MR environment. The experiment concluded with an additional fMRI session at Day 7 where subjects were tested on each item once. Fig. 5 depicts sample trials from all of the experimental stages. The experimental tasks were programmed and presented with E-prime 2.0 (Schneider, Eschman & Zuccolotto, 2002).

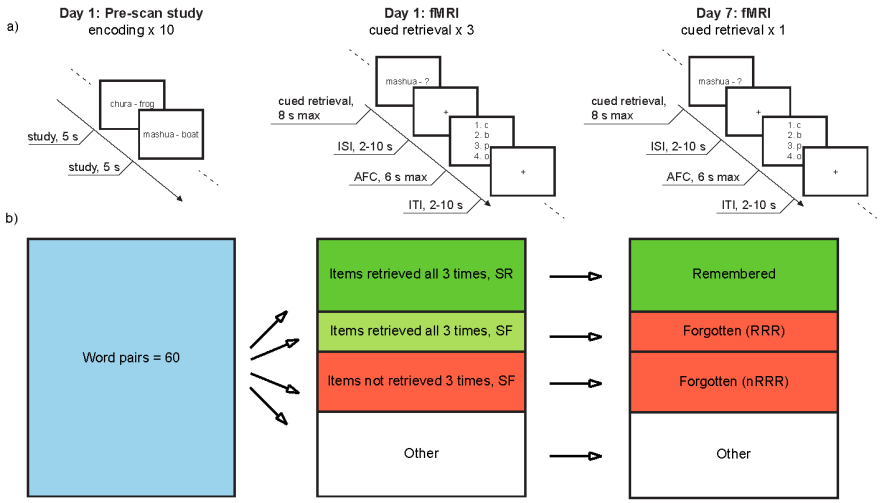

Figure 5. A schematic overview of Study III. a) The experimental procedure b) how items were sorted post-hoc into the different categories included in the analysis. Abbreviations: ISI = inter-stimulus interval; AFC = alternative forced choice; ITI = inter-trial interval; SR = subsequently recalled; SF subsequently forgotten; RRR = items retrieved all three times; nRRR = items not retrieved three times. Other = items not included in the other categories.

Day 1: Scanner

During the scanner session, each participant performed three repetitive test runs. The order of presentation of the items was uniquely randomly selected for each participant. The procedure for each event is illustrated in Fig. 5a. During testing, the Swahili word was shown for eight seconds, and within this time, participants were asked to respond by pressing a four-button key-pad with their fingers on the right hand to indicate if they encountered a word they "Knew was correct" (right index finger), "Believed was correct" (right middle finger) or if they "Did not retrieve a word" (right ring finger). To verify their responses, the participants were asked to reply the second letter in the Swedish word.

Day 7: Scanner and subsequent memory tasks

One week after the initial fMRI session, all participants returned for an additional fMRI session. Participants were again tested on the word-pairs, but only once, using the same procedure as in the initial scanning session (see Fig. 5). Following the scanning session, all subjects completed a post-paper confirmatory test where participants were asked to fill in the Swedish word of those they classified as remembering in the scanner.

Item classification included in the analysis

For the purpose of simplicity, I will present the item categories that were of interest in this study. Since the main focus was on successful repeated retrieval at Day 1, as well as on the effects of those items retrieved one week later, items were post hoc sorted into categories based on performance on Day 7. Items successfully and repeatedly retrieved three times on Day 1 are abbreviated as RRR. Some of those items would be forgotten at Day 7, and those were labeled as RRR_SF. Others would be correctly recalled at Day 7 and were labeled as RRR_SR. For the exact details about the different item categories, see Figure 5 or the original research article.

Results

Behavioral results

The validity of the responses that were judged as "Knew was correct", and which were given the correct second letter in the scanner, was highly correlated to the post-scan confirmatory test performance: $r = .99$, $p < .01$. Since the current study did not include a study condition, we could not statistically analyze the data in terms of a testing effect. However, inspection of the data revealed that 86% of the words correctly recalled at day 7 were those that were successfully repeatedly retrieved three times at Day 1, and an additional 11% of the words were judged as "Knew is correct" and given the correct second letter two times on Day 1, adding up to a total of 97%. These behavioral data confirm that successful repeated retrieval is critical for the testing effect.

Imaging results

The aim of Study III was to examine functional brain-activity responses that relate to the retention of repeated retrieval practice, as measured during repeated retrieval, as well as one week later. The results will be presented in the following order: first, I will report changes in activity as a function of long-term retention (Fig. 6a-b). Second, the RSA was used for the right posterior parietal cortex. Finally, an interaction effect for the repetition effects in relation to subsequent memory will be presented. Before we turn over to the results, I would like to remind the reader that the reported results of the differences in brain activity from Day 1 concerns items with the same behavioral outcome (i.e. RRR). The only difference is whether the items will be recalled (RRR_SR) or forgotten (RRR_SF) one week later.

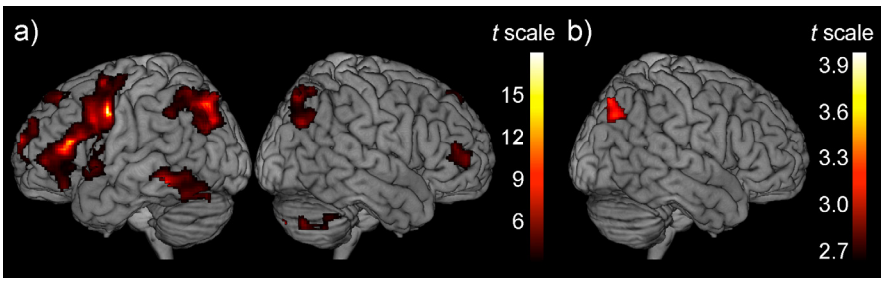

Figure 6. Day 7 and Day 1 BOLD responses reflecting long-term retention due to repeated testing. a) Day 7: Differences in BOLD brain activity for RRR_SR versus RRR_SF b) Day 1: BOLD-signal changes in relation to successful long-term retention (RRR_SR versus RRR_SF).

First, on Day 7, contrasting words subsequently recalled (RRR_SR) with those forgotten (RRR_SF) at Day 7 showed significantly higher BOLD brain activity in several regions including the bilateral posterior parietal cortex, the left: inferior temporal cortex, striatum, inferior, and superior prefrontal cortex, as well as the right cerebellum (see Fig. 6a). Next, when examining brain activity at Day 1 by contrasting the post-hoc sorted categories, RRR_SR and RRR_SF, across repetitions, significant outcomes were noted in two right parietal clusters that merged into one cluster at a more liberal statistical threshold (see Fig 6b). Activity in those regions was significantly higher for the RRR_SR items compared to the RRR_SF items already at Day 1, despite the same behavioral outcome (i.e. RRR). Moreover, those two clusters fell within the same right parietal region engaged at Day 7 (see Fig 6a) and they were therefore selected as the ROI for the RSA analysis.

Second, the RSA was used to investigate whether repeated testing was associated with retrieval variability or consistency. As can be seen in Fig. 7 (below), the results revealed that successful repeated retrieval, which lays the foundation for long-term retention, is characterized by low neural pattern similarity in the right posterior parietal region. What does this mean? It means that the average correlation in this region was lower for the items that were subsequently recalled (RRR_SR) when compared to the forgotten items (RRR_SF) and to those items not repeatedly recalled (nR) during Day 1. It is important to note that the lower correlation in the right posterior parietal cortex found for the items subsequently recalled compared to those forgotten does not mean that the brain activity level in this region was lower as indicated by the traditional BOLD analysis.

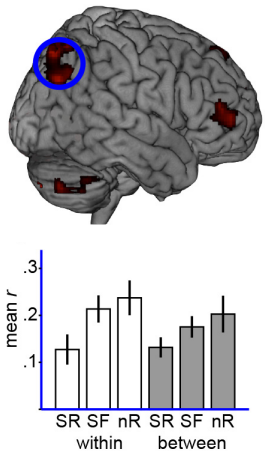

Figure 7. Repeated retrieval in relation to pattern similarity in the right posterior parietal ROI. At Day 1, within-items pattern similarity was lower for items subsequently recalled compared to those forgotten (RRR_SR vs RRR_SF).

Note. White bars = within items in that category (e.g. mashua-mashua-mashua); Grey bars = between items in that category (e.g. mashua-wingu-theluji).

Finally, as can be seen in Fig. 8a (below), a significant repetition by subsequent memory interaction was found in the left dlPFC. The functional brain activity level in this region differed with respect to whether the item was recalled or forgotten one week later (see Fig 8b). Again, despite the fact that both item categories (RRR_SR and RRR_SF) were successfully repeatedly retrieved three times, different pattern in activity levels were already evident in the left dlPFC at Day 1 based on retrieval success at Day 7.

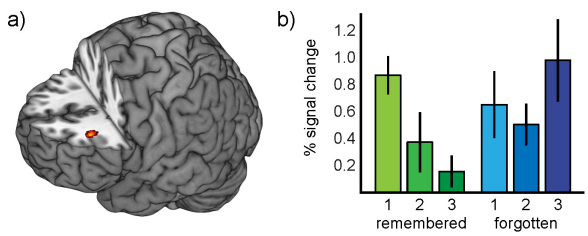

Figure 8. Day 1 BOLD-signal repetition effects in relation to successful long-term retention. a) A repetition x subsequent memory effect: left dlPFC. b) Changes in BOLD brain activity across repetitions.

Since both the left dlPFC and the right posterior parietal region were identified as important, we further analyzed Day 7 fMRI BOLD activity in those two regions for words that were retrieved faster compared to those that were retrieved more slowly (i.e. RRR_SR_fast versus RRR_SR_slow). The results showed lower dlPFC activity (p =.045) but higher parietal activity (a non-significant trend: p = .056) for faster successful retrieval, compared to slower responses which showed the reversed pattern.

Short discussion

To summarize, with respect to traditional BOLD-level analyses, we first identified one brain region that was particularly important for forming stable memories during repeated retrieval: the right superior parietal cortex. Next, with the RSA, we observed that repeated retrieval induced durable memory representations due to representational variability in this brain region; specifically, pattern similarity in the right parietal cortex across repeated retrieval was lower for retrieved words that were subsequently remembered, as compared to those that were forgotten. The reduced activity in left dlPFC across repetitions indicates less need of executive cognitive control processes for words that were to be successfully recalled one week after learning. At Day 7, higher activity in the parietal region and lower activity in the left dlPFC were found for faster responses compared to slower.

The results from Study III showed that the mnemonic benefits with test-enhanced learning for producing durable memories is influenced by multiple neurocognitive processes with different functional aspects that cooperate to establish the formation of a durable memory. The function of a specific region should always be considered within the neural network it is a part of. Based on the pattern similarity analysis, the variability of brain activity in the right posterior parietal region across repetitions can be a sign of semantic elaboration and the altering of possible representations in the brain. These cognitive operations are supported by a parallel process that encompasses the reduction of cognitive control, as evident by the reductions in activity in the left dlPFC across repetitions. As a consequence of these neurocognitive processes, the to-be-learned material was well consolidated one week later, and the memory has become semantic, as evident of the higher activity at Day 7 for words subsequently recalled in bilateral posterior parietal cortex, left inferior temporal cortex, striatum and left IFG. That interpretation was further confirmed by investigating the functional brain activity at Day 7 in relation to how well learned the words had been, as indicated by slow and fast response time (Keretzstes

, et al., 2013; MacLeod & Nelson, 1984; van den Broek, Takashima, Segers & Verhoven, 2013). In conclusion, test-enhanced learning in Study III provide evidence for the theoretical explanations that promotes semantic elaboration (Carpenter, 2011; Pyc, 2009).

General discussion

The main objectives of this thesis were to investigate whether beneficial effects of test-enhanced learning with feedback can be demonstrated in authentic educational contexts with the use of actual course materials (Study I and II), as well as to explain the underlying neuro-cognitive processes involved in successful test-enhanced learning (Study III). Further, the association between working memory capacity and the degree of learning was of particular interest (Study I and II).

The results confirmed the generalizability of test-enhanced learning in authentic settings when using course material during the progress of a course. These results were found for different kinds of materials: introductory psychology key concepts and mathematics, and across age cohorts: university students and children in the fifth grade. Furthermore, the influence of working memory capacity was related to age group. For children in the fifth grade, WMC was found to be a predictor of the transfer tasks immediately after practice. WMC did not predict performance for young adults. The neurocognitive processes involved during retrieval practice, which lays the foundation for the testing effect, have been suggested to be induced by increased semantic elaboration and reduced executive demands. These findings will be discussed and elaborated upon in relation to previous research. I will begin by discussing the behavioral studies (Study I and II); and I will then address the neuroimaging study (Study III), and conclude with a summary of all three studies.

Behavioral evidence for test-enhanced learning

Together, the results from both Study I and II provide empirical evidence that the mnemonic benefits of test-enhanced learning can be regarded as robust even in authentic settings. As referred to in the introduction, several factors and theoretical explanations have been put forward to explain the benefits associated with test-enhanced learning. I will begin by briefly highlighting some of the main findings reported in Study I and II, related to the theoretical explanations and factors suggested to influence the testing effect-paradigm.

In Study I, the aim was to investigate whether repeated testing with feedback benefits learning compared to restudy of introductory psychology key concepts within an educational context in a sample of undergraduate students. In Study I, we were able to investigate the testing effect, as we included a study condition that is commonly used as an active control group.

The results confirmed prior findings in that repeated testing with feedback is superior when compared to restudy in the long term. In Study I, despite the fact that the material used was integrated in the course, we found support for a testing effect by using computer-based tasks immediately after the lecture. Learning is a cumulative process, and the knowledge of key concepts is important for students in order to gain a better conceptual understanding of the subject at hand. To cite Willingham (2009) "factual knowledge must precede skill" (Willingham, 2009, p. 25); in fact, some subjective reflections were made during the data collection that corresponded very well with that citation by Willingham. Even though it was not an official research question, and there were no scientific results, several students reported that after they had learned the key concepts, they found it easier to read the chapter when they studied on their own. In that sense, repeated testing with feedback of central key concepts may have some indirect effects in terms of improving comprehension and one's further understanding of the concepts. That suggestion would be in line with prior research suggesting that knowledge about key concepts is effective for reading comprehension (e.g. McDaniel & Pressley, 1989; Beck, Perfettii & McKeown, 1982). In a society characterized by the rapid development of technology, supportive learning by computer-based tasks can easily be arranged by teachers (Linn, Gerard, Ryoo, McElhaney, Liu & Rafferty, 2014), and the students can easily take quizzes either on-line after the lecture or interleaved during the lecture (Freeman et al., 2014; Szpunar, Khan & Schacter, 2013; Pennebaker. Gosling & Ferrel, 2013).

Contrary to the typical interaction effect found in the testing effect literature, in Study I, we also found a significant testing effect at the immediate test. As far as I know, this is the first study that has demonstrated an immediate testing effect using authentic material included in the curricula during the progress of a course. Carpenter, Pashler, Wixted, and Vul (2008), as well as Kornell, Bjork and Garcia (2011), have shown immediate benefits of the testing effect but, unlike the present study (Study I), this was not achieved in a real educational context. The results in Study I for the immediate testing effect can best be explained in terms of the bifurcation model (Kornell, Bjork & Garcia, 2011). The bifurcation model posits that testing without feedback creates a bifurcated item distribution in which items that are retrieved are high in memory strength, and items that are not retrieved are low in memory strength. When participants are provided with feedback, memory strength becomes high enough to surpass a specific threshold and the information becomes recallable (Kornell et al., 2011). Studied items are moderate in terms of memory strength. The results from the immediate test in the current study are in line with the tenets of this model. The present study included correct answer feedback that was provided independent of whether

the response was correct or not, which possibly resulted in items being placed above the threshold at the immediate test (i.e. the items were recallable). It should be highlighted that the proposed term "memory strength" is rather vaguely defined, and the proposed model has not been so well investigated and replicated. More recent studies have found support for an immediate testing effect, but only when considering the factor retrieval success (Jang et al., 2012; Rowland & DeLosh, 2014). Across several experiments, Rowland and DeLosh (2014) found support for an immediate testing effect, but only when initial retrieval was successful or when feedback was provided to compensate for retrieval failures (see Jang et al., 2012 for similar findings). We did not analyze the data using an approach conditionalized (Rowland & DeLosh, 2014), but related to the improvement observed during the learning phase, most items were probably successfully retrieved.

Based on prior research and on the findings from Study I, specifically, that testing is superior compared to study, we decided to not include a study condition in Study II. Instead the aim of Study II was to examine how repeated testing with feedback benefits learning and transfer of mathematical learning in relation to the encoding variability hypothesis in a sample of fifth-grade students. The term variability in the current study was defined as practicing the same conceptual problem and solving procedure, but with several different tasks. It has been argued that encoding variability during repeated testing should facilitate transfer to a larger extent because the number of potential retrieval routes increases, and therefore facilitates correct retrieval (but see Butler, 2010).

In contrast to the results from the study by Butler (2010), our results showed a significant interaction, between task and group at the immediate test, which confirmed the encoding variability hypothesis (see McDaniel & Masson, 1985 for similar findings). That is, practicing the same amount of time, but with different tasks, produced more transfer when compared to just taking the same tasks. Why did our results differ from those reported by Butler? One possible explanation could be based on the experimental design. In Butler's (2010) study, the subjects only practiced three consecutive times, they did not include an immediate test, and all subjects returned one week later for a final test. Our results pertained to the immediate test, and, therefore the outcome might not be comparable; instead, our study is more similar to that of McDaniel and Masson (1985), in which subjects were tested one day later (i.e. relatively close in time from the intervention). In that sense, it is tempting to speculate that the beneficial effect of encoding variability has some temporal boundaries in terms of time or durability.

For the practiced tasks, the performance level at the immediate test did not reveal any significant differences between the groups, and the performance level was roughly equivalent, with the Same-Test group performing slightly better (mean proportion correct: .88 versus .83). In line with our results, a recent study by McDermott, Agarwal, D'Antonio, et al. (2014) also examined the effects of re-phrasing the questions during the intervention phase. The results showed that re-phrasing the questions did not affect the results and a testing effect was still present when compared to restudy (McDermott, Agarwal, D'Antonio, Roediger & McDaniel, 2014; see Butler, 2010 for related findings). Based on those results, together with ours from Study II, if the purpose is to learn something specific, the beneficial effects of test-enhanced learning might not be sensitive to how quizzes are initially formulated.

Dunlosky, Rawson, Marsh et al. (2013) stressed that it is important to evaluate the durability of learning across time. In line with that, we followed up by examining how durable the initial learning activity was in terms of retention three days and five weeks after the initial learning session. Interestingly, and a little bit surprisingly, a significant interaction was evident showing that performance between the groups differed with regard to the type of task or, more specifically, the prior transfer task. No differences were found for the practiced tasks. However, for the prior transfer tasks, the Variable-Test group performance level decreased (from .71 to .58) whereas the Same-Test group remained stable (from .72 to .71). Why did the Variable-Group continue to decrease in performance from the immediate test? One possible explanation could be framed in terms of Bjork and Bjorks (1992) 'new theory of disuse', which suggests that a memory representation is characterized by two different strengths: storage strength and retrieval strength. According to Bjork and Bjork (1992), the critical aspect for long-term retention is storage strength which, in our study, can be understood in terms of how well the mathematical principles had been acquired, as indicated by performance at the delayed tests. Retrieval strength is the temporary accessibility of information in the short-term, with less emphasis on long-term retention (Bjork & Bjork, 1992). In that case, why did the performance level in the Same-Test group remain stable at the delays, while participants performed worse at the immediate test?

In my view, this could be related to both retrieval effort and semantic processing. The retrieval effort hypothesis (Pyc & Rawson, 2009), posits that more effortful retrievals are better for memory than less effortful retrievals. As indicated by performance level and that WMC was found as a predictor, the transfer tasks were probably perceived as effortful at the immediate test, and in that sense, they might have influenced subsequent learning. We know from the literature that one single test improves performance at a later

retention test relative to restudy (Roediger & Karpicke, 2006a). Speculatively, the massive practice of fewer tasks (i.e. the learning phase) for the Same-Test group could potentially have made retrieval less cognitive effortful during the initial learning session, and instead, the participants could allocate more cognitive recourses to mathematical reasoning, which provides some support for the elaborative retrieval hypothesis (Carpenter, 2009). Consequently, at the immediate test, the more effortful retrievals related to the transfer tasks (in combination with elaboration during the intervention) might have influenced how well the to-be-learned material was retained, which provides some support for the retrieval effort hypothesis (Pyc & Rawson, 2009). It should be noted that several control analyses were performed to ensure that WMC did not differ between the groups; the influence of WMC was thus equal.

The theoretical explanations suggested above do not necessarily stand in conflict with one another. The testing effect is an effect (i.e. product) produced by test-enhanced learning (i.e. process). The theoretical explanations might complement each other by emphasizing different processes that lay the foundation for the testing effect. In my view, it seems plausible to suggest that both retrieval effort and memory elaboration contribute to the testing effect. Although the same tasks were used at the immediate test and at the delays, so the transfer tasks could no longer be treated as "pure transfer". Conversely, both groups had been exposed to the transfer tasks once, so the only difference encountered was due to the initial learning condition.

Regarding age, the majority of studies examining the testing effect have used university students. In Study I, we had undergraduate students participate. In Study II, we investigated the effects of test-enhanced learning in a sample of fifth-grade children. Even if there are some studies that have investigated the testing effect in both younger and older children, there is still a lack of sufficient evidence to draw conclusions regarding the boundaries associated with test-enhanced learning in a young population. Given that young children in this age group have not yet mastered efficient study strategies, the inclusion of test-enhanced learning as a regular school activity may not only improve learning but also teach children to regulate their own learning efficiently.

In Study II, we used mathematics. Within the testing effect literature, mathematics is woefully understudied; nevertheless important. Compared to other countries, Swedish children in the fifth grade receive a small number of tests during their compulsory education (Organisation for Economic Co-operation and Development - OECD, 2005; The Swedish National Agency for

Education, 2007). International comparisons like TIMSS have shown that a large number of young Swedish students also fail to achieve the national educational goals in mathematics (Mullis, Martin, Foy, et al., 2008). Consequently, the Swedish government has allocated many resources to help identify factors that can improve mathematical learning in terms of pedagogical activities and interventions. In light of this, and contrary to the suggestion made by Freeman et al. (2014), mathematical textbooks still serve as the primary basis for lessons in the fourth grade (Mullis, Martin, Foy, et al., 2008). Finding evidence-based pedagogical methods to improve learning is not only a question of scientific interest; it is definitively a question of interest for the educational community.

Results from Study I showed that repeated testing is beneficial when compared to restudy regardless of WMC. Furthermore, across time, the results in Study I provided no support for the "rich get richer" notion, which means that individuals with high WMC benefit most from testing with feedback. However, the sample consisted of a fairly homogenous group consisting of young adults. Therefore, the interpretations of these finding should be taken with caution. In Study II, despite the fact that WMC was measured two years earlier, we found that WMC predicted performance on the transfer tasks at the immediate test, but there were no differences between the test groups. That being said, those indications would also fit well within the interpretation of the theoretical explanations for the findings obtained in Study II. The results of our study are in line with those of prior studies, which showed that, particularly for children in this age, individual differences in WMC can predict scholastic performance (Alloway & Alloway, 2010; Bayliss, Jarrold, Gunn & Baddeley, 2003; Hitch, Towse & Hutton, 2001).

Interestingly, WMC was only found to be a significant predictor for performance on the transfer tasks, and also only at the immediate test; this was not found for any other tasks or sessions. Why? Speculatively, in our Study II, we did not have a study condition. Given that we had two test groups, we know that repeated testing is an active learning activity (regardless of test format). We know that testing is cognitively demanding because it forces you to retrieve information from memory, or at least to try to retrieve. We can probably be sure for that initial testing taxes our executive functions. In Study II, individual differences in WMC was related to performance on the transfer tasks, but there was no evidence for that influence regarding performance on the practiced tasks. Test-enhanced learning might be especially beneficial for those with lower WMC. Speculatively, if we would have had a study group, would WMC have predicted their performance across time and for all tasks? Perhaps, but,

unfortunately, I cannot say based on our results and study design. Future studies should definitely include measurements of WMC to best develop the most optimized design for producing learning. In sum, both Study I and II are unique, as they include measurements of WMC as an index for cognitive ability. This issue has, in some sense, been ignored but it has recently been identified as a "gap" in the current literature (Dunlosky, Rawson and Marsh et al. 2013, p. 35). It should be noted that we also provided evidence for that testing as a learning method can be easily applied with the support of computer-based tasks and with use of the clickers technique.

Neuro-cognitive evidence for test-enhanced learning

The results from Study III provides a unique contribution to the growing literature related to the testing effect phenomenon. This is the first study that reported repetition effects and how they are related to long-term retention, as measured one week later. In Study III, the aim was to investigate whether repeated testing was associated with retrieval variability or consistency, and also to identify the brain regions related to long-term retention as measured one week later.

As is clear from the literature, repeated successful retrieval is a key factor that lays the foundation for the testing effect. The main question was: why? In Study III, we focused on the effects of retrieval practice in terms of differentiating items that were repeatedly successfully retrieved in relation to whether they would be subsequently recalled or forgotten one week later. This was investigated in relation to functional BOLD brain activity. The results revealed that two key regions were identified as important in laying the foundation for long-term retention following test-enhanced learning: the left dlPFC and the right posterior parietal cortex. Those regions of interest will be discussed separately, and I will begin by discussing our findings in relation to prior imaging studies that have focused on the effects related to test-enhanced learning.

In Study III, the right posterior parietal cortex showed a significantly different activity pattern both at Day 1 and Day 7, which was dependent upon retrieval success in the long term. Both at Day 1 and 7, higher activity in this region was found for items that were successfully repeatedly retrieved and subsequently remembered at Day 7 compared to those that were subsequently forgotten. All of the three imaging studies that explicitly examined the testing effect reported higher activity in the right parietal region. Wing et al. (2013) found increased coupling between the HC and the right inferior parietal cortex during initial successful testing, and van den Broek et al. (2013) found significantly higher activity in a right inferior

parietal region (rather close to the region reported by Wing et al. 2013) for the tested items that would be successfully recalled one week later. Moreover, Keresztes et al. (2013) used a between-subjects design and was the only study that scanned participants at the delayed test. A significant interaction effect for learning condition and retention interval revealed that the pattern of activity in the right superior parietal cortex differed for the study and test condition across time (Keresztes et al., 2013). For the study condition, there was a significant decrease from the immediate to the delayed test but for the test condition, there were no significant differences in the activity levels in that region across time. These results correspond well with our findings, even if the specific location differed, high activity in the PPC during retrieval practice, as well as one week later, was found in Study III. The finding by Wing et al. (2013) indicated that the parietal region is recruited early in the acquisition phase and based on both van den Broek et al. (2013), Keresztes et al. (2013) and our finding, high activity in parietal cortex is important both in the short and long-term.

It should be highlighted that those studies (Keresztes et al., 2013; van den Broek et al., 2013; Wing et al. 2013), which focused on contrasting study and test, only scanned participants once either by a within-subjects or between-subjects design, whereas we used a within-subjects design that included two scanning sessions. In addition, the other studies had a study condition, which we did not have; despite that, in Study III, the right parietal region turned out to be an important region of interest. Liu et al. (2014) used an intermixed test/restudy design and the results showed that the right PPC was significantly more active during successful retrieval at test one when compared to those not retrieved, but the results also revealed that the right PPC was significantly more active for items successfully retrieved and subsequently recalled compared to those forgotten at test two (Liu et al., 2014). This provides further support for our findings.

Another finding by Liu et al. (2014) was that the right PPC was significantly engaged during successful retrieval and identified as crucial for later memory but, despite that, not engaged during the initial study (Liu et al., 2014). In the same vein, Vestergren and Nyberg (2014) found that the right inferior parietal cortex was significantly less active during subsequent encoding for items previously tested when compared to those prior studied. Despite that the specific location of the parietal regions did not overlap exactly, the region reported by Vestergren and Nyberg (2014) was more or less the same as Wing et al. (2013) reported as being significantly more active for tested items during successful retrieval when compared to study in their hippocampal connectivity interaction. The reason why this is interesting is because, as suggested by Nelson et al. (2013), the parietal

region might be sensitive for retrieval practice per se, and this might be one reason to why testing is superior when compared to study. The recruitment of parietal cortex initial during learning will lay the foundation for later retrieval success. According to our results and those of others that have explicitly studied the testing effect, the parietal region is a key region involved in the mnemonic benefits associated with testing. What is going on in this region?

In terms of the pattern similarity analysis, the results from Study III, contribute with novel information by showing that the important role of the right posterior parietal region is due to its greater neural pattern dissimilarity across successful repeated retrieval. How can we assume that? The pattern similarity in the right posterior parietal region was lower for words that were subsequently recalled when compared to words repeatedly recalled and forgotten, and those that were never repeatedly retrieved. This result indicates that repetition-induced variability of brain activity in this right posterior parietal region is characteristic of processes that leads to long-term retention, as measured one week later; this does not demonstrate consistency (Xue et al., 2010).

The retrieval-induced variability found in the present study is in line with neurocognitive accounts that suggest that the "semanticization" during memory formation occurs during multiple learning trials, which both enhances and facilitates the consolidation process (Henke, 2010). Intense retrieval practices can enhance the consolidation processes at several levels (i.e. synaptic and system levels), not only by stabilizing the neural representation of what has been learned, but also creating associations and retrieval links (Alberini, 2005; Dudai & Eisenburg, 2004; Lee 2009; Sara, 2000). It is therefore tempting to speculate that test-enhanced learning "speeds up" the neurocognitive processes by triggering the consolidation processes that are paralleled with recruiting regions that are known to be important in learning and long-term memory. Specifically, the angular gyrus (AG), as part of the parietal cluster, has been suggested to play a role in integrating different aspects of semantic information into larger units (Binder et al., 2009; Bonner, Peelle, Cook, & Grossman, 2013; Lau et al., 2008). The induced variability in our right parietal cluster might reflect an intense triggering of the consolidation processes that are recruiting AG as a cross-modal hub to integrate different aspects of semantic information by creating associations and retrieval links necessary for durable memory formation (Alberini, 2005, Binder, et al., 2009; Draganski, Gaser, Kempermann et al., 2006; Dudai & Eisenburg, 2004; Lau et al., 2008; Lee 2009; Seghier, 2013; Sara, 2000).

The neurocognitive accounts correspond well with the semantic elaboration view. The elaborative retrieval hypothesis posits that retrieval during memory search for a target also activates associated candidates related to the target (Carpenter, 2009). The difference between the present data and those of Xue et al. (2010) may reflect the functional heterogeneity between repeated encoding/study and retrieval/test.

To date, of all the imaging studies that have focused on the testing effect phenomenon, we are the only study that has examined how repeated retrieval across repetitions is related to functional changes in brain activity. As was evident in the repetition by subsequent memory interaction, another key region identified in Study III was the left dlPFC. Despite the same behavioral outcome Day 1 (RRR), the pattern of functional brain activity across repetitions was related to whether the words would be successfully remembered or forgotten one week later.

The dlPFC has been identified as a region that is involved in cognitive control processes, with higher activity in this region indicative of increased demands on retrieval selection mechanisms, such as the resolution of response conflicts (see e.g. Badre & Wagner, 2004). The left dlPFC is engaged as an "updating node" of newly learned material that is not yet stable (Johnson, Raye, Mitchell, Greene & Anderson, 2003). Specifically, the recruitment of the left dlPFC has been suggested as being particularly important during the initial encoding, as it plays a supervisory role in distributing and integrating information required for the formation of a long-lasting memory trace (Rossi, Innocenti, Polizotto et al., 2010). The proposed function of the left dlPFC corresponds well with the findings of our study with respect to parallel cognitive processes that occur in different regions- across repetitions there is less need for cognitive control processes, as is evident by reduced dlPFC activity. This is related to increased representation in the parietal region as shown by high levels of activity in this region both on Day 1 and Day 7 for words that were successfully recalled one week later. Conversely, words that were not subsequently recalled required the engagement of the left dlPFC over repetitions as demonstrated by the activity levels in the dlPFC across repetitions; and this is probably due to less integrated representations in the right parietal cortex.

Prior studies have found that AG is more active during the recollection of strong memories when compared to weaker memories (Kim, 2010; Seghier, 2013: Vilberg & Rugg, 2012), which fits well with the results from Day 7 in the current study. Additional regions that showed higher activity for words subsequently remembered on Day 7 were predominantly located in left lateralized regions, including the inferior temporal region, the prefrontal

cortex, and the caudate, as well as some right-sided regions including the cerebellum and the parietal region. Together, the higher degree of activity in those regions likely reflects higher cognitive effort during semantic retrieval one week after learning (Cabeza & Nyberg, 2000b; Hedden & Gabrieli, 2010; Scimeca & Badre, 2012). The inferior frontal gyrus (IFG) is involved in cognitive processes such as semantic processing (Bokde, Tagamets, Friedman & Horwitz, 2001) and item selection when several representations are available during retrieval (Badre & Wagner, 2007; Lau, Phillips & Poeppel, 2008). Two of the neuroimaging studies that explicitly investigated the testing effect reported significantly higher activity in left IFG for the tested items that were subsequently recalled (Keresztes et al., 2013; Wing et al., 2013). In addition, Vestergren and Nyberg (2014) found support for that higher activity in this region was predictive for tested items that were later remembered. Wing et al. (2013) found an increased coupling with the HC and IFG during initial testing, which was predictive for later remembering; this corresponds well with the finding by Liu et al. (2014), in that the bilateral IFG involved in retrieval success during test one was also predictive for retrieval success at test two (Liu et al., 2014).

The study by Keresztes et al. (2013) was the only one that scanned subjects one week after the initial learning took place. Extending their analysis to the whole brain revealed a significant interaction between learning condition and retention interval. Keresztes et al. (2013) found significantly higher activity in left IFG for the tested items that were subsequently recalled compared to the studied items that were recalled one week after learning. Those results correspond with ours despite the differences in design. In addition, the left IFG has been suggested as a "semantic working memory system" involved in semantic processing (Martin & Chao, 2001). Semantic processing, as evident by increased activity in the left IFG, is also partly dependent upon the recruitment of left-sided temporal and medial frontal regions (Boekde et al., 2001). Higher activity in the left temporal cortex indicates memory access that could potentially reactivate well consolidated memories (Habib & Nyberg, 2008).

In line with this, we also found higher activity in the left temporal cortices and a superior frontal region at Day 7, suggesting that words prior tested had become strong semantic representations in memory. We also found significantly higher activity in the striatum, and that region has been suggested to be involved in controlled semantic memory retrieval (Scimeca & Badre, 2012), and is well known as part of the executive working memory system. Indeed, both Wing et al. (2013) and van den Broek et al. (2013) reported higher activity in this region for testing compared to study. Moreover, Wing et al (2013) found support for that striatal activity predicted

subsequent retrieval success, but only for tested items and not for study indicating that testing "triggers" the fronto-striatal network early in the acquisition phase.

Together, our results reveal that several regions involved in the "semantic working memory system" (Martin & Chao, 2001, p. 198) are crucial for correct recall one week after test-enhanced learning. In contrast to the other studies, Study III is unique because we scanned the participants both at Day 1 and Day 7, which makes our results, together with those of other studies reliable.

A summary and reflections on the studies included in the thesis

The findings from Study I and II provide evidence for the broad applicability of test-enhanced learning. The results clearly point to the direction that our current knowledge about test-enhanced learning can be directly integrated within authentic classroom learning. Additional findings in Study II included the fact that WMC could be seen as a predictor for the transfer tasks at the immediate test, which clearly shows that individual differences in working memory do matter - especially with respect to children. These results are interesting and are related to the findings from Study III. To go back to the idea of the testing effect phenomenon, the aim and recommendation is to apply test-enhanced learning during *initial* learning because it improves one's ability to learn material over time. That is, we need to consider the temporal interval in terms of time. The results from Study III clearly point to the direction that initial testing is executive demanding, but it also triggers the enrichment of semantic representations in the brain. Those enriched representations will become stored in semantic storage regions and the representational concept is therefore accessible one week later, as evident by the activation of regions known to be involved in human long-term semantic memory. Of course, one week following retrieval practice, the executive demands are apparent, but together with the recruitment of semantic storage regions, that material is successfully retrieved/learned. We know from a wealth of studies that working memory is highly predictive of scholastic performance, and we also roughly know which brain regions working memory is comprised by. As demonstrated in Study III, retrieval practice will tax these regions, but as a consequence of retrieval practice, these practice sessions will transform information into a strong semantic memory, which increases the likelihood with which students will pass their courses and continue to build up their knowledge, especially as the courses become more advanced.

Another interesting point of speculation refers to "inside-outside the brain thinking". What do I mean by that? What I mean is that commonly, we think that applying re-formulated questions will affect the quality of learning in a positive direction. According to our results from Study II and III, this might not be the case. In Study II, we found support for encoding variability immediately after practice for the transfer tasks, but when we followed up learning 3-days and five weeks later, the Variable-Test group continued to decrease in performance compared to the Same-Test group. Why? This is what I meant by "inside-outside the brain thinking" According to the results in Study III, completing the same quizzes over and over again actually introduces variability in brain activity which then serves as the key for producing long-lasting memories. It is not how we design the questions that is important; what is important, is that we understand what happens in the brain, because it is the brain we use when we learn and when we are faced with situations that require the use of that specific information.

So to conclude, why should fMRI be bused for educational purposes? As evident from the results in Study III, fMRI has the potential to compare psychological theories that predict the same outcomes *but* differ in terms of the hypothesized mechanism that is used to explain that outcome. Moreover, fMRI can capture processes that are not evident in behavior, as illustrated in Study III. In that sense, using fMRI gave us the possibility to better explain *why* testing is beneficial. My personal recommendation is that testing should be included as a regular daily activity in school because it improves memory by enabling semantic elaboration in the brain. This, in turn, produces durable learning in terms of the accessibility to semantic representations stored in the brain. Use testing as a *tool to improve* learning, not only *of* learning.

Limitations and future directions

Below I will raise some limitations from the different studies. They will be raised together, and not one by one, as some of the issues are common across all. Finally, I will raise some thoughts for the future.

In Study I and II, we collected data in authentic settings, which simply means that there was some loose of experimental control. It is also important to differentiate between research as part of the course (e.g. influence on high stake tests, grades) and research that is conducted nominally in a classroom setting (which is what we did). The latter situation is much more similar to laboratory research; it has much the same experimental control, but in a somewhat different setting. In addition, in both Study I and II, we did control for some study habits, attitudes to testing, and judgment of learning,

but those are not reported here. In Study I, we did not have a pure testing condition. This would have been valuable to provide more ”pure” information about the testing effect by disentangling the testing effect from the feedback effect. It should be noted that studies that have examined the testing effect have not solely used testing alone during learning; rather they have alternated study and recall periods within a cycle (Karpicke & Roediger, 2007; Roediger & Karpicke, 2006a; 2006b). One cycle is commonly cited as one learning event. A cycle can be designed as study/test/study/test (=STST) which, in reality, means four repetitions within one cycle (Karpicke & Roediger, 2007; Roediger & Karpicke, 2006a).

In Study II, due to the fact that we collected data within the original schedule, it was not possible to randomly assign the children into the different groups; however, the teacher responsible for the children was consulted and confirmed that no individual or teaching differences were present. The sample size in Study II was small and it would have been preferable to have a larger sample both for increased statistical power and also to be able to draw some more generalized conclusions. In Study II, we did not correct the data to account for guesses in the statistical analysis; such an adjustment might be preferable when using MC questions. Hence, both groups were exposed for the MC test-format, so that this would not have changed the results. In Study II, it would have been informative to have a pure study condition to further evaluate how the different testing designs are related to a pure study condition in terms of being able to examine the testing effect per se. The number of participants available did not allow us to do that. In addition, as we replaced the original lecture, we also had some ethical issues to deal with related to the responsibility to provide methods that promotes learning.

Regarding WMC, given that both undergraduate students and children were included in Study I and II, the differences in the degree to which WMC influenced learning seemed quite clear and straightforward. In Study I, the undergraduate students were young adults who were rather high skilled, and they could not be regarded as representational for a more general population. It should be stressed that in Study II, despite the fact that WMC was measured two years earlier, WMC was shown as a predictor for performance. This result clearly indicates that this variable is an important one to consider, particularly among fifth-grade children. One weakness, for Study I and II was that different measurements were used to assess WMC in both studies, which, of course, limits the ability to further compare them. Therefore, no conclusions can be drawn; however, recommendations can be made. Future studies should address this issue more directly, as a primary

research question, because it can be informative when consider the most optimal design, for both research and educational practice.

In Study III, we used the fMRI technique to better understand what happens in the brain during repeated retrieval. Using such an advanced technology is, of course, related to some boundaries. One such limit was related to the pre-scan phase where we had participants study the word-pairs for ten rounds. Based on a pilot study this amount of pre-study was necessary to be able to conduct the analysis we wanted; but in light of the testing effect paradigm, it would have been preferable if we could have had the participants study in the scanner as well. For those familiar with the fMRI technique, you can certainly appreciate that it would not be possible (i.e. in terms of time) to engage in this type of trials also. For those of you that are not familiar with the fMRI technique, it is all about time. It would not be ethical to have participants laying in the scanner for such a long period of time. Of course, we cannot say anything about the testing effect per se because we did not include a study condition. However, in line with prior research, successful repeated retrieval is a key that underlies the testing effect. In that sense, as imaging studies within this paradigm are rare, it seems logical to begin by investigating those factors that have already been identified as important.

The number of participants, as well as the number of items, included in the analysis in Study III is also a weakness. Even if most of the participants fulfilled the criteria for some of the item categories of interest, they did not meet all of the criteria, which was necessary for the analysis. This is unfortunately the reality; we cannot manipulate the number of items that each subject correctly recalls. For those familiar with the fMRI technique, you all understand what a fine-graded analysis we have done. We decided to retain our strict inclusion criteria at the cost of the number of participants included. Including participants who did not fulfill the criteria would have resulted in unwanted noise in the data, which is not recommended. The other way around, a large sample size is of course preferable to be able to generalize the results and improve the statistical power. In fact, despite the small sample size, the effects were strong and robust and they can therefore be regarded as reliable. In sum, we do not claim that we can generalize our results; we do claim, however, that there is certainly an indication of the mnemonic benefits with repeated retrieval, but this requires further replication before more general conclusions can be made. What is clear is that using fMRI provides novel information that is not found in behavioral data.

One interesting aspect to consider for the future concerns whether intense learning per se results in changes in synaptic efficiency; this might underlie

changes in structural plasticity. Brain plasticity is a term that has become widely used within psychology and neuroscience, but it is not so easy to define. The term "plasticity" is often referred to as either structural and/or functional changes in the human adult brain (May, 2011), though there is less knowledge regarding the underlying factors that precede those changes (Zatorre, Fields & Johansen-Berg, 2012; but see Draganski, Gaser, Kempermann et al., 2006; Mårtensson, Eriksson, Bodammer et al., 2012 for examples). Related to that, we know that working memory is important for learning, and there is evidence for individual differences in WMC as predictive for scholastic performance (Alloway & Alloway, 2010; Bayliss, Jarrold, Gunn & Baddeley, 2003; Hitch, Towse & Hutton, 2001; Nyroos & Wiklund-Hörnqvist, 2012).

In line with the development of research tools (such as fMRI) it has become possible to further investigate the neural correlates of working memory. This has also shown that the underlying neural correlates of working memory, in some sense, may be task dependent, but also that the neural regions involved rather incorporate multiple cortical regions (e.g. prefrontal, parietal) and subcortical regions (e.g. striatum) and not a one-to-one mapping between structure and function. Those regions are also frequently reported in the imaging studies presented here with respect to test-enhanced learning. Thus, test-enhanced learning might be a valuable learning strategy for those suffering from working memory impairment (Pastötter, Weber & Bäuml, 2013; Sumowski, Leavitt, Cohen, Paxton, Chiaravalloti, & DeLuca, 2013). In that sense, test-enhanced learning can be a form of working memory training, but instead of providing general cognitive training (memory tasks per se), it concerns specific content that can be learned, and which is related to one's education. To conclude, findings from cognitive neuroscience may provide a significant possibility for both educators and students to develop efficient learning methods with a shared common goal: to foster durable learning.

Brain-based teaching – why?

As was evident from the studies included in this thesis, test-enhanced learning can be easily transferred into the classroom. It can be used with different materials and age cohorts, and unique information can be transformed from neuroscience data to practical recommendations. The findings from Study III, provided unique information and contributed to our understanding of the neurological basis of test-enhanced learning with both practical and theoretical implications. Understanding the complexity of the neurobiology that serves as the foundation for learning and memory in humans might be one of the most challenging and intriguing phenomena in

science (Mareschal, Butterworth & Tolmie, 2013). As evident from the results in Study III, our understanding of the brain is integral for our understanding of how learning takes place. In turn, knowledge about how memory and learning interacts will (hopefully) influence how learning activities are designed and applied by educators within an educational context (Freeman et al., 2014; Mareschal, Butterworth & Tolmie, 2013). Despite that, empirical findings from cognitive psychology (particularly memory and learning) have had less of an impact on the educational field (Freeman et al., 2014; Matlin, 2002; Newcombe, 2002). Educators might better aid students' learning by adopting learning activities that are based on empirical evidence (such as how the human brain works) in order to subsequently practice brain-based teaching in the classroom (Sigman, Peña, Goldin & Ribeiro, 2014; Freeman et al., 2014; Goswami, 2004; Maitlin, 2002). According to the Swedish National Agency for Education, it is of particular relevance that educational practices are based upon scientific evidence. This thesis contribute with that. I will end up by citing Goswami (2004): "educational and cognitive psychologists need to take the initiative, and think 'outside the box' about how current neuroscience techniques can help to answer educational questions" (Goswami, 2004, p.12). We did that. The results in this thesis provide behavioral and neuro-cognitive evidence for the power of test-enhanced learning.

References

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

Agarwal, P.K., D'Antonio, L.D. Roediger, H.L., McDermott, K. B. & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory and Cognition*, *3*, 131-139. doi: 10.1016/j.jarmac.2014.07.002

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

Alberini, C. M. (2005). Mechanisms of memory stabilization: are consolidation and reconsolidation similar or distinct processes? *Trends in Neurosciences Review*, *28*, 51-56. doi: 10.1016/j.tins.2004.11.001

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology, 106*, 20-29. doi:10.1016/j.jecp.2009.11.003

Aron, A.R., Gluck, M.A. & Poldrack, R.A. (2006). Long-term test–retest reliability of functional MRI in a classification learning task. *NeuroImage, 29*, 1000–1006. doi:10.1016/j.neuroimage.2005.08.010

Bacon, F. (2000). *Francis Bacon: The New Organon*. Cambridge, England: Cambridge University Press. (L. Jardine & M Silverthorne, transl. Original work published 1620)

Baddeley, A.D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*, 417-423. doi: 10.1016/S1364-6613(00)01538-2

Badre, D., & Wagner, A. D. (2007). Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia, 45*(13), 2883-2901. doi: 10.1016/j.neuropsychologia.2007.06.015

Badre, D., Wagner, A.D. (2004). Selection, integration, and Conflict Monitoring: Assessing the Nature and Generality of Prefrontal

Cognitive Control Mechanisms. *Neuron*, *41*, 473-478. doi: 10.1016/S0896-6273(03)00851-1

Bahrick, H.P. & Hall, L.K. (2005). The importance of retrieval failures to long-term retention:A metacognitive explanation of the spacing effect. *Journal of Memory and Language, 52*, 566-577. doi: 10.1016/j.jml.2005.01.012

Bangert-Drowns, R.L., Kulik, J.A. & Kulik, C.C. (1991). Effects of Frequent Classroom Testing. *Journal of Educational Research, 85,* 89–99.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4)*,* 612-637. doi: 10.1037/0033-2909.128.4.612

Bayliss, D. M., Jarrold, C., Gunn, D. M., & Baddeley, A. D. (2003). The Complexities of Complex Span: Explaining Individual Differences in Working Memory in Children and Adults. *Journal of Experimental Psychology: General*, 132(1), 71-92. doi: 10.1037/0096-3445.132.1.71

Beck, I.L., Perfetti, C.A., & McKeown, M.G. (1982). The effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology, 74*, 506–521.

Binder, J.R., Desai, R.H. (2011). The neurobiology of semantic memory *Trends in Cognitive Science, 15*, 527-536. doi: 10.1016/j.tics.2011.10.001

Binder, J.R., Desai, R.H., Graves, W.W. & Conant, L.L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*, 2767-2796. doi: 10.1093/cercor/bhp055

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

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.). *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35-67). Hillsdale, NJ: Erlbaum.

Bokde, A.L.W., Tagaments, M.-A., Friedman, R.B., & Horwitz, B. (2001). Functional interactions of the inferior frontal cortex during the processing of words and word-like stimuli. *Neuron*, *30*, 609–617. doi: 10.1016/S0896-6273(01)00288-4

Borst, J. P. & Anderson, J. R. (2013). Using model-based functional MRI to locate working memory updates and declarative memory retrievals in the fronto-parietal network. *PNAS*, *110*, xxx –xxx. doi: 10.1073/pnas.1221572110

Bonner, M.F., Peelle, J.E., Cook, P.A. & Grossman, M. (2013). Heteromodal conceptual processing in the angular gyrus. NeuroImage, 71, 175-186. doi. 10.1016./j.neuroimage.2013.01.006

Brewer, G. A. & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language, 66*, 407-415. doi:10.1016/j.jml.2011.21.09

Brosvic, G.M. & Epstein, M. L. (2007). Enhancing learning in the introductory course. *The Psychological Record, 57*, 391-408.

Brosvic, G. M., Epstein, M. L., Cook, M. J., & Dihoff, R. E. (2005). Efficacy of error for the correction of initially incorrect assumptions and of feedback for the affirmation of correct responding: Learning in the classroom. *The Psychological Record, 55,* 401–418.

Buckner, R. L. (1998). Event-Related fMRI and the Hemodynamic Response. *Human Brain Mapping*, *6*, 373-377. doi: 10.1002/(SICI)1097-0193(1998)6:5/6<373::AID-HBM8>3.0.CO;2-P

Buckner, R.L., Wheeler, M.E. Sheridan, M.A. (2001). Encoding Processes during Retrieval Tasks. *Journal of Cognitive Neurosciences*, *13*, 406-415. doi: 10.1162/08989290151137430

Butler, A. C. (2010). Repeated testing produces improved transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1118–1133. doi:10.1037/ a0019902

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback enhances retention of low confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *34*, 918-928. doi: 10.1037/0278-7393.34.4.918

Butler, A.C. & Roediger, H.L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36,* 604-616. doi: 10.3758/MC.36.3.604

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

Cabeza, R. (2008). Role of parietal regions in episodic memory retrieval: The dual attentional processes hypothesis. *Neuropsychologica*, *46*, 1813-1827. doi:10.1016/j.neuropsychologia.2008.03.019

Cabeza, R., Ciaramelli, E., Olson, I. R., & Moscovitch, M. (2008). The parietal cortex and episodic memory: An attentional account. *Nature Reviews Neuroscience, 9*, 613-625. doi:10.1038/nrn2459

Cabeza, R & Nyberg, L. (2000a). Neural basis of learning and memory: functional neuroimaging evidence. *Current opinion in Neurology*, *13*, 415-421.

Cabeza, R & Nyberg, L. (2000b). Imaging Cognition II: An Empirical Review of 275 PET and fMRI Studies. *Journal of Cognitive Neuroscience, 12*, 1-47. doi: 10.1162/08989290051137585

Carpenter, S.K. & Kelly, J.W. (2012). Tests enhance retention and transfer of spatial learning. *Psychological Bulletin Review, 19*, 443-448. doi: 10.3758/s13423-012-0221-2

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: Learning, Memory, and Cognition, 37*, 1547–1552. doi:10.1037/a0024140

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory & Cognition, 35*, 1563–1569. doi: 10.1037/a0017021

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology, 23*, 760–771. doi:10.1002/acp.1507

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

Craik, F.I.M. & Lockhart, R. S. (1972). Levels of Processing: A Framework for Memory Research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684.

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.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal learning and Verbal Behavior, 19*, 450-466.

D'Amico, A., & Guarnera, M. (2005). Exploring working memory in children with low arithmetical achievement. *Learning and Individual Differences, 15*, 189-202.

Daniel, D.B. (2012). Promising principles: Translating the science of learning to educational practice. *Journal of Applied Research in Memory and Cognition, 1*, 251-253. doi: 10.1016/j.jarmac.2012.10.004

Danker, J. F. & Anderson, J. R. (2010). The ghosts of brain states past: Remembering reactivates the brain regions engaged during encoding. *Psychological Bulletin, 136*, 87-102. doi: 10.1037/a0017937

D'Esposito, M. (2007). From cognitive to neural models of working memory. *Philosophical Transactions of the Royal Society, 362*, 761-772. doi: 10.1098/rstb.2007.2086

Draganski, B. Gaser, C, Kempermann, G., Kuhn, H.G., Winkler, J., Büchel, C. & May, A. (2006). Temporal and Spatial Dynamics of Brain Structure Changes during Extensive Learning. *The Journal of Neuroscience*, *26*, 6314-6317. doi: 10.1523/JNEUROSCI.4628-05.2006

Duchastel, P. C. & Nungester, R. J. (1982). Testing effects measured with alternate test forms. *Journal of Educational Research*, *75*, 309-313.

Dudai, Y. (2004). The Neurobiology of Consolidation, Or, How Stable is the Engram*? Annual Review of Psychology, 55*, 51-86. doi: 10.1146/annurev.psych.55.090902.142050

Dudai, Y. & Eisenberg, M. (2004). Rites of passage of the engram: reconsolidation and the lingering consolidation hypothesis. *Neuron*, *44*, 93-100. doi: 10.1016/j.neuron.2004.09.003

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. *Psychological Science in the Public Interest, 14*(1), 4-58. doi: 10.1177/1529100612453266

Eichenbaum, H.B., Yonelinas, A.P. & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, *30*, 123-152. doi: 10.1146/annurev.neuro.30.051606.094328

Eriksson, E., Kalpouzos, G. & Nyberg, L. (2011). Rewiring the brain with repeated retrieval: A parametric fMRI study of the testing effect. *Neuroscience Letters, 505, 36-40*. doi:10.1016/j.neulet.2011.08.061

Fazio, L. K., Huelser, B. J., Johnson, A., & Marsh, E. J. (2010). Receiving right/wrong feedback: Consequences for learning. *Memory, 18*, 335–350. doi:10.1080/09658211003652491

Freeman, S., Eddy, S. L. McDonough, M., Smith, M.K., Okoroafor, N., Jordt, H & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*, 8410-8415. doi: 10.1073/pnas.1319030111

Gathercole, S.E. & Alloway, T.P. (2008). *Working memory and Learning*. *A Practical Guide for Teachers*. Sage Press, London.

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

Genovese, C.R., Lazar, N.A. & Nichols, T. (2002). Tresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate. *NeuroImage*, *15*, 870-878. doi: 10.1016/nimg.2001.1037

Gholson, B., Dattel, A. R., Morgan, D., & Eymard, L. A. (1989). Problem solving, recall, and mapping relations in isomorphic transfer and nonisomorphic transfer among preschoolers and elementary school children. *Child Development*, *60*, 1172-1187. doi: 10.2307/1130791

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

Gluck, M.A., Mercado, E., Myers C.E. (2008). (Eds) Learning and Memory. From Brain to Behavior (1th Ed., p. 39). Worth Publishers, New York.

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L. & Tabbers, H. K. (2014). The effect of retrieval practice in primary school vocabulary learning. *Applied Cognitive Psychology, 28*, 135-142. doi:10.1002/acp.2956

Goswami, U. (2004). Annual Review. Neuroscience and education. *British Journal of Educational Psychology, 74*, 1-14.

Habib, R. & Nyberg, L. (2008). Neural Correlates of Availability and Accessibility in Memory, *Cerebral Cortex*, *18*, 17201726. doi:10.1093/cercor/bhm201

Halpern, D. F., & Hakel, M. D. (2002). Learning that lasts a lifetime: Teaching for long-term retention and transfer. *New Directions for Teaching and Learning*, *89*, 3-7. doi: 10.1002/tl.42

Hannula, D. E. & Ranganath, C. (2008). Medial temporal lobe activity predicts successful relational memory binding. *The Journal of Neuroscience*, *28*, 116-124. doi: 10.1523/JNEUROSCI.3086-07.2008

Hashimoto, T., Usui, N., Taira, M., & Kojima, S. (2010).Neural enhancement and attenuation induced by repetitive recall, *Neurobiology of Learning and Memory*, *96*, 143-149. doi: 10.1016/j.nlm.2011.03.008

Hattikudur, S. & Postle, B.R. (2011). Effects of Test-Enhanced Learning in a Cognitive Psychology Course. *Journal of Behavioral and Neuroscience Research*, *9*, 151-157.

Hedden, T. & Gabrieli, J.D.E. (2010). Shared and selective neural correlates of inhibition, facilitation, and shifting processes during executive control. *NeuroImage*, *51*, 421-431. doi: 10.1016/j.neuroimage.2010.01.089

Henke, K. (2010). A model for memory systems based on processing modes rather than consciousnessNature Reviews Neuroscience, 11, 523-532- doi: 10.1038/nrn2850

Hitch, G.J., Towse, J.N. & Hutton, U. (2001). What Limits Childrens' Working Memory Span? Theoretical Accounts and Applications for Scholastic Development. *Journal of Experimental Psychology: General, 130,* 184-198. doi: 10.1037//0096-3445.130.2.184

Huettel, S.A., Song, A.W. & McCarthy, G. (2008). *Functional Magnetic Resonance Imaging* (2nd Ed), Sunderland, USA; Sinauer Associates Inc.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83*, 340–344.

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

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability. *Quarterly Journal of Experimental Psychology, 65,* 962-975. doi: 10.1080/17470218.2011.638079

James,W. (1890). *The principles of psychology* (Vols. 1 & 2). New York: *Holt*.

Johnson, M.K., Raye, C.L., Mitchell, K.J., Greene, E.J. & Anderson, (2003). FMRI evidence for an Organization of Prefrontal Cortex by Both Type of Processes and Type of Information. *Cerebral Cortex*, *13*, 265-273. doi: 10.1093/cercor/13.3.265

Jonsson, B. Wiklund-Hörnqvist, C, Nyroos, M. & Börjesson A. (2014). Self-reported memory strategies and their relationship to immediate and delayed text recall and working memory capacity *Education Inquiry*, *5*, 22850, doi: 10.3402/edui.v5.22850

Kang, S. H. K., McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review, 18*, 998–1005. doi: 10.3758/s13423-011-0113-x

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. (2014). Learning With Retrieval-Based Concept Mapping. *Journal of Educational Psychology*, *106*, 849–85. doi: 10.1037/a0035934

Karpicke, J. D., Blunt, J. R., Smith, M.A. & Karpicke, S.S. (in press). Retrieval based learning: The need for guided retrieval in elementary school children. Journal of Applied Research in Memory and Cognition. 2014. doi: 10.1016/j.jarmac.2014.07.008

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. *Science, 331*(6018), 772-775. doi: 10.1126/science.1199327

Karpicke, J.D., Butler, A.C. & Roediger, H.L. (2009). Metacognitive Strategies in Student Learning: Do Students Practice Retrieval When Study on Their Own? *Memory, 17*, 471-479. doi: 10.1080/09658210802647009

Karpicke, J.D. & Roediger,H.L. (2008). The Critical Importance of Retrieval for Learning, *Science*, *319*, 266. doi: 10.1126/science.1152408

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

Keresztes, A., Kaiser, D., Kovács, G. & Racsmány, M. (2013). Testing Promotes Long-Term Learning via Stabilizing Activation Patterns in a Large Network of Brain Areas. *Cerebral Cortex*, *24*, 3025-3035. doi: 10.1093/cercor/bht158

Kim, A.S.N. (2011). Neural activity that predicts subsequent memory and forgetting: a meta-analysis of 74 fMRI studies. *NeuroImage, 54*, 2446–2461. doi: 10.1016/j.neuroimage.2010.09.045

Kim, H. (2010). Dissociating the roles of the default-mode, dorsal, and ventral networks in episodic memory retrieval. *NeuroImage*, *50*, 1648-1657. doi: 10.1016/j.neuroimage.2010.01.051

Kornell, N., Bjork, R.A. & Garcia, M.A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*, 85-97. doi: 10.1016/j.jml.2011.04.002

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 989–998. doi: 10.1037/a0015729

Kornell, N. & Son, L.K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*, 493-501. doi: 10.1080/09658210902832915

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

Kriegeskorte, N., Mur, M. & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 1-28. doi: 10.3389/neuro.06.004.2008

Kromann, C. B., Jensen, M. L. & Ringsted, C. (2009). The effects of testing on skills learning. *Medical Education*, *43*, 21–27.doi: 10.1111/j.1365-2923.2008.03245.x

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research*, *47*(1), 211–232. doi: 10.3102/00346543047002211

Larsen, D.P., Butler, A.C. & Roediger, H.L. (2009). Repeated testing improves long-termretention relative to repeated study :a randomised controlled trial. Medical Education, 43, 1174-1181. doi:10.111/j.1365-2923.2009.03518.x

Lau, E. F., Phillips, C. & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Review Neuroscience*, *9*, 920-933. doi:10.1038/nrn.2532

Lee, J.L. (2009). Reconsolidation: maintain memory relevance. *Trends in Neuroscience*, *32*, 413-420. doi: 10.1016/j.tins.2009.05.002
Leeming, F.C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology*, *29*, 210–212.
Lipko-Speed, A., Dunlosky, J. & Rawson, K.A. (2014). Does testing with feedback help grade-school children learn key concepts in science? *Journal of Applied Research in Memory and Cognition*, *3*, 171-176. doi: 10.1016/j.jarmac.2014.04.002
Linn, M.C., Gerard, L., Ryoo, K., McElhaney, K. Liu, O.L. & Rafferty, A.N. (2014). Computer-Guided Inquiry to Improve Science Learning. *Science*, *344*, 155-156. doi: 10.1126/science.1245980
Liu, X.L., Liang, P., Li, K. & Reder, L.M. (2014). Uncovering the Neural Mechanisms Underlying Learning from Tests. *PLOS ONE*, *9*, e92025 (1-7). doi: 10.1371/journal.pone.0092025
Liu, W. C., & Stengel, D. N. (2011). Improving student retention and performance in quantitative courses using clickers. *International Journal for Technology in Mathematics Education, 18*, 51-58.
Lyle, K. B. & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology*, *38*, 94–97.doi: 10.1177/0098628311401587
MacLeod, C. M., & Nelson, T. O. (1984). Response latency and response accuracy as measures of memory. *Acta Psychologica, 57*(3), 215-235. doi: 10.1016/0001-6918(84)90032-5
Mareschal, D., Butterworth, B. & Tolmie, A. (2013). Educational Neuroscience. Wiley-Blackwell (9781118725894)
Martin, E. (1968). Stimulus meaningfulness and paired associate transfer: An encoding variability hypothesis. *Psychological Review*, *75*, 421-441.
Martin, A. & Chao, L. L. (2001). Semantic memory and the brain: structure and processes. *Current Opinion in Neurobiology, 11*, 194–201. doi: 10.1016/S0959-4388(00)00196-3
Martin, S.J., Grimwood, P.D., Morris, R.G. (2000). Synaptic plasticity and memory: an evaluation of the hypothesis. *Annual Review of Neuroscience, 23*, 649–711. doi: 10.1146/annurev.neuro.23.1.649
Martyn, M. (2007). Clickers in the classroom: an active learning approach. *Educause Quarterly*, *30*, 71-74.
Matlin, M. W. (2002). Cognitive Psychology and College-Level Pedagogy: Two Siblings That Rarely Communicate. *New Directions for Teaching and Learning*, *89*, 87-103. doi: 10.1002/tl.49
May, A. (2011) Experience-dependent structural plasticity in the adult human brain. *Trends Cogn. Sci. 15*, 475–482. doi: 10.1016/j.tics.2011.08.002

Mayer, R.E. (2004). Teaching of Subject Matter. *Annual Review of Psychology, 55*, 715-744. doi: 10.1146/annurev.psych.55.082602.133124

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27,* 360–372. doi:10.1002/acp.2914

McDaniel, M. A., Wildman, K.M. & Anderson, J.L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition*, *1*, 18-26. doi: 10.1016/j.jarmac.2011.10.001

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

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representation through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 371–385. doi: 10.1037/0278-7393.11.2.371

McDaniel, M.A. & Pressley, M. (1989). Keyword and context instruction of new vocabulary meanings: Effects on text comprehension and memory. *Journal of Educational Psychology, 81*, 204–213.

McDermott, K.B., Agarwal, P. K., D'Antonio, L., Roedgier, H.L. & McDaniel, M. (2014). Both Multiple-Choice and Short-Answer Quizzes Enhance Later Exam Performance in Middle and High School Classes. *Journal of Experimental Psychology: Applied*, *20*, 3-21. doi:10-1037/xap0000004

Melton, A.W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning & Behavior*, *9*, 596-606. doi: 10.1016/S0022-5371(70)80107-4

Meyer, A.N., & Logan, J.M. (2013). Taking the testing effect beyond the college freshman: Benefits for lifelong learning. *Psychology and Aging, 28,* 142-147.

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

Mullis, I.V.S., Martin, M.O., Foy, P., Olson, J.F., Preuschoff, C., Erberber, E., et al. (2008). TIMSS 2007 International Mathematics Report. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mårtensson, J., Eriksson, J., Bodammer, N.C., Lindgren, M., Johansson, M., Nyberg, L. & Lövden, M. (2012). Growth of language related brain

areas after foreign language learning. *NeuroImage*, *63*, 240-244. doi:10.1016/j.neuroimage.2012.06.043

Nelson, S.M. Arnold, K.M., Gilmore, A.W. & McDermott, K. B. (2013). Neural Signatures of Test-Potentiated Learning in Parietal Cortex. The Journal of Neuroscience, 33, 11754-11762. doi: 10.1523/JNEUROSCI.0960-13.2013

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

Newcombe, (2002). Biology is to Medicine as Psychology is for Education. True or False? *New Directions for Teaching and Learning*, *89*, 9-18. doi: 10.1002/tl.43

Nyberg, L., Habib, R., McIntosh, A.R. & Tulving, E. (2000). Reactivation of encoding-related brain activity during memory retrieval. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 11120-11124. doi: 10.1073/pnas.97.20.11120

Nyroos, M. & Wiklund-Hörnqvist, C. (2011). The association between working memory and educational attainment as measured in different mathematical subtopics in the Swedish national assessment: primary education. *Educational Psychology*, *32*, 1-18. doi: 10.1080/01443410.2011.643578 .

OECD. (2005). *Education at a glance: OECD indicators*. Centre for Educational Research and Innovation, Paris: Organisation for Economic Co-operation and Development.

Paller, K.A. & Wagner, A.D. (2002). Observing the transformation of experience into memory, *Trends in Cognitive Sciences*, *6*, 93–102. doi: 10.1016/S1364-6613(00)01845-3

Pashler, H., Cepeda, N.J., Wixted, J.T. & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 3-8. doi: 10.1037/0278-7393.31.1.3

Pastötter, B., Weber, J., Bäuml, K-H, T. (2013). Using Testing to Improve Learning After Severe Traumatic Brain Injury, *Neuropsychology, 27,* 280-285. doi: 10.1037/a0031797

Pennebaker J.W., Gosling S.D. & Ferrell J.D. (2013). Daily Online Testing in Large Classes: Boosting College Performance while Reducing Achievement Gaps. *PLoS ONE*, *8*, e79774. doi:10.1371/journal.pone.0079774.

Poldrack, R.A. (2008). The role of fMRI in Cognitive Neuroscience: where do we stand? *Current Opinion in Neurobiology, 18*, 223-227. doi: 10.1016/j.conb.2008.07.006

Poldrack, R.A., Mumford, J.A., & Nichols, T.E. (2011). Handbook of Functional MRI Data Analysis. New York: Cambridge University Press.

Purves, D., Augustine, G.J., Fitzpatrick, D., Hall, W.C., LaMantia, A.-S., White, L.E. (2011). Synaptic Plasticity, in: Neuroscience (5th Ed). Sinauer Associates, Inc., Sunderland, Massachusetts.

Prince, S.E., Tsukiura, T. & Cabeza, R. (2007). Distinguishing the Neural Correlates of Episodic Memory Encoding and Semantic Memory. *Psychological Science*, *18*, 144-151. doi: 10.1111/j.1467-9280.2007.01864.x

Pyc, M. A. & Rawson, K. A. (2010). Why Testing Improves Memory: Mediator Effectiveness Hypothesis**.** *Science*, *15*, 335. doi:10.1126/science.1191465

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

Rawson, K.A. & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: how much is enough? *Journal of Experimental Psychology: General, 140*, 283–302. doi: 10.1016/j.learninstruc.2011.08.003

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology Applied*, *15*, 243-257. doi: 10.1037/a0016496

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

Roediger, H. L., & Karpicke, J. D. (2006a). Test-Enhanced Learning: Taking memory tests improves long-term memory. *Psychological Science, 17*, 249-255. doi: 10.1111/j.1467-9280.2006.01693.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

Roediger, H.L. & Marsh, E.J. (2005). The positive and negative consequence of multiple-choice testing. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 1155–1159. doi: 10.1037/0278-7393.31.5.1155

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 233–239. doi: 10.1037/a0017678

Rosburg, T., Johansson, M., Weigl, M. & Mecklinger, A. (2014). How does testing affect retieval related processes? An event-related potential (ERP) study on the short-term effects of repeated retrieval. *Cognitive Affective Behavioral Neuroscience*, X, XX-XX. doi: 10.3758/s13415-014-0310-y

Rossi, S., Innocenti, I., Polizotto,N.R., Feurra, M., DeCapua, A., Ulivelli, M., Bartalini, S. & Cappa, S.F. (2010). Temoral Dynamics of Memory Trace Formation in the Human Pre Frontal Cortex, *Cerebral cortex*, *21*, 368 – 373. doi: 10.1093/cercor/bhq103

Rowland, C.A. (2014, August 25). The Effect of Testing versus Restudy on Retention: A Meta-Analytic Review of the Testing Effect. *Psychological Bulletin*, doi: 10.1037/a0037559

Rowland, C. A. & DeLosh, E.L. (2014). Mnemonic benefits of retrieval practice at short retention intervals. *Memory*, Advance online publication. doi: 10.1080/09658211.2014.889710

Ryan, L., Cox, C., Hayes, S.M. & Nadel, L. (2008). Hippocampal activation during episodic and semantic memory retrieval: Comparing category production and category cued recall. *Neuropsychologia*, *46*, 2109-2121. doi: 10.1016/j.neuropsychologia.2008.02.030

Sara, S.J. (2000). Commentary – Reconsolidation: Strengthening the shaky trace through retrieval. *Nature Review Neuroscience*, *1*, 212-213. doi: 10.1038/35044575

Scimeca, J.M. & Badre, D. (2012). Striatal Contributions to Declarative Memory Retrieval. *Neuron*, *75*, 380-392. doi: 10.1016/j.neuron.2012.07.014

Schneider, W., Eschman, A. & Zuccolotto, A. (2002). E-Prime user's guide. Pittsburgh: Psychology Software Tools Inc.

Seghier, M.L. (2013). The Angular Gyrus: Multiple Functions and Multiple Subdivisions. *The Neuroscientist*, *1*, 43-61. doi: 10.1177/1073858412440596

Shapiro, A. M. & Gordon, L. T. (2013). Classroom clickers offer more than repetition: Converging evidence for the testing effect and confirmatory feedback in clicker-assisted learning. *Journal of Teaching and Learning with Technology*, *2*, 15-30.

Sigman, M., Peña, M., Goldin, A.P. & Ribeiro, S. (2014). Neuroscience and education: prime time to build the bridge. *Nature Neuroscience Review*, 17, 497- 502. doi: 10.1038/nn.3672

Skolverket (2007). *PIRLS 2006 Läsförmåga hos elever i årskurs 4: i Sverige och i världen* [PIRLS 2006 reading comprehension in pupils in grade 4: in Sweden and the world; in Swedish]. Rapport 305. Stockholm: Fritzes Kundservice.

Slamecka, N.J. & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 592-604.

Smith, M. A. & Karpicke, J. D (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory*, 22, 784-802, doi: 10.1080/09658211.2013.831454

Spaniol, J., Davidson, P., Kim, A., Han, H., Moscovitch, M., Grady, C., 2009. Event-related fMRI studies of episodic encoding and retrieval: meta-analyses using activation likelihood estimation. *Neuropsychologia*, *47*, 1765–79. doi: 10.1016/j.neuropsychologia.2009.02.028

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

Stenlund, T., Sundström, A. & Jonsson, B. (2014). Effects of repeated testing on short- and long-term memory performance across different test formats. *Educational Psychology: An International Journal of Experimental Educational Psychology, xx,* xxx-xxx, doi: 10.1080/01443410.2014.953037

Stenlund, T., Jönsson, F. & Jonsson, B. Memory and learning effects from group discussions: Influential factors and a comparison with individually practice testing. Under revision *Journal of Applied Research in Memory and Cognition*

St. Clair – Thompson, H. L. & Gathercole, S.E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *The quarterly Journal of Experimental Psychology*, *59*, 745 – 759. doi: 10.1080/17470210500162854

Squire, L.R. (2004). Minireview. Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory, 82,* 171-177. doi:10.1016/j.nlm.2004.06.005

Sumowski, J.F., Leavitt, V.M., Cohen, A., Paxton, J., Chiaravalloti, N.D., & DeLuca, J. (2013). Retrieval practice is a robust memory aid for memory-impaired patients with MS. *Multiple Sclerosis Journal, 19*, 1943–1946. doi: 10.1177/1352458513485980

Szpunar, K. K., Khan, N.Y. & Schacter, D.L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *PNAS*, *110*, 6313-6317. doi: 10.1073/pnas.1221764110

Thompson, C.P., Wenger, S.K. & Bartling, C.A. (1978). How Recall Facilitates Subsequent Recall: A Reappraisal. *Journal of Experimental Psychology*, *4*, 210-221. doi: 10.1037/0278-7393.4.3.210

Thorndike, E.L. (1906). *The principles of teaching based on psychology*. New York, NY: A.G. Seiler.

The Swedish National Agency for Education: Trends in International Mathematics and Science Study (TIMSS). (2007). Report 323. Stockholm: Skolverket.

Tulving, E. (1989). Remembering and Knowing the Past. *American Scientist, 77*, 361-367.

Tulving, E. (1972). Episodic and semantic memory. In: E. Tulving & W. Donaldson (Eds.), *Organization of memory* (1th Ed., pp.382 - 402). Academic Press. New York and London.

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.

Ullman, H., Almeida, R., & Klingberg, T. (2014). Structural maturation and brain activity predict future working memory capacity during childhood development. *The Journal of Neuroscience*, *34*, 1592-1598. doi: 10.1523/JNEUROSCI.0842-13.2014

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37,* 498–505. doi:10.3758/BF03192720

van den Broek, G.S. E., Takashima, A., Segers, E. Fernández, G. & Verhoeven, L. (2013). Neural correlates of testing effects in vocabulary learning. *Neuroimage*,*78*, 94-102. doi: 10.1016/j.neuroimage.2013.03.071

van den Broek, G. S., Segers, E., Takashima, A., & Verhoeven, L. (2013). Do testing effects change over time? Insights from immediate and delayed retrieval speed. *Memory*, *22,* 803-812. doi: 10.1080/09658211.2013.831455

Vaughn, K. E., Rawson, K.A. (2011). Diagnosing criterion-level effects on memory: What aspects of memory are enhanced by repeated retrieval? *Psychological science, 22*, 1027-1031, doi: 10.1177/0956797611417724

Vestergren, P. & Nyberg, L. (2014). Testing alters brain activity during subsequent restudy: Evidence for test-potentiated encoding. *Trends in Neuroscience and Education*, *3*, 69-80. doi: 10.1016/j.tine.2013.11.001

Vilberg, K. L. & Rugg, M. D. (2012). The Neural Correlates of Recollection: Transient Versus Sustained fMRI Effects. *The Journal of Neuroscience*, *7*, 15679-15687. doi:10.1523/JNEUROSCI.3065-12.2012

Wagner, A.D., Shannon, B.J., Kahn, I. & Buckner, R.L. (2005). Parietal lobe contributions to episodic memory retrieval. *Trends in cognitive sciences*, *9*, 445-453. doi:10.1016/j.tics.2005.07.001

Wechsler, D. (1999). *Wechsler Adult Intelligence Scale*, *WAIS-III NI* (3rd ed.). Stockholm: Pearson Assessment.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children*, *WISC-IV* (4th ed.). Stockholm: Pearson Assessment.

Wheeler, M.A., & Roediger, H.L. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, *3*, 240–245. doi: 10.1111/j.1467-9280.1992.tb00036.x

Willingham, D. T. (2009). Why don't students like school? San Francisco, CA: Jossey-Bass.

Wing, E. A., Marsh, E. J. & Cabeza, R. (2013). Neural correlates of retrieval-based memory enhancement: An fMRI study of the testing effect. *Neuropsychologia*,*12*, 2360-2370. doi: 10.1016/j.neuropsychologia.2013.04.004

Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology*, *55*, 10-16. doi: 10.1111/sjop.12093

Worsley, K.J., Andermann, M., Koulis, T., MacDonald, D & Evans, A.C. (1999). Detecting changes in non-isotropic images. *Human Brain Mapping*, *8*, 98-101. doi: 10.1002/(SICI)1097-0193(1999)8:2/3<98::AID-HBM5>3.0.CO;2-F

Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J. A., & Poldrack, R. A. (2010). Greater neural pattern similarity across repetitions is associated with better memory. *Science*, *330*, 97–101. doi: 10.1126/science.1193125

Yan, Thai & Bjork, R. A. (2014). Habits and beliefs that guide self-regulated learning: Do they vary with mindset? *Journal of Applied Research in Memory and Cognition*, *3*, 140-142, doi: 10.1016/j.jarmac.2014.04.003

Zatorre, R.J., Fields, R.D. & Johansen-Berg, H. (2012). Plasticity in gray and white: neuroimaging changes in brain structure during learning. *Nature Neuroscience*, *15*, 528-536. doi: 10.1038/nn.3045