Effects of retrieval and articulation on memory

Max Larsson Sundqvist

Academic dissertation for the Degree of Doctor of Philosophy in Psychology at Stockholm University to be publicly defended on Friday 2 June 2017 at 13.00 in David Magnussonsalen (U31), Frescati Hagväg 8.

Abstract

Many would agree that learning occurs when new information is stored in memory. Therefore, most learning efforts typically focus on encoding processes, such as additional study or other forms of repetition. However, as I will outline in this thesis, there are other means by which to improve memory, such as retrieval practice in the form of tests. Testing memory has a reinforcing effect on memory, and it improves retention more than an equal amount of repeated study – referred to as the testing effect – and it has been assumed that retrieval processes drive this effect. Recently, however, this assumption has been called into question because of findings that suggest that articulation, that is, the act of providing an explicit response on a memory test, may play a role in determining the magnitude of the testing effect. Therefore, in three studies, I have examined the effects of retrieval and articulation on later retention, in an attempt to ascertain whether the testing effect is entirely driven by retrieval, or if there are additive effects of articulation. I have also explored possible boundary conditions that may determine when, and if, the effects of retrieval and articulation become selective with respect to memory performance. In all three studies, participants studied paired associates and were tested in a cued recall paradigm after a short (~5 min) and a long (1 week) retention interval, and retrieval was either covert (i.e., responses were retrieved but not articulated) or overt (i.e., responses were retrieved and articulated).

In Study I, I demonstrated that uninstructed covert retrieval practice (by means of delayed judgments of learning) produced a testing effect (i.e., improved memory relative to a study-only condition) similar to that of explicit testing, which supports the idea that the testing effect is mainly the result of retrieval processes. In study II, I compared memory performance for covert and overt testing, and found partial support for a relative efficacy in favor of overt retrieval, compared to covert retrieval, although the effect size was small. In Study III, I further explored the distinction between different response formats (i.e., covert retrieval vs. various forms of overt testing), specifically handwriting and keyboard typing. I also examined the relative efficacy of covert versus overt retrieval as a function of list order (i.e., whether covert and overt retrieval is practiced in blocks or random order) and its manipulation within or between subjects. The results of Study III were inconclusive insofar as a relative efficacy of covert versus overt retrieval, with respect to later retention, could not be demonstrated reliably. The list order manipulations did not appear to affect covert and overt retrieval selectively. More importantly, in cases where a relative efficacy was found, the effect size was again small.

Taken together, the three studies that of thesis indicate that the benefit of testing memory appears to be almost entirely the result of retrieval processes, and that articulation alone adds very little – if anything – to the magnitude of the testing effect, at least in cued-recall paradigms. These findings are discussed in terms of their theoretical implications, as well as their importance for the development of optimal teaching and learning practices in educational settings.

Keywords: *testing effect, covert retrieval, overt retrieval, judgments of learning, delayed JOL effect, response format, retention interval.*

Stockholm 2017
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-141851

ISBN 978-91-7649-736-4
ISBN 978-91-7649-737-1

Department of Psychology

Stockholm University, 106 91 Stockholm

EFFECTS OF RETRIEVAL AND ARTICULATION ON MEMORY

Max Larsson Sundqvist

Effects of retrieval and articulation on memory

Max Larsson Sundqvist

ISBN 978-91-7649-736-4

Printed in Sweden by US AB, Stockholm 2017
Distributor: Department of Psychology, Stockholm University

An expert is a person who has made all the mistakes which can be made in a very narrow field.

–Niels Bohr

Abstract

Many would agree that learning occurs when new information is stored in memory. Therefore, most learning efforts typically focus on encoding processes, such as additional study or other forms of repetition. However, as I will outline in this thesis, there are other means by which to improve memory, such as retrieval practice in the form of tests. Testing memory has a reinforcing effect on memory, and it improves retention more than an equal amount of repeated study – referred to as the testing effect – and it has been assumed that retrieval processes drive this effect. Recently, however, this assumption has been called into question because of findings that suggest that articulation, that is, the act of providing an explicit response on a memory test, may play a role in determining the magnitude of the testing effect. Therefore, in three studies, I have examined the effects of retrieval and articulation on later retention, in an attempt to ascertain whether the testing effect is entirely driven by retrieval, or if there are additive effects of articulation. I have also explored possible boundary conditions that may determine when, and if, the effects of retrieval and articulation become selective with respect to memory performance. In all three studies, participants studied paired associates and were tested in a cued recall paradigm after a short (~5 min) and a long (1 week) retention interval, and retrieval was either covert (i.e., responses were retrieved but not articulated) or overt (i.e., responses were retrieved and articulated).

In Study I, I demonstrated that uninstructed covert retrieval practice (by means of delayed judgments of learning) produced a testing effect (i.e., improved memory relative to a study-only condition) similar to that of explicit testing, which supports the idea that the testing effect is mainly the result of retrieval processes. In study II, I compared memory performance for covert and overt testing, and found partial support for a relative efficacy in favor of overt retrieval, compared to covert retrieval, although the effect size was small. In Study III, I further explored the distinction between different response formats (i.e., covert retrieval vs. various forms of overt testing), specifically handwriting and keyboard typing. I also examined the relative efficacy of covert versus overt retrieval as a function of list order (i.e., whether covert and overt retrieval is practiced in blocks or random order) and its manipulation within or between subjects. The results of Study III were in-

conclusive insofar as a relative efficacy of covert versus overt retrieval, with respect to later retention, could not be demonstrated reliably. The list order manipulations did not appear to affect covert and overt retrieval selectively. More importantly, in cases where a relative efficacy was found, the effect size was again small.

Taken together, the three studies that of thesis indicate that the benefit of testing memory appears to be almost entirely the result of retrieval processes, and that articulation alone adds very little – if anything – to the magnitude of the testing effect, at least in cued-recall paradigms. These findings are discussed in terms of their theoretical implications, as well as their importance for the development of optimal teaching and learning practices in educational settings.

Keywords: testing effect, covert retrieval, overt retrieval, judgments of learning, delayed JOL effect, response format, retention interval

Sammanfattning på svenska

Att lära sig nya saker är till stor del beroende av vår förmåga att lagra information i minnet. Därför är det väldigt vanligt att människor, som vill lära sig (och komma ihåg) något, brukar lägga mycket tid och ansträngning på att lägga saker på minnet genom instudering och repetition, exempelvis genom att läsa kurslitteratur om och om igen inför en kommande tentamen. I den här avhandlingen diskuterar jag andra och bättre sätt att stärka minnet, exempelvis genom att öva sig på att plocka fram information ur minnet genom s.k. självtestning. Att testa minnet har en förstärkande effekt på den information som testas. Detta är ett välkänt fenomen som kallas *testeffekten*, och det som gör det intressant är att långtidsminnet förbättras mer av att försöka plocka fram information ur minnet än av en motsvarande mängd instudering av samma material eller information. Tidigare har det antagits att det är just framplockningen av information ur minnet som är den bakomliggande processen som driver testeffekten, men på senare tid har även andra möjliga processer implicerats, däribland *artikulation*. Med artikulation avses någon explicit form av uttryck för den information som testas, exempelvis genom att säga ett svar högt, eller skriva ner det med papper och penna, eller på ett tangentbord för den delen. Idén är att artikulation kan påverka hur stark testeffekten blir, på så vis att information som både plockas fram ur minnet, och därefter artikuleras på något vis, förstärks mer än information som enbart plockas fram men inte artikuleras. I tre studier har jag därför undersökt denna frågeställning genom att jämföra effekterna av framplockning med effekterna av artikulation för att undersöka vad det egentligen är som driver testeffekten. Är det enbart framplockning, eller tillför artikulation något ytterligare till effekten och dess styrka? Jag har också närmare undersökt möjliga omständigheter under vilka testeffekten påverkas i olika utsträckning av framplockning respektive artikulation.

I samtliga experiment har deltagare studerat listor innehållandes ordpar (t.ex., *piano – fingrar*) för att senare bli testade. De visas det ena ordet (*piano*) och skall försöka minnas och ange det andra ordet (*fingrar*). Deltagarnas minne testades dels efter fem minuter, och dels igen efter en vecka, och testernas *svarsformat* manipulerades så att de var antingen explicita (d.v.s., deltagarna försökte plocka fram rätt svar ur minnet och artikulerade det sedan) eller implicita (d.v.s., deltagarna fick plocka fram informationen men aldrig uttryckligen artikulera den).

Av Studie I framgick att framplockning som inte följs av artikulation ledde till en testeffekt på ett liknande sätt som i typiska testeffektsexperiment (d.v.s., där både framplockning och explicit artikulation förekommer). Med andra ord förstärktes minnet för information mer av att enbart plockas fram ur minnet (men inte artikuleras) än av att åter studera samma information under en motsvarande mängd tid.

Studie II visade att deltagarna kom ihåg fler av de ordpar som både plockats fram och artikulerats, än de ordpar som enbart plockats fram men inte artikulerats. Detta antyder att artikulation förstärker testeffekten utöver vad som redan åstadkommits av enbart framplockning, men effektstorleken var liten.

I Studie III undersöktes skillnader mellan olika former av artikulation, specifikt att skriva för hand jämfört med att skriva på ett tangentbord, med avseende på den fördel som eventuellt existerar för explicita test jämfört med implicita test. Därutöver undersökte jag också om denna fördel beror på huruvida deltagare blir testade på olika svarsformat (d.v.s., explicit eller implicit) i slumpmässig ordning eller i block av än det ena formatet, än det andra, och dessutom huruvida denna manipulation gjordes inom eller mellan individer. Resultaten från Studie III är tvetydiga såtillvida att en fördel för ett svarsformat över ett annat (d.v.s., att explicita test förbättrar minnet mer än implicita test) inte kunde påvisas på ett tillförlitligt sätt. Att testa de olika svarsformaten i block eller slumpmässig ordning verkar inte ha haft någon särskild betydelse för hur de förbättrar minnet, och heller inte huruvida denna manipulation gjordes inom eller mellan individer. I de fall där en skillnad fanns mellan de olika svarsformaten var effektstorleken återigen liten.

Sammantaget visar avhandlingens tre studier att den förstärkning i långtidsminnet, som kommer av att testa minnet under inlärning, drivs nästan uteslutande av framplockningsprocessen i sig, och att artikulation verkar tillföra väldigt litet till styrkan på testeffekten. Dessa resultat diskuteras i termer av deras teoretiska implikationer, men de är också särskilt intressanta med avseende på utformandet av instuderingstekniker och utlärningsmetoder vars syfte är att maximera inlärning och hågkomst. Därför har resultaten också praktisk betydelse för alla de studenter som önskar göra det mesta av den som ägnas åt studier, och budskapet är enkelt: Se till att testa minnet under inlärning genom att öva på att plocka fram viktig information ur minnet – då fastnar den också bättre, oavsett om den artikuleras eller inte!

Acknowledgements

First of all, I wish to thank my supervisor, Fredrik Jönsson, for having the patience, perseverance, and enthusiasm needed to see this thing through. Despite my complete inability to meet deadlines (or to plan ahead for that matter), you have managed to keep me on track for some seven years, which in itself is an awesome feat. For this, I extend my deepest gratitude and condolences. I would have never reached this point if not for your sense of urgency, even at times when I could not see why. Thank you for looking out for me, and for showing me the ropes.

I also wish to thank my co-supervisor, Timo Mäntylä, for always finding the time for last-minute (or even last-second) questions and discussions, and also for being living proof that professors are not as absent-minded and nutty as they are usually made out to be – your effortless know-how is something to aspire to. If Fredrik is the caring father-figure in our scientific family, you're undoubtedly the cool uncle.

Ivo Todorov, my sibling-at-work, thank you for being a good friend in all aspects of the word. Thanks for every day of it. 'Growing up' together at the department of psychology has been a pleasure and a privilege, however I am fairly certain we have yet to actually grow up. Let's not.

Tina Sundelin, please come back! Since you left, I have never been more productive, and the work/play asymmetry has shifted to unthinkable levels. Thank you for introducing me to parallel play.

Thank you, Sebastian Cancino, for bringing just the right amount of common sense to the lunch table discussions, or should I say, dissections. Having you as a confidant has meant a lot.

Diana Sanchez Cortes, it's probably just the oxytocin talking, but you are the sunshine of this department. Thank you for brightening the spirit of everyone around you, not least of all me.

Azade Azad, thank you for staying true to the cause. It may not be you or I who will krossa patriarkatet, but I am certain your daughters will be on the front line when it finally happens.

Anders Sand, thank you for making almost any conversation interesting by first taking a contrary position and then waiting to see what happens. Thank you also for letting me pick your brain from time to time.

Anna Blomkvist, thank you for being the most no-nonsense and yet playful person this department has ever seen, and probably ever will. You treat science as a laughing matter, and rightly so.

Henrik Nordström, thank you for tirelessly throwing the frisbee in my direction. Clearly, one of us is a dog, but let's never figure that out.

Thank you, Nichel Gonzalez, for your unwavering devotion to the art of double-entendre. Always remember, a steak pun is a rare medium well done.

To my colleagues at the department, thank you all (including, but not limited to): Andreas Jemstedt for allowing the word *zen* to assume physical form, Kristina Karlsson for caringly keeping my blood sugar levels in check, Joel Gruneau-Brulin for attachment anxiety alleviation, Veit Kubik for countless discussions and helpful suggestions, Tanaz Molapour for putting up with me and Ivo for so long, Maria Larsson for your help but more importantly your sense of humor, Johan Willander for always having a minute to spare, Torun Lindholm for your help and support, Ingrid Ekström for always bringing a smile to my face, Mats Nilsson for being the best teacher ever, Ann Fridner for your kindness and encouragement, Jesper Alvarsson for bringing me south of the border, Marie Gustafsson Sendén for sharing my passion for food and wine, Gustaf Törngren for broaching any topic and making it fascinating, Håkan Fischer for your extraordinary combination of modesty and (unfathomable) knowledge, Stephan Baraldi for encouraging me to hone my teaching skills, Marta Zakrewska for staying up so late at parties, Artin Arshamian for much-needed ideas and input, Pehr Granqvist for good banter, Annika Lantz for offering insights into the unfamiliar and intimidating domain of organizational psychology, and Jan Dalkvist for introducing me to experimental psychology in the first place. Thank you Philip Gustafsson, Hellen Vergoossen, Maja Wall, Stina Cornell Kärnekull, Elmeri Syrjänen, Linda Rämö, Monika Karlsson, Magdalena Skarp, Wenche Gros, Bo Heffler, and Shahin Foladi, for making the department of psychology such a nice place to work.

I would also like to thank my friends and family for almost never asking me about anything that has to do with work. You guys make every aspect of life so enjoyable and meaningful. Thank you, Pontus, Nikolina, and Jacob, for blurring the line between friendship and kinship.

Åke and Anna-Lotta, thank you for your unconditional love and support. You still ask me what it is I do, exactly, and I love you both for it. Besides, I often wonder the same thing myself!

Isabella, every day with you is the best. I adore you.

List of Studies

This doctoral thesis is based on the following studies:

I. Larsson Sundqvist, M., Todorov, I., Kubik, V., & Jönsson, F. U. (2012). Study for now, but judge for later: Delayed judgments of learning promote long-term retention. *Scandinavian Journal of Psychology*, 53(6), 450–454. http://doi.org/10.1111/j.1467-9450.2012.00968.x

II. Jönsson, F. U., Kubik, V., Sundqvist, M. L., Todorov, I., & Jonsson, B. (2014). How crucial is the response format for the testing effect?. *Psychological Research*, *78*(5), 623–633. http://doi.org/10.1007/s00426-013-0522-8

III. Larsson Sundqvist, M., Mäntylä, T., & Jönsson, F. U. (2017). Testing the testing effect: the relative efficacy of covert versus overt retrieval. (Submitted to journal.)

Contents

Abbreviations

ESP	Encoding-specificity principle
JOL	Judgment of learning
LOP	Levels of processing
LTM	Long-term memory
RI	Retention interval
SFP	Self-fulfilling prophecy
STM	Short-term memory
TAP	Transfer-appropriate processing

Introduction

During the course of our lives, we learn a great deal. Some things we learn because they are essential to our survival, others we learn because we are curious and desire knowledge. Some things we learn because it is expected of us to do so, and still others we learn whether we like it or not. No matter the reason for learning, all learned skills, abilities and knowledge can reasonably be thought of as information stored in memory. This also means that everything we know is contained in our memory. So, when thinking back on our childhood, providing the correct answer to a question, or remembering how to ride a bike, we access and act upon that stored information by retrieving it from memory. Naturally, we can be more or less successful in our attempts to access information in memory, which is why the concept of memory is also very much associated with the notion of testing it.

Memory testing occurs practically everywhere and all the time, because it is an essential part of modern educational systems. Standardized tests are by far the most common way of assessing educational performance, and typically, such tests require students to memorize relevant information in order to get the questions right. As will be discussed in this thesis, the act of testing memory is not only a matter of coming up with the right answers to questions (i.e., retrieving and articulating the correct information), it also has the power to strengthen memory for that information which is tested, often referred to as the *testing effect* (e.g., Roediger & Karpicke, 2006a). Most theoretical accounts explain the testing effect as the result of retrieval processes, but as testing typically involves both retrieval and articulation of a response, it is necessary to disentangle their individual contributions to the testing effect. In other words, it is important to know whether the testing effect, as reported in many previous studies, is the result of retrieval processes alone, or if the act of articulation also serves to enhance memory to some extent. Articulation, here, refers to the act of providing an explicit response (such as saying it out loud or writing it down) in reference to a task that requires retrieval of information, such as a cued recall test.

Moreover, it is possible that different ways of testing memory (e.g., instructed vs. non-instructed tests that dictate whether retrieval will be incidental or intentional) can enhance memory to different extents, meaning that some forms of testing may be superior to others. Specifically, this thesis aims to

establish what drives the testing effect by comparing effects of retrieval and articulation on later memory performance. Different forms of retrieval and articulation will also be examined in order to assess their relative efficacy.

Knowledge about the testing effect, and what drives it, is important because it has the potential to vastly improve the effectiveness of efforts to both learn and teach certain types of information, meaning that there are considerable implications for the development of optimal learning and teaching practices. For example, whereas tests (e.g., written exams) in many cases are a preferred means of assessing students' performance, the use of tests as a memory enhancer (and thus, a learning aid) still remains largely overlooked in most curricula.

Memory and learning, a brief overview

As this thesis is mainly concerned with memorial effects of retrieval and articulation, and the processes they entail, the general concept of human memory, in terms of theories and models, will not be exhaustively reviewed here. In broad terms, there are two major approaches to human memory; one that regards it as a number of processes that occur either independently or in interaction with each other (e.g., Foster & Jelic, 1999), and the other describing memory as a system of interrelated components localized within the brain (e.g., Surprenant & Neath, 2009). This thesis makes no attempt to compare, unify or contrast the two, because it relates primarily to the processing approach, and much less so to the systems approach. In fact, only a few memory systems are referred to in this thesis, namely primary memory, which is working memory or short-term memory (STM; e.g., Baddeley & Hitch, 1974; Baddeley, 2000; see Cowan, 2008 for a distinction between the two), and secondary memory, which is long-term memory (LTM; e.g., Schacter, Wagner & Buckner, 2000). However, for the purpose of clarity, it should be mentioned that LTM is not a single memory system, but is instead composed of declarative and non-declarative forms of memory, where declarative memory contains episodic and semantic information that can be explicitly expressed, such as information about events or places or facts, and non-declarative memory stores perceptual and procedural representations that cannot be explicitly expressed, such as skills and habits, priming, and perceptual memory (Squire, 2004). For example, you may know how to ride a bike, but that knowledge is not stored in memory as a verbal instruction on how to ride a bike, but instead as a learned set of skills that can only be performed rather than described with words.

Among these memory systems, the most relevant to this thesis are the declarative forms of memory, namely episodic and semantic memory. These

forms of memory are both implicated in the tasks that are typically involved in the context of education and learning. Episodic memory contains composite information of personal experiences and information about events in time and space (i.e., when and where something happened), whereas semantic memory stores non-contextual information such as facts (e.g., the name of a street) and concepts (e.g., the meaning of words; Squire, 2004). In the studies presented in this thesis, participants have learned paired associates (i.e., word pairs of varying relatedness), and this learning draws upon both semantic and episodic memory functions. Specifically, the degree of relatedness between paired associates determine the inherent difficulty in learning a particular word pair, such that lower relatedness (i.e., weak associates) makes the word pair more difficult to learn. As evidenced by Elwood (1997), word pairs that are difficult to learn draw more upon episodic than semantic memory functions. For this reason, when using the term *memory performance* in this thesis, I am referring primarily to episodic retention and to some extent semantic retention.

Recently, cognitive psychology has seen a shift towards a more embodied view of human memory and cognition (see Wilson, 2002), which may call into question some of the models and theories that are generally agreed upon at present time. I will return to some of these theories later in this thesis, and discuss their implications for the testing effect and the processes that appear to drive it. Rather than delineating the current state of affairs in the field of cognitive psychology, this thesis will mainly present theories and findings that pertain directly to memory effects that usually occur in a typical learning environment, that is, an environment in which information needs to be encoded, stored, and accessed at a later point in time. For instance, a student who prepares for a final exam is likely to go through all of these processes in order to perform well. The information, or knowledge, that is needed for a passing grade will first need to be acquired, supposedly through attending classes and lectures, taking notes, and reading course literature. It must then be stored or retained, possibly by rehearsing and re-reading information, along with other attempts at memorizing it. Finally, the information must be successfully retrieved from its storage in memory, and transferred onto paper in the form of answers to the questions in the exam. The actions and steps described above can broadly be divided into three processes, namely *encoding*, *retention*, and *retrieval*, which are described below. This division is also useful in terms of the scope of this thesis: it is important to know whether the processes that underlie the testing effect are mainly associated with encoding, retention, or retrieval, and in the case of a combination of these three, an understanding of how the testing effect is affected individually by each of them.

Encoding

Encoding refers to the process of converting information from sensory inputs, such as vision and hearing, or from different forms of cognitive processing, to units or constructs that may then be stored in memory. When we learn new information, such as a fact about something, our working memory is occupied by that information, and a mental construct that represents the content of that information is created which can then be stored more permanently in long-term memory. From point of view of this thesis, the effects of articulation should be regarded as encoding effects (or possibly re-encoding effects). For instance, rehearsing a text by reading it out loud is a form of articulation that becomes part of the encoding process, and other encoding processes also contain an articulatory component, such as writing by hand, or performing an action (e.g., Cohen, 1981). In all three studies of this thesis, I used cued recall tests with different response formats, meaning that whereas the retrieval processes were highly similar, the articulatory processes were different. In Study I, for example, retrieved information was never articulated at all, whereas in Study II, a condition with no articulation was directly compared to a condition in which information was explicitly articulated after retrieval. Moreover, Study III featured two kinds of explicit (i.e., overt) articulation of retrieved information, namely writing by hand and typing on a computer keyboard.

There are a number of memorial effects that can be observed during (and because of) encoding, such as the *levels-of-processing effect* (LOP; Craik & Lockhart, 1972; Craik, 2002), which states that the level or depth at which information is processed determines how strongly that information will be encoded into memory. Shallow processing, which is based on visual or auditory attributes of the information, leads to weaker memory traces, whereas deeper processing, which entails semantic processing, produces memory traces that are more resilient to decay.

Another memorial encoding effect is the *generation effect* (e.g., Jacoby, 1978; Slamecka & Graf, 1978), which states that information that is somehow generated during encoding is better remembered than information that is only read. For instance, participants who read words and generate synonyms to them will remember those words better than participants who only read the words. This can be explained in part by the LOP account, but the generation effect emerges also for shallow forms of processing, meaning that the process of generation itself has a reinforcing effect on memory.

Closely related to the generation effect is the *production effect* (e.g., MacLeod et al., 2010; Ozubko & MacLeod, 2010), whereby articulation during encoding, for example by saying a word out loud, can enhance memory

compared to reading it silently. While these two effects may seem practically identical, they do differ in the sense that generation is a conscious, effortful processing of information, whereas production often is not. That is, reading a word out loud does not, cognitively speaking, entail more processing of information than does reading it silently. And yet, both have the potential to enhance or improve memory. I will return to the implications of this observation later on in this thesis.

Retention

Retention is the preservation and storage of existing information in memory. As most of us can attest, memory retention is far from flawless, and typically decays or decreases with time. Research on retention and forgetting dates as far back as the late 19th century, when Herman Ebbinghaus (1885) showed that memory decay follows a curve that is steep at first and then gradually levels out until memory declines no more after enough time has passed. Whatever is remembered, after all that time, is likely to not be forgotten at all. In this thesis, retention is of crucial importance, as some of the effects observed and investigated here do not emerge until after some time has passed. This period of time is often referred to as the retention interval (RI), that is, the amount of time during which certain information has been retained prior to retrieval. Forgetting, then, is simply a measure of how much information has *not* been retained during that time. By measuring memory performance at different retention intervals, such as five minutes or a week in the studies that comprise this thesis, one can observe how much information has been retained in memory during that time, and whether different encoding or retrieval conditions are differentially affected at these retention intervals.

Retrieval

Retrieving information from memory means bringing it to mind from LTM and into primary memory, and becoming consciously aware of that information (rather than having it unconsciously stored in memory), and possibly processing it further. According to Tulving and Thomson's (1973) *encoding-specificity principle* (ESP), memory retrieval is dependent on cues and their connection to the target information, such that the retrieval cue is stored alongside the information itself during encoding. The level of congruency or overlap between the cue and the target information is what determines the probability of successful retrieval, meaning that memory is (in part) context-dependent. Another consequence of this congruence is that it becomes rather difficult to fully dissociate or disentangle the processes that underlie encoding and retrieval, respectively, because one has the potential to affect the other and vice versa. Put differently, if the act of retrieval also entails (new)

encoding processes (e.g., by means of how it was articulated) it seems very difficult to ascertain what the effects are of retrieval only, and indeed, this is also reflected in Bjork's (1975) ideas on retrieval as a "memory modifier".

A similar congruency between encoding and retrieval is also found in the *transfer-appropriate processing hypothesis* (TAP; e.g., Bransford, Franks, Morris & Stein, 1979), whereby successful retrieval is dependent on the overlap (or transfer between) the processes engaged during encoding and the processes engaged during retrieval. Indeed, Godden and Baddeley (1975) found that divers, who studied word lists while underwater or on dry land, and later recalled words freely in either environment, performed best when the learning and testing environments were the same, again indicating a context-dependent aspect of successful retrieval. It should be noted, however, that effects of ESP and TAP are typically small (see Smith & Vela, 2001; for a meta-analysis).

Retrieval underlies what we colloquially refer to as *recalling (*or *recollecting)* or *recognizing* something or someone. For instance, knowing for certain that a stimulus (e.g., a word) has been presented before entails the process of matching existing information with information retrieved from memory (e.g., Mandler, 1980), and thus, we recognize the stimulus. It should be noted that recognition is not synonymous to recollection, because recognition is typically determined by the level of perceived familiarity of a stimulus, rather than the successful retrieval of relevant information pertaining to that stimulus. Recollection, on the other hand, cannot take place without successful retrieval of the sought-after information (see Yonelinas, 2002; for a discussion on the distinction between recollection and familiarity). If you were asked to name the capital of Belarus, you could have a strong sense of familiarity to the correct answer (Minsk), because you know you have heard it before, but still not be able to retrieve it from memory and provide a correct answer. Thus, we see that while the perceived familiarity can be high, that does not guarantee that retrieval will be successful. In all three studies of this thesis, memory performance has been measured by cued recall tests in which a cue word is used to retrieve a target word, meaning that memory performance reflects retrieval success rather than perceived familiarity. Moreover, as this thesis is particularly concerned with effects that occur during (and because of) retrieval, they will be outlined in greater detail in the next section.

Effects of retrieval on memory

According to Bjork (1975), retrieval is not merely a process by which information is gathered from memory storage to then be used in some way, and

later returned to storage. Instead, the act of retrieval also modifies the state of that information in memory, and specifically, the degree of modification corresponds to the level or depth of processing involved at retrieval. In other words, if some information is either difficult to produce (see Bjork, 1994) or retrieve from memory, or requires high degrees of cognitive elaboration (cf. the LOP effect; Craik & Lockhart, 1972), it will also be more robustly re-encoded into memory and subsequently form a stronger memory trace after it has been (successfully) retrieved. However, desirable difficulties and the degree of elaborative encoding at retrieval cannot fully account for the memory enhancement that occurs after successful retrieval. For instance, Karpicke and Smith (2012) found that repeated retrieval practice led to better retention of word pairs than an equal amount of elaborative encoding, meaning that retrieval itself has some impact on memory. This retrieval-induced modification of memory is found in several well-known memory phenomena, although in different ways and in different fields of research, which are described below.

Findings from the testing effect literature

The testing effect (see Roediger & Karpicke, 2006a; for a review; Adesope, Trevisan and Sundararajan, 2017; for a meta-analysis) is a robust memory phenomenon that appears when memory is tested and participants retrieve and overtly articulate information that is needed to perform the test. More specifically, testing memory appears to be more beneficial to later retention than an equal amount of restudy (e.g., Carpenter & DeLosh, 2006; Carrier & Pashler, 1992; Karpicke & Roediger, 2007; Karpicke & Roediger, 2008; McDaniel, Anderson, Derbish & Morrisette, 2007; Nungester & Duchastel, 1982; Roediger & Karpicke, 2006b; Wheeler, Ewers & Buonanno, 2003; Toppino & Cohen, 2009; see Figure 1). Although the testing effect has mostly been studied within the context of verbal learning, where the material to be learned is usually word lists containing word pairs (i.e., paired associates; e.g., Karpicke & Roediger, 2007; Pashler, Cepeda, Wixted, & Rohrer, 2005), it has also been demonstrated for a number of other learning materials, including longer text materials (e.g., Kang, McDermott & Roediger, 2007), pictures (Wheeler & Roediger, 1992) and even visuospatial information in maps (Carpenter & Pashler, 2007). Moreover, the testing effect also appears for many testing formats, such as free recall (e.g., Zaromb & Roediger, 2010), cued recall (e.g., Carrier & Pashler, 1992), and also short-answer and multiple-choice tests (e.g., Kang, McDermott & Roediger, 2007; McDaniel, Roediger & McDermott, 2007).

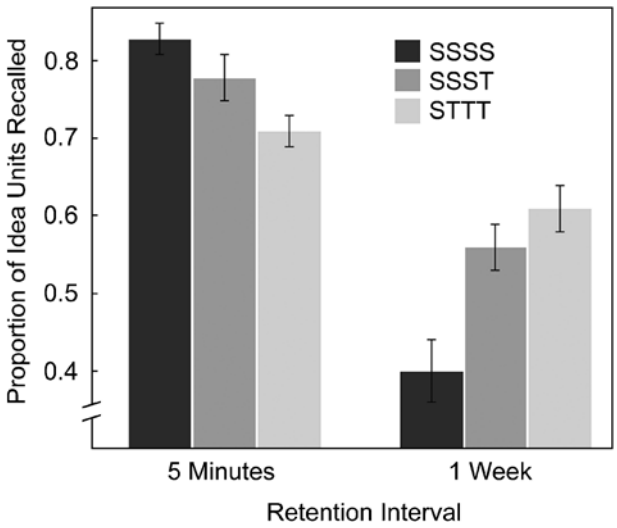


Figure 1. *Mean proportion of idea units recalled on a final test 5 minutes or 1 week after learning as a function of learning condition. The shorthand condition labels indicate the order of study (S) and test (T) periods. Error bars represent standard errors of the means. From ©Henry Roediger and Jeffrey Karpicke (2006b).*

The testing effect is particularly surprising as it appears even when no corrective feedback is provided during testing (see Kornell, Bjork & Garcia, 2011). For instance, in cued recall testing there is typically a cue word and a target word, and the task is to memorize the association between the two. During testing, only the cue word is shown and this triggers a retrieval attempt for the target word. If retrieval is successful, the target word is expressed, and the memory for that cue-target association is reinforced via the testing effect. If retrieval is unsuccessful, no answer can be given and it also follows that there can be no memory reinforcement by means of a testing effect. Despite this limitation, testing is still more beneficial to memory than a comparable amount of additional study, which provides access to the entire material, rather than only what was successfully retrieved during testing. Corrective feedback (i.e., seeing the correct answer after submitting a response during testing) boosts the testing effect for two reasons: i) it produces testing effects also for items that were not successfully retrieved, and ii) it reduces the risk of reinforcing false information in memory by means of a testing effect (see Agarwal, Karpicke, Kang, Roediger & McDermott, 2008; for a discussion on this).

Why, then, is testing so beneficial for later memory performance, and even without feedback? As evidenced by Roediger and Karpicke (2006a) in their review of the testing effect, there is no single answer that will definitively put this question to rest, but I have attempted to summarize here the different theoretical accounts that have been put forth to date.

In line with Dempster's (1996) *retrieval hypothesis,* there are a number of ideas (many of which have already been mentioned briefly) that may account for the testing effect, and as the name suggests, they all share the assumption that the memorial benefit stems from the act of retrieval. This could occur by means of a generation effect (e.g., Jacoby, 1978) or an increased level of elaboration of the memory trace (cf., Craik & Tulving, 1975), as both entail more effortful retrieval. It could also be that retrieval produces an increase in the number of retrieval routes that lead to the relevant information (e.g., Bjork, 1975; McDaniel & Masson, 1985), comparably to the LOP effect. As demonstrated by Gardiner, Passmore, Herriot and Klee (1973), greater retrieval effort (as measured by the time it took participants to retrieve information) appears to be associated with better recall performance on a later test. In sum, these theories predict that the magnitude of the testing effect should increase with the difficulty, effort and elaboration associated with the retrieval attempt, and there are a number of findings that support this notion (e.g., Carpenter & DeLosh, 2006; Glover 1989; Kornell, Bjork & Garcia, 2011; Toppino & Cohen, 2009; Wheeler, Ewers & Buonanno, 2003).

On a general note, the presence of a testing effect is not always established in the same fashion everywhere. For instance, some would argue that a testing effect is demonstrated by a response format × retention interval (RI) interaction (cf., Kornell, Bjork & Garcia, 2011; Rowland, 2014), specifically because it dissociates the effects of studying and testing (Toppino & Cohen, 2009), whereas others would say that a testing effect manifests as the superiority of a testing condition relative to a study-only condition, more or less regardless of the RI (i.e., a main effect of response format), meaning that tests do not necessarily reduce the rate of forgetting but rather should improve memory altogether (cf., Slamecka & Katsaiti, 1988). As argued by Wheeler, Ewers and Buonanno (2003), studying may reinforce memory representation of an item, but testing instead enhances the retrieval process itself, meaning that the effects of studying and testing may not become dissociated until some time has passed. Indeed, in a meta-analysis of 272 independent effects from 118 separate experiments on the testing effect, Adesope, Trevisan and Sundararajan (2017) found that the effect size of the testing effect was larger when the RI was on the order of days, than when it was less than 24 hours. Therefore, in this thesis I define the testing effect practically as a response format × RI interaction.

From point of view of the TAP hypothesis, the testing effect could be explained as the result of a larger congruency between the conditions during learning and the conditions at final testing. In other words, the testing effect occurs because the processes required to perform well during retrieval practice are the same processes that are required to perform well at the final test, whereas the processes involved in restudying are not the same, or at least not as similar. Thus, the magnitude of the testing effect becomes a function of this similarity – the more similar the conditions, the larger the testing effect. However, the support for this notion is not unequivocal, and despite studies that do confirm a TAP effect (e.g., Nungester & Duchastel, 1982; McDaniel & Fisher, 1991), there are some that find precisely the opposite (see Kang et al., 2007).

Nonetheless, in a meta-analytic review by Rowland (2014), the initial–final test match did not reliably increase the magnitude of the testing effect, meaning that the TAP theory cannot account for its occurrence. The retrieval (effort) theories, however, gathered support in the sense that recall tests (which entail more effortful and elaborative retrieval) yielded larger testing effects than recognition tests. The findings above explain very well why testing should reinforce memory and improve later retention, and virtually all accounts converge on the idea that retrieval produces the memorial benefit. However, this notion can be found also outside of the testing effect literature, because retrieval processes also underlie other memorial effects that do not pertain to testing. As I will try to illustrate in the next section, retrieval is often uninstructed and incidental, and an indirect consequence of metacognitive judgments.

Findings from research on metacognition and metamemory

Metacognition is the focus of Study I of this thesis, and it refers to higher-order thinking and appraisal processes that serve to either monitor or regulate ongoing cognitive processes (e.g., Schraw, 1998). Etymologically speaking, it translates as “knowledge about knowledge” or “thinking about thinking”, meaning that it involves not our thoughts per se, but rather our awareness and knowledge about those thoughts. Similarly, metamemory refers to the (metacognitive) monitoring of memory processes such as learning.

While there are many different ways in which metacognitive judgments can occur and influence our decisions, the first study in this thesis is particularly concerned with so-called *judgments of learning* (JOL; e.g., Leonesio & Nelson, 1990; Nelson & Narens, 1994) and the *delayed JOL effect* (Dunlosky & Nelson, 1992; Jang & Nelson, 2005; Nelson, Narens & Dunlosky, 2004), and this is because the delayed JOL effect has a close relationship to the testing effect at the level of memory processes. Although this relationship

will be more thoroughly explained later on, it should be noted that they both entail the attempted retrieval of some relevant information. Granted, what happens with that information after it has been retrieved may differ for many reasons, but they both entail retrieval-induced memorial effects.

Judgments of learning are metacognitive assessments of the degree to which something has been learned. If I return to the student preparing for a final exam, I could argue that his or her decision to either keep studying or instead go to bed is likely influenced by a JOL, that is, an assessment of the degree to which something has been learned (and whether or not that degree of learning is sufficient). This way, a JOL is basically equivalent to asking the question "Do I know this?", or in more prospective terms, "Will I remember this on the day of the exam?". Obviously, these judgments are not perfect, meaning that JOLs can be more or less accurate. Accuracy is typically evaluated by calculating an item-by-item Goodman-Kruskal gamma correlation between JOLs and final recall performance (see Nelson, 1984). Given the often-prospective nature of JOLs, a higher accuracy is desirable because it indicates that the student will likely remember what he or she also expected to remember, giving the JOL a predictive value towards future memory performance. Conversely, low JOL accuracy reflects low predictive value, meaning that the student remembers more or less than expected. As metacognitive assessments serve to not only monitor but also regulate cognitive processes – in this case, the learning efforts – high JOL accuracy is assumed to be beneficial for optimal learning behavior because the learning decisions are based on assessments with high predictive value towards future memory performance.

So, if high JOL accuracy is desirable because of its predictive value, what exactly makes these judgments more accurate and predictive? The single-most important factor to increased JOL accuracy is temporal delay, and that is why the *delayed* JOL effect is called exactly that. In terms of predictive value, delayed JOLs are considerably more accurate than immediate JOLs, that is, JOLs that are made after some time has passed instead of during or immediately after learning. The reason for this is still a matter of debate (see Rhodes & Tauber, 2011, for a review), but virtually all accounts share one fundamental assumption: A delay (between learning and JOL) eliminates the possibility that the JOL becomes primarily based on information available in STM (see Nelson & Dunlosky, 1991).

Thus, the accuracy or predictive value of an immediate JOL is low, and for two reasons, i) it does not take into account the fact that working memory stores are impermanent and decay rapidly, and ii) even if the relevant information was indeed encoded into LTM, the detrimental effects of time (i.e., forgetting) have likely been disregarded when making an immediate JOL. A

delayed JOL, on the other hand, takes place when the information is no longer available in working memory, and therefore, is reliant on successful retrieval from LTM to begin with. Moreover, as this retrieval attempt is made at a delay, forgetting may have already taken place, which makes it more likely that forgetting effects are factored into the JOL. Therefore, delayed JOLs are based on information that, if successfully retrieved, will probably also be successfully retrieved again at a later point in time. If the information cannot be retreived, this too serves as a diagnostic purpose in the sense that it is very unlikely that it would somehow reappear into memory at a later time, thus making the JOL all the more accurate.

Therefore, delayed JOLs are based on information that, if successfully retrieved, will probably also be successfully retrieved again at a later point in time. If the information cannot be retrieved, this too serves as a diagnostic purpose in the sense that it is very unlikely that it would somehow reappear into memory at a later time, thus making the JOL all the more accurate. Indeed, Dunlosky and Nelson (1992) investigated the accuracy for different forms of JOLs and found that delayed JOLs were more accurate than immediate JOLs (see also Rhodes & Tauber, 2011, for a review), but only for JOLs that were based on a cue rather than the entire information that was to be learned. So, if the information were a word pair (e.g., *elephant – memory)*, the JOL would then be based on only the cue word (i.e., *elephant*), thus prompting an attempted retrieval for the target word (i.e., *memory*).

Returning to Bjork's (1975) idea of retrieval as a modifier of information stored in memory, another explanation for the delayed JOL effect is found in the *self-fulfilling prophecy* theory (SFP; Spellman and Bjork, 1992). The basic premise of this theory is that a delayed JOL is essentially a diagnostic test of LTM, which involves an attempted retrieval of the sought-after information, and if that retrieval attempt is successful, I should expect to observe the same memorial effects as in other contexts where retrieval occurs, such as memory testing. For this reason, this retrieval-based memory enhancement (i.e., the testing effect) also partially explains the correlation between delayed JOLs and later recall performance; information that cannot be retrieved is not reinforced via the testing effect and is given a low JOL, whereas information that is successfully retrieved is reinforced via the testing effect and also given a high JOL. Put simply, the delayed JOL effect is partly the result of a testing effect inherent to the nature of delayed JOLs and the retrieval attempts they elicit.

Study I of this thesis will investigate whether repeated, delayed JOLs do indeed improve later retention relative to an equal amount of additional study, in much the same way that repeated testing does, or put differently, that retrieving information covertly (by means of delayed JOLs) does indeed

produce a testing effect, as posited by the SFP theory. If so, it would expand the body of research that reports testing effects in covert retrieval settings (e.g., Carpenter, Pashler & Vul; 2006; Carpenter & Pashler, 2007; Kang, 2010) to include not only instructed covert retrieval but also covert retrieval, which is unintentional or incidental because of how delayed JOLs are performed.

The distinction between instructed and incidental retrieval is important for our understanding of the extent to which delayed JOLs should be expected to produce a testing effect. For instance, Jönsson, Hedner and Olsson (2012) compared memory performance for items that were learned through study followed by either instructed self-testing or delayed JOLs, and found that both testing and making JOLs produced reliable testing effects, and more importantly, that they did not differ in magnitude.

Using a similar design, Tauber, Dunlosky and Rawson (2015) compared explicit tests with delayed JOLs, and found that whereas retrieval practice (by means of cued recall tests) boosted memory performance on a final test, delayed JOLs did not. The authors explain this difference in terms of retrieval effort, meaning that retrieval practice causes more effortful retrieval attempts than do delayed JOLs, and that JOLs are based primarily on cue familiarity and not whether the retrieval attempts are successful. The difference in retrieval effort is reflected very well in the response latencies for the two conditions, with lower latencies (i.e., less time spent retrieving information) for the delayed JOL conditions compared to the test conditions. An interesting observation, here, is the fact that exposure times for the study and testing conditions were equated in the study by Jönsson, Hedner and Olsson. (2012), whereas the Tauber et al. (2015) study did not place any time restraints on the testing conditions. Given the disparate results of these studies, perhaps the difference can be explained in terms of more effortful retrieval for explicit testing (compared to JOLs) but only when allowed for, in terms of time.

Another key distinction between the previous studies discussed here is whether the act of making a JOL is performed once or repeatedly. In many delayed JOL studies, JOLs are only made once for each item, whereas testing effect experiments typically feature repeated tests for each item. Repeated testing sessions effectively act as a form of feedback for participants in the sense that they may recall mistakes made during one testing session and use that information to guide their learning efforts in subsequent study and testing of the learning material (see Kang, McDermott & Roediger, 2007). Whereas Tauber, Dunlosky and Rawson (2015) used a single delayed JOL trial, Jönsson, Hedner and Olsson (2012) used three, which might explain why only the latter found a testing effect as the result of delayed JOLs.

Effects of articulation on memory

Now that we have a basic understanding of how memory is affected by retrieval alone, let us also consider the effects of articulation, that is, the act of overtly expressing or verbalizing information that has been retrieved from memory. Although the testing effect seems to be mainly the result of processes involved at retrieval, there is also evidence that explicit articulation of information serves to enhance its retention. For instance, Gardiner, Passmore, Herriot and Klee (1977) investigated memory for words as a function of either writing them down or saying them out loud (or both) during learning, and found that while there were no differences between the groups that had either spoken or written down the words, the group that had engaged in both types of articulation were superior in terms of word recognition. At first glance, this finding seems intuitive and could for example be explained by Dempster's (1996) *amount-of-processing* account of the testing effect. However, since the amount of exposure to the material was equal across all groups, this explanation is not satisfactory (see also Roediger & Karpicke, 2006a). While the findings of Gardiner et al. (1977) do not suggest that different forms of articulation affect memory differentially, they clearly demonstrate that articulation may affect later memory performance. Gardiner et al. (1977) argued that different forms of articulation would create qualitatively different memory traces in terms of their visual, auditory, and kinesthetic attributes. In other words, saying a word out loud will produce different auditory aspects of the memory trace for that word, compared to having written it down on paper.

Moreover, it could be argued that articulatory processes may be present even when no explicit articulation takes place. For instance, the phonological loop in Baddeley's (2000; 2001) working memory model includes a process of subvocal articulation, meaning that articulation (in this case, speech) is foregone by subvocal rehearsal of the information that is about to be articulated. From this point of view, even *thinking* about the answer to a question includes an articulatory component, which makes retrieval and articulation effectively indistinguishable from one another. Therefore, in this thesis, I will only discuss effects of overt, explicit articulation, and not effects of subvocal articulation.

The production effect and the generation effect

A similar line of evidence for articulatory effects on memory comes from the production effect, which has already been described briefly. When explicitly articulating information, such as a word, a verbal cue is likely created in a way that would not happen for information that was never articulated, and this cue facilitates retrieval of that information in the future (see MacLeod et

al., 2010). Typically, the production effect is observed during encoding rather than retrieval, but as the basic premise of the testing effect is that there is a relationship between retrieval and encoding (in the sense that retrieval efforts may alter the memory state of some information and thus re-encode it differently back into LTM), it may have relevance also for our understanding of the testing effect. For instance, if the retrieval attempts that are made when testing memoy (during learning) can be considered additional learning (i.e., encoding) opportunities, it is reasonable to suspect that the production effect may appear under testing conditions that entail retrieval and articulation (and thus, encoding) of the information that is tested. If so, I should expect to find testing effects of greater magnitude for testing that involves not only retrieval but also articulation.

The generation effect also posits better memory performance as a result of articulation, for instance by generating synonyms or elaborating on the to-be-learned material in some way (e.g., Hirshman & Bjork, 1988; see Bertsch, Pesta, Wiscott & McDaniel, 2007, for a review). However, Karpicke & Zaromb (2009) pointed out that whereas the generation effect often appears under circumstances where generation is incidental (i.e., the articulation of some information is merely a means of completing some other task, such as generating synonyms), the testing effect instead entails both retrieval and articulation in contexts where these processes are their own goals (e.g., providing the correct answer to a question). So, while highly similar, the generation effect and the testing effect differ by whether retrieval is incidental or intentional (although it should be noted that learning or encoding itself is instructed and therefore intentional), and, perhaps more interestingly, the effects of retrieval appear to be paramount to the effects of articulation in terms of memorial enhancement. As mentioned earlier, this distinction is also relevant to our understanding of how and why JOLs produce testing effects, precisely because JOLs too include an element of retrieval that is unintentional, meaning that retrieval effort is likely low compared to explicit testing. In turn, this can also explain why it seems difficult to find testing effects with JOLs instead of overt testing (e.g., Tauber, Dunlosky & Rawson, 2015; but see Jönsson, Olsson & Hedner, 2012).

Findings from research on embodied cognition, haptics, and writing

Embodied cognition refers to the notion that many cognitive processes reflect not only what happens in the brain, but also how we use our bodies to physically interact with the world around us (e.g., Wilson, 2002). For example, the *enactment effect* (e.g., Cohen, 1981; Engelkamp & Krumnacker,

1980; Zimmer & Cohen, 2001) explains why memory for action phrases (e.g., "throw the stone") is better when these actions are physically carried out, or *enacted*, compared to only reading the phrases. There are several theoretical explanations for this effect, one being the semantic link between the object (e.g., stone) and the action, which is typically a verb (e.g., throw), meaning that the enactment of the action phrase also entails additional semantic elaboration of the action phrase and its meaning. In other words, the action performed also reflects the meaning of the action phrase. If the object is instead paired with an unrelated action verb (e.g., "drink the stone"), the enactment effect disappears (e.g., Zimmer & Engelkamp, 2003).

From point of view of the testing effect, it could be argued that testing in many cases includes a motoric component of articulation, for instance by typing or writing an answer, whereas covert retrieval does not. However, as seen above, the enactment effect is contingent on the extent to which the action physically represents, or relates to, the semantic content of the action phrase. Therefore, in a setting where paired associates are tested, the act of typing or writing a word is not necessarily the same as physically articulating the meaning of that word, meaning that there should be no enactment effect for such tasks. Nonetheless, Kubik, Olofsson, Nilsson and Jönsson (2015) found a testing effect for action phrases in a cued recall paradigm, however only for items that used nouns (i.e., the objects) as cues and not verbs.

When it comes to writing things down, there is a colloquial understanding among many that it is beneficial for memory to do so (cf., Naka & Naoi, 1995). The general idea, for many, is that the act of writing something down will simply make it "stick" in a way that it otherwise would not. If so, why does articulation benefit memory, and how? Also, do different forms of articulation benefit memory differently? Mangen, Anda, Oxborough and Brønnick (2015) investigated various forms of articulation and their effects on later retention, specifically by comparing handwriting to typing on a computer keyboard as well as a tablet PC keyboard (i.e., a keyboard displayed on a touch screen), and found that free recall performance was better for words that had been written by hand than words that were typed on computer or tablet keyboards. However, other similar studies have not found such an advantage (e.g., Vaughn, Schumm, & Gordon, 1992) for handwriting versus typing.

In a review, Mangen and Velay (2010) conclude that handwriting is qualitatively different from typing in several ways. For example, the tracing of letters that occurs during handwriting is uniquely linked to each particular letter that is written, whereas pressing a keyboard button involves essentially identical movements for each button, meaning that handwriting involves a higher

level of embodied cognition than does typing. And as testing effects have been observed for materials that involve embodied cognition, I should expect that higher levels of embodied cognition would produce larger testing effects. Some support for this notion can be found in Longcamp, Boucard, Gilhodes and Velay (2006), who found that characters and letters learned through typing were less accurately recognized at a later test than handwritten information. However, as this study used recognition tests, meaning they measured familiarity rather than retention – paired with the fact that the learning material consisted of single letters and characters, and not words or more complex forms of information – it remains to be established whether the typing/handwriting distinction is relevant also in a testing effect paradigm.

Moreover, there is a higher degree of congruency between the visual and sensorimotor content of handwritten information than information that has been typed, for example on a keyboard. Again, this is because the cerebral representation of a letter or a word contains information about its visual attributes as well as the sensorimotor components that are required to physically construct that letter or word (e.g., Kato et al., 1999; Matsuo et al., 2003). When typing a word on a keyboard, a large portion of those visual attributes and sensorimotor components are lost because the act of pressing a button on a keyboard is neither unique for each individual letter (except for the placement of the letter on the keyboard layout) in terms of the movement of the hands, nor is it linked to any of the visual attributes of the letter or word itself. This can be likened to the previously discussed ideas on retention as a function of the number of available retrieval routes that lead to specific information (e.g., Bjork, 1975; McDaniel & Masson, 1985). By that logic, information that is accessible through visual *and* sensimotoric retrieval routes should have a higher probability of successful retrieval than information with fewer retrieval routes, and thus be better retained.

In addition to the content of cerebral representations of letters and words, Mangen and Velay (2010) also point to the difference in processing speed involved in typing and handwriting, with handwriting being the slower and more arduous of the two. If I combine these findings with Bjork's (1975; 1994) ideas on effortful processing and desirable difficulties, I expect that i) the act of articulation (in any form) may entail more effortful processing and thus lead to better memory performance, and ii) in testing situations, where both retrieval and articulation typically occur, I expect that their effects are additive with respect to the magnitude of the testing effect, and finally iii) handwriting should be more beneficial for later memory performance than typing because it is more embodied. This last distinction is considered particularly in the first two experiments of Study III in this thesis.

Covert versus overt retrieval – is there a relative efficacy?

At this point, I have reviewed memorial effects associated with retrieval and articulation, and it would seem the testing effect is implicated in learning contexts where both retrieval *and* articulation takes place (e.g., cued recall tests during learning), but also in contexts where *only* retrieval and no articulation takes place (e.g., delayed JOLs). A typical testing effect experiment includes overt retrieval (i.e. testing), which involves i) the attempt to retrieve some relevant information, and given that retrieval is successful; ii) the overt articulation of that information. For example, this can be done by writing it down on a paper (Carpenter & DeLosh, 2006), by typing it on a keyboard (Toppino & Cohen, 2009), or by saying it out loud (Kuo & Hirshman, 1996). In covert retrieval, however, information is retrieved from memory but never explicitly articulated. An obvious difference here is that only overt retrieval provides the possibility of determining memory accuracy (i.e., whether or not the participants' responses are correct) in a cued or free recall paradigm, which may also explain why covert retrieval has received so little attention in previous studies.

The difference between covert and overt retrieval can be easily illustrated by comparing the questions "Do you know what year Picasso was born?" and "What year was Picasso born?"; both entail attempted retrieval of relevant information, but the former is answered by a simple *yes* or *no*, whereas the latter also involves the overt articulation of the relevant information (i.e., 1881). This example also demonstrates the degree of similarity (in practical terms) between covert and overt retrieval (in the sense that someone who answers *yes* is likely very prepared to also answer *1881*), and that the task itself (i.e., the way that a test is carried out, what instructions are used, etc.) will determine whether participants engage in covert or overt retrieval.

In many learning environments, like that of a student preparing for an exam, I would argue that covert retrieval is equally commonplace as overt retrieval. For example, many use so-called flashcards as a learning tool, and besides the benefit of spaced repetition (i.e., repeated study or exposure to a learning material that is spread out over several occasions, rather than all at once; e.g., Ausubel & Youssef, 1965), flashcards also produce a testing effect because they elicit a retrieval attempt (e.g., Kornell & Son, 2009). This retrieval attempt can be either overt (e.g., by saying the answer out loud) or covert (i.e., by thinking of the answer but not articulating it), but regardless of retrieval mode, the testing effect will boost memory for that information which was successfully retrieved. This is not to say that all retrieval modes are equivalent and equal, but it demonstrates that both covert and overt retrieval strategies are both likely to be employed in typical learning environments.

The question, then, is whether covert and overt retrieval (during learning) produce testing effects of equal magnitudes. While it is known that different forms of memory testing may benefit memory to different extents (e.g., recognition tests produce a weaker testing effect than does free and cued recall tests; see Roediger et al., 2010), relatively little is known about the relative efficacy of different response formats, that is, the retrieval mode used at testing that takes place during learning.

Although there are studies that have found a testing effect for covert testing procedures for various materials (e.g., Carpenter, Pashler & Vul; 2006; Carpenter & Pashler, 2007; Kang, 2010), hardly any have directly compared testing effects produced by either covert or overt retrieval, meaning that still very little is known about their relative efficacy. From the few studies that have directly compared covert and overt retrieval in a paired-associates and cued-recall paradigm (including the second and third study of this thesis), there appears to be no clear-cut conclusion to draw. For example, Putnam and Roediger (2013) found, in only one of three experiments, an effect of response format (i.e., whether retrieval was covert or overt at testing in a learning phase) such that overt retrieval was more beneficial for later retention than covert retrieval. It should also be noted that Experiment 1 of their study did not replicate a testing effect to begin with, and likely this was because the restudy condition was confounded by JOLs that participants made after each item (i.e., the restudy condition was, in fact, also a covert testing condition). In four experiments by Smith, Roediger and Karpicke (2013), no relative efficacy was found for covert versus overt retrieval in a free recall paradigm. Given such inconclusive findings, it seems worthwhile to further test whether covert and overt retrieval produce testing effects of comparable magnitudes. This is the aim of Study II of this thesis.

Boundary conditions of the testing effect

One aspect of testing effect experiments that has, perhaps, received too little attention thus far, is the question of how (and if) methodological differences between testing effect experiments produce different memorial benefits. For instance, does it matter whether lists used in testing effect experiments are tested in blocks or in random order? Should I expect identical testing effects for within- or between-subject designs?

Rowland, Littrell-Baez, Sensenig and DeLosh (2014) examined list order effects (pure vs. mixed), and found that the testing effect was unaffected by list order. However, while list order may not affect the testing effect itself, it can have implications for the magnitudes of testing effects produced by different response formats. This is directly related to the expectations that par-

ticipants have of the tasks performed in a testing effect experiment. For instance, if all items in a list are tested overtly, participants are likely to become aware of this fact, and thus take that into account when performing the test. In effect, this could cause participants to employ different retrieval strategies, or thresholds in terms of metacognitive judgments, for different forms of testing. If, on the other hand, items are tested overtly and covertly at random, participants have no way of anticipating the response format that will be used, which effectively ensures that the retrieval processes are identical for all items until the point of articulation. And for the purposes of disentangling effects of retrieval and articulation, that is precisely what is needed. Partial support for this argument can be found in Jonker, Levene and MacLeod (2014), who found production effects (i.e., better retention for participants who read words out loud than those who read them silently), but only for random lists and not blocked lists. Granted, these findings do not pertain to effects of retrieval per se, but they demonstrate how different list orders can lead to differences in memory performance.

Another important methodological consideration is that of the manipulation of variables within or between subjects. In much the same way that a random list order ensures that one form of testing does not affect another, the within- or between-subjects manipulation of this list order may also selectively affect the way participants engage in memory testing. If participants are tested *both* randomly and in blocks of covert and overt retrieval, there is undoubtedly the possibility that one mode of testing can influence the other, for instance in such a way that participants treat the two tasks as if they were more similar than they are in fact. Huff, McNabb and Hutchison (2015) investigated free recall performance in a false memory paradigm, and found that blocked lists led to better retention than random lists, but only for within-subject design. Thus, I see evidence that within-subject manipulations are susceptible to this type of transfer effects, whereas between-subject designs are not, and this is particularly relevant to the list order manipulation discussed above.

Overall aims

Based on the foregoing review, the main objectives of the studies that comprise this thesis were to examine the following:

I – does covert retrieval produce a testing effect?

It is known from research on the delayed JOL effect that JOLs elicit a retrieval attempt such that participants covertly test memory diagnostically before providing the JOL estimate (e.g., Spellman & Bjork, 1992, Nelson & Dunlosky, 1991). Those retrieval attempts are uninstructed and spontaneous. Moreover, this effect is more pronounced when the JOL is made at a delay

after learning (cf., Rhodes & Tauber, 2011), which effectively makes the conditions (under which a delayed JOL is made) very similar to those present during typical memory testing. Thus, and according to Spellman and Bjork's (1992) self-fulfilling prophecy, the delayed JOL effect can be regarded as the result of a testing effect produced by a retrieval attempt. Therefore, the first aim of this thesis, and the focus of Study I, is to investigate whether uninstructed, covert retrieval (as elicited by delayed JOLs) do indeed produce a testing effect (i.e., benefit later retention relative to equal amounts of additional study).

II – is there a relative efficacy of overt versus covert retrieval?

Let us also ponder whether there are any effects of articulation on memory performance that add to those already produced by retrieval. As already examined, the production effect and the generation effect both posit memorial benefits as the result of articulatory processes, and although it matters whether retrieval is intentional or incidental (cf. Karpicke & Zaromb, 2010), it remains unclear whether articulation must also be intentional for such enhancing effects to occur in a similar way. Originally, the testing effect was thought of as an effect that pertains only to the act of retrieval (e.g, Bjork, 1975; 1994). However, more recent findings (e.g., Putnam & Roediger, 2013; Smith, Roediger & Karpicke, 2013) call into question whether there is also an articulatory component to this effect. Therefore, the second aim of this thesis, and the focus of Studies II and III is to ascertain whether covert and overt retrieval produce testing effects of comparable magnitudes, specifically by using designs where the *only* difference between the covert and overt retrieval conditions is the act of articulation (rather than exposure times or other procedural differences, which was further explored in Study III).

III – does the overt response format matter?

I have presented evidence from research on embodied cognition, writing, and haptics which show that, regardless of the testing effect per se, the act of articulation is in some cases an embodied form of cognition associated with qualitatively different cerebral representations of the visual and sensorimotor content of the information being articulated, depending specifically on the mode of articulation. For this reason, the third aim of this thesis, and the focus of Study III, is to further investigate the effects of different forms of articulation with respect to the testing effect. Compared to the second aim, which pertains to the relative efficacy of covert versus overt retrieval, this third aim relates directly to the relative efficacy of different modes of articulation, specifically handwriting versus typing on a computer keyboard.

IV – the testing effect magnitude as a function of experimental designs

There are a number of methodological considerations, relevant to the relative efficacy of covert and overt retrieval, which have been outlined above. For

this reason, this thesis aims to examine the role of list order, and whether it is manipulated within or between subjects, with respect to the relative efficacy of covert versus overt retrieval. Specifically, list order is manipulated because it may affect the way participants perceive and anticipate the learning task itself, and by comparing these list orders in both within- and between-subjects designs, I can ascertain whether any differences in final memory performance is due to transfer effects from one list order to another (i.e., blocked to random or vice versa).

Empirical studies

Study I

Aim

Study I examines the SFP hypothesis (Spellman & Bjork, 1992) of the delayed JOL effect, namely that delayed JOLs can be regarded as covert learning opportunities, and as such, they should produce a memorial benefit relative to a study-only condition (i.e., a testing effect). The aim was to compare the effect of making delayed JOLs on later retention with that of an equal amount of restudy, in a paired-associates paradigm.

Procedure

87 participants were divided into four groups in a 2 × 2 between-subjects design, with learning groups (study-only vs. delayed JOL) and RI (5 minutes vs. one week) as the independent variables, and final cued-recall performance as the dependent variable. The learning material consisted of 40 Swahili-Swedish word pairs adapted from Nelson and Dunlosky's (1994) Swahili-English norms. In each study session, every word pair was displayed individually, in random order, on a computer screen for 8 seconds. Each session had a 30-second distractor task between them. The study-only groups performed four consecutive study sessions, whereas the JOL groups alternated between study sessions and JOL sessions (in which only the cue word was shown, and participants were asked "How certain are you that you will remember the word in [5 minutes or 1 week]?" and responded on a percentage scale from 0 – 100) for a total of four sessions. After the initial study and JOL sessions, a final cued recall test was administered after either five minutes or one week, depending on the RI.

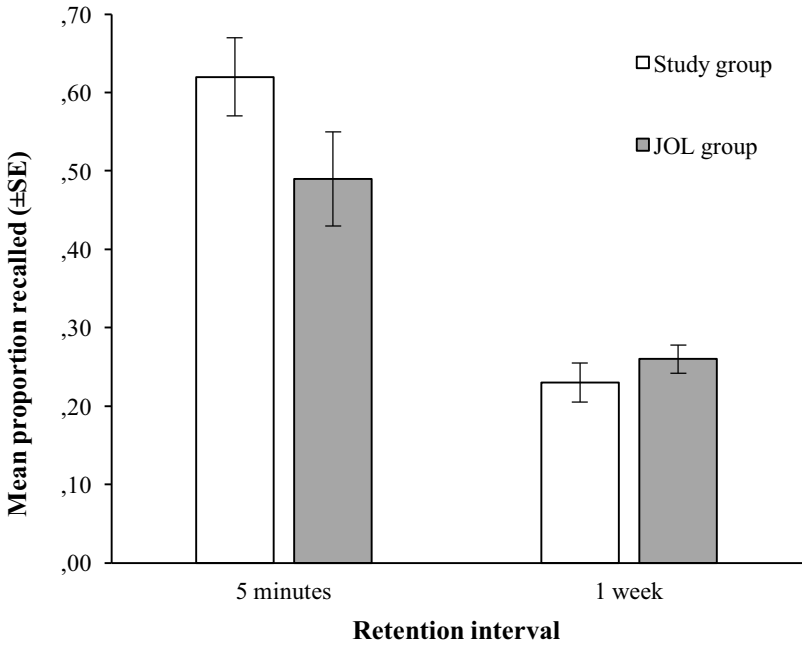


Figure 2. *Mean recall performance across learning group and retention interval. Error bars denote the standard errors of the means.*

There was a significant interaction between group and RI, $F_{1, 88} = 5.31$, $\eta_p^2 = .06$, $p = .02$. The JOL groups remembered less than the study-only groups after five minutes ($p = .045$; Cohen's $d = .62$), and slightly, but non-significantly more after the long RI ($p = .31$; $d = .31$). Although none of the JOL groups remembered significantly more than their respective study-only counterpart, the interaction clearly shows that the information was better retained over time in the JOL group, compared to the study group (see Figure 2). In terms of the difference between the two RIs for each learning condition (i.e., the amount of information that was forgotten in a week), the effect size was much larger for the study groups ($d = 2.13$), indicating a much faster rate of forgetting than in the JOL groups ($d = 1.02$). The item-by-item gamma correlations for the JOL groups were high (1st JOL: $M = .79$, $SD = .23$; 2nd JOL: $M = .90$, $SD = .10$), meaning that the participants made highly predictive metacognitive assessments.

Conclusion

Study I replicated the learning condition × RI interaction that is typically found in testing effect experiments, however using delayed JOLs as a form of covert testing instead of overt testing (e.g., cued recall tests). These findings support the SFP hypothesis because they demonstrate that delayed JOLs may produce a testing effect similar to that of overt testing (as evidenced by the interaction). This, in turn, indicates that delayed JOLs are likely to elicit covert retrieval attempts that constitute additional learning opportunities (see the General Discussion for more on this).

Study II

Aims

As demonstrated in Study I, covert retrieval (in the form of delayed JOLs) appears to produce testing effects similar to those of overt retrieval, which is the typical form of retrieval in testing effect experiments. The aim of Study II, then, was to examine whether covert and overt retrieval led to comparable retention relative to a study-only condition (Exp. 1) and each other (Exp. 2). Methodologically, the only difference between covert and overt retrieval is that overt retrieval entails retrieval and articulation of information, whereas covert retrieval involves only retrieval of information, but no explicit articulation of it. In order to ensure that the covert and overt retrieval conditions were as comparable as possible, Study II did not use JOLs to elicit retrieval attempts, but instead instructed participants to retrieve the target information either silently or by articulating their response. The relative efficacy of these response formats (i.e., the difference in magnitude of the testing effects they produce) can thus establish whether the testing effect is driven entirely by retrieval processes (in which case there should be no relative efficacy) or if the act of articulation has anything to add to it (in which case I should expect that the testing effect is more pronounced for overt retrieval than covert retrieval).

Procedure

The overall design was highly similar for Experiments 1 and 2 of this study, the main difference being that response format was manipulated either between (Exp. 1) or within (Exp. 2) subjects. The change to a statistically more powerful within-subjects design was due to the small effect sizes typically reported in similar studies (as was also the case in Exp. 1). Both experiments

used two RIs (~19 minutes vs. one week) as a within-subjects independent variable.

Experiment 1

97 participants were divided into three groups (overt testing, n = 33; covert testing, n = 33; study only, n = 31). The learning material consisted of 40 word pairs, of which 20 were the same Swahili-Swedish translations as in Study I, and the other 20 were Swedish association norms from Shaps, Johansson & Nilsson (1976). In each study session, every word pair was displayed individually, in random order, on a computer screen for 6 seconds. Each session had a 30-second distractor task between them. In each testing session, only the cue word was displayed, and the task was to retrieve the target word from memory. For the covert retrieval condition, this was done silently, and for the overt retrieval condition, participants were instructed to type the target word using a keyboard. All groups performed a total of six sessions of either consecutive study (study-only group) or alternating study and testing sessions (3 of each; covert/overt retrieval depending on group), with distractor tasks between sessions. All participants performed cued-recall tests after ~19 minutes and again after one week.

Experiment 2

Similarly to Experiment 1, 40 participants learned 48 word pairs, 24 Swahili-Swedish translations and 24 Swedish association norms. Each participant performed six learning sessions of alternating study (6s per item) and testing, where each testing session contained blocks of 24 items tested covertly and 24 items tested overtly, with full counterbalancing and randomization of items and blocks. Final cued recall tests were administered at ~19 minutes and again after one week.

Results

Experiment 1

Table 1a. *Mean (SD) cued recall performance as a function of retention interval and response format in Experiment 1.*

Response format	Retention interval	
	Short	Long
Study-only	.81 (.19)	.57 (.19)
Covert	.83 (.15)	.62 (.19)
Overt	.83 (.16)	.67 (.19)

There was no main effect of response format, however there was a response format × RI interaction ($F_{2, 94}$= 4.50, η_p^2 = .09, p = .01) which was driven mainly by differences between the study-only condition and the other two testing conditions after a week. Although the overt testing group performed significantly better than the study-only group (t_{62} = 2.08, p = .04), the covert testing did not differ reliably from either group (see Table 1a).

Experiment 2

As evident in Table 1b, there was a main effect of response format, such that overt testing led to significantly better retention than covert testing (M_{overt} = .66, SD_{overt} = .17; M_{covert} = .63, SD_{covert} = .17; $F_{1, 39}$ = 10.76, η_p^2 = .22, p = .002).

Table 1b. *Mean (SD) cued recall as a function of retention interval and response format in Experiment 2.*

Response format	Retention interval	
	Short	Long
Covert	.69 (.16)	.54 (.18)
Overt	.72 (.18)	.58 (.19)

Conclusion

In addition to replicating a testing effect, Study II also demonstrated that the response format (i.e., covert or overt retrieval practice) affects later memory performance. Overt testing led to better retention than covert testing, although in only the second of two experiments, and the effect size was small (d = 18). Thus, it remains to be established whether covert and overt retrieval are indeed comparable in terms of the testing effects they produce, and if so, whether the testing effect is entirely driven by retrieval processes.

Study III

Aims

Study III attempts to further disentangle the effects of retrieval and articulation on later retention. From study I, I learned that covert retrieval produces a testing effect comparable to that of overt retrieval, however Study II indicated that there is a relative efficacy of these two response formats. The aims of Study III, then, are fourfold:

Aim I

To replicate a testing effect via a response format × RI interaction. This was done only to ensure that the chosen learning material would indeed produce a reliable testing effect. For the same reason, only Experiments 1 and 2 included the study-only condition necessary for that comparison.

Aim II

To compare retention following covert and overt retrieval. This was done to replicate, at least in part, the findings of Study II. This comparison was present in all four experiments of Study III.

Aim III

To compare different forms of overt retrieval with respect to later retention. Because of studies that predict both equal (e.g., Putnam & Roediger, 2013; Smith, Roediger & Karpicke, 2013) and different (e.g., Longchamp, Boucard, Gilhodes & Velay, 2006; Mangen & Velay, 2010) magnitudes of the testing effect as a function of response format, the inclusion of overt response formats, associated with varying degrees of embodied cognition, was needed to either support or refute these theories.

Aim IV

To investigate the effects of list order (blocked vs. random) as well as the manipulation of this order (within- vs. between-subjects), in such a way that the effects of retrieval and the effects of articulation can be understood separately. This exploratory manipulation was made in an attempt to uncover possible mediators or moderators of the testing effect, but also because there are previously discussed grounds for expecting such methodological aspects to affect the magnitude of the testing effect in ways that do not pertain to the testing effect itself, such as how participants perceive and anticipate tasks involved in memory testing.

Procedure

All four experiments of Study III used a word list consisting of 48 word pairs taken from the Swedish Association norms by (Shaps, Johansson & Nilsson, 1976) and an initial learning phase consisting of three study sessions in which each word pair was displayed individually on a computer screen for ~5 seconds (6s in Exp. 1, 5s in Exp. 2–4), with short distractor tasks between sessions. The learning phase was followed by a testing phase,

which consisted of additional testing sessions that will be described for each individual experiment.

Experiment 1 and 2

For 32 participants, the word list was randomly divided into four lists of 12 items and tested in different ways depending on response format condition in a within-subjects design. For the study-only condition, all word pairs were displayed individually for 12 seconds. For the covert testing condition, only the cue word was shown and participants asked to retrieve the target word and wait until 12 seconds had passed. The two overt testing conditions (i.e., typing and writing) were identical to the covert condition, except participants were also instructed to articulate their answer within 12 seconds, either by typing it using a keyboard, or by writing it down with pen and paper. After a five-minute distractor task, a final cued recall test for half of all items was given, and after one week, the other half of the items were tested in the same way.

For the 33 participants in Experiment 2, the procedure was identical to that of Experiment 1, except the test phase contained three consecutive testing sessions instead of one.

Experiment 3

In the testing phase, which consisted of three consecutive testing sessions, 42 participants were tested covertly and overtly, in blocked and random list orders, using a within-subjects design. No study-only condition was included, and the overt testing condition only included typing. The test duration for each item was set to 10 seconds. The word list was randomly divided into four lists of 12 items and tested differently, depending on response format condition. For the blocked condition, all items were tested in blocks of covert and overt lists (with counterbalancing), whereas in the random condition, all 48 items were tested covertly and overtly in random order. The final cued recall tests were identical to Experiments 1 and 2.

Experiment 4

The design of Experiment 4 was identical to that of Experiment 3, except for the list order condition, which was manipulated between subjects. Thus, half of the 64 participants were tested covertly and overtly in blocks, and the other half in random order for all items.

Results

Experiment 1

As evident from Table 2, there was a significant main effect of response format ($F_{3,93} = 5.98,\ \eta_p^2 = .16,\ p = .001$), and a response format × RI interaction ($F_{3,93} = 8.97,\ \eta_p^2 = .22, p = .001$). Sidak post-hoc comparisons show that both the main effect and interaction are driven mainly by the recall performance of the study-only condition relative to the other three testing conditions, which did not differ significantly at the long RI.

Table 2. *Mean (SD) cued recall performance as a function of response format and retention interval in Experiment 1.*

Response format	Retention interval	
	Short	Long
Study-only	.80 (.27)	.29 (.22)
Covert	.77 (.27)	.52 (.32)
Type	.85 (.16)	.47 (.32)
Write	.76 (.26)	.50 (.31)

Experiment 2

As in Experiment 1, there was a significant main effect of response format, ($F_{3,96} = 3.73,\ \eta_p^2 = .10,\ p = .0$), as well as a response format × RI interaction, ($F_{3,96} = 20.35,\ \eta_p^2 = .39, p = .001$). Again, these effects were mainly driven by the study-only condition and its difference from the other three response formats (see Table 3).

Table 3. *Mean (SD) cued recall performance as a function of response format and retention interval in Experiment 2.*

Response format	Retention interval	
	Short	Long
Study-only	.85 (.18)	.32 (.20)
Covert	.72 (.24)	.61 (.29)
Type	.77 (.22)	.58 (.27)
Write	.72 (.23)	.55 (.30)

Experiment 3

There was no main effect of list order, and the main effect of response format was near-significant ($F_{1,41} = 3.54$, $\eta_p^2 = .08$, $p = .067$). As before, the response format × RI interaction was significant (5 min: $M_{covert} = .76$; $SD_{covert} = .22$; $M_{overt} = .74$; $SD_{overt} = .22$; 1 week: $M_{covert} = .56$; $SD_{covert} = .27$; $M_{overt} = .66$; $SD_{overt} = .25$; $F_{1,41} = 9.34$, $\eta_p^2 = .19, p = .004$), suggesting that overt retrieval leads to better retention at the long RI. The response format × list order interaction was also significant ($F_{1,41} = 4.20$, $\eta_p^2 = .09, p = .047$), and it was driven by differences at the short RI. Specifically, there appears to be an advantage for blocked lists when testing overtly, and an advantage for random lists when testing covertly (see Fig. 3).

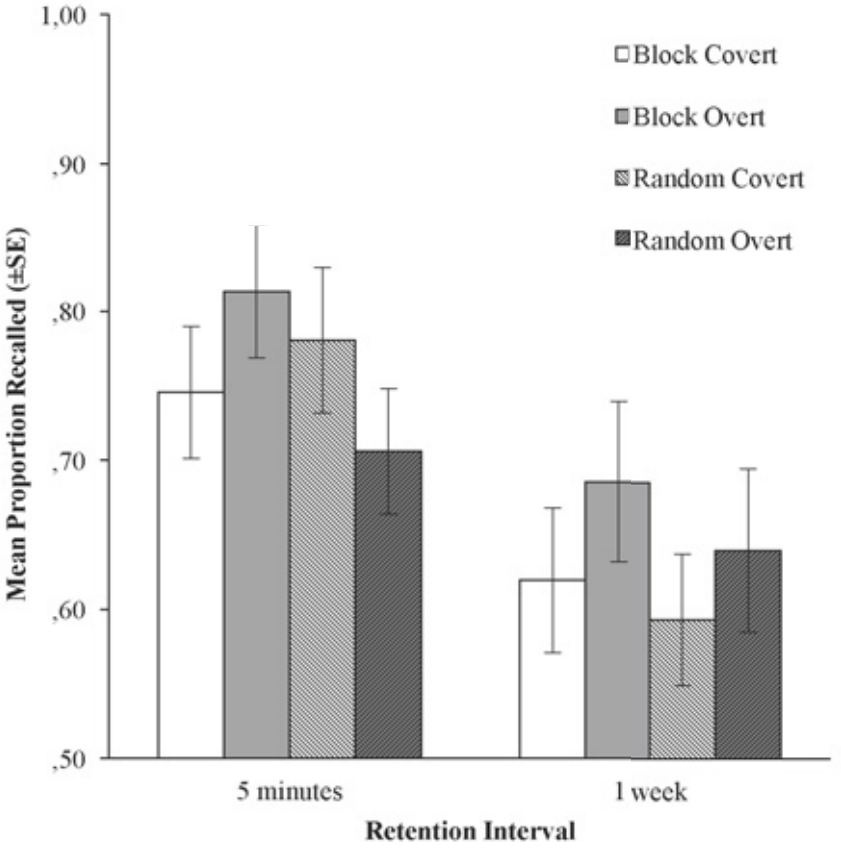


Figure 3. *Mean cued recall performance for items tested, covertly and overtly, in blocks and random order, at the two RIs. Error bars represent the standard errors of the means. Note that the Y axis begins at .50.*

Experiment 4

There were no significant main effects of either response format or testing order, however as in previous experiments, the response format × RI interaction was significant (5 min: M_{covert} = .79; SD_{covert} =.20; M_{overt} = .77; SD_{overt} = .21; 1 week: M_{covert} = .62; SD_{covert} =.21; M_{overt} = .68; SD_{overt} = .21; $F_{1,62} = 7.95$, $\eta_p^2 = .11$, $p = .006$), which indicates a relative efficacy in favor of overt versus covert retrieval at the long RI (see Fig. 4).

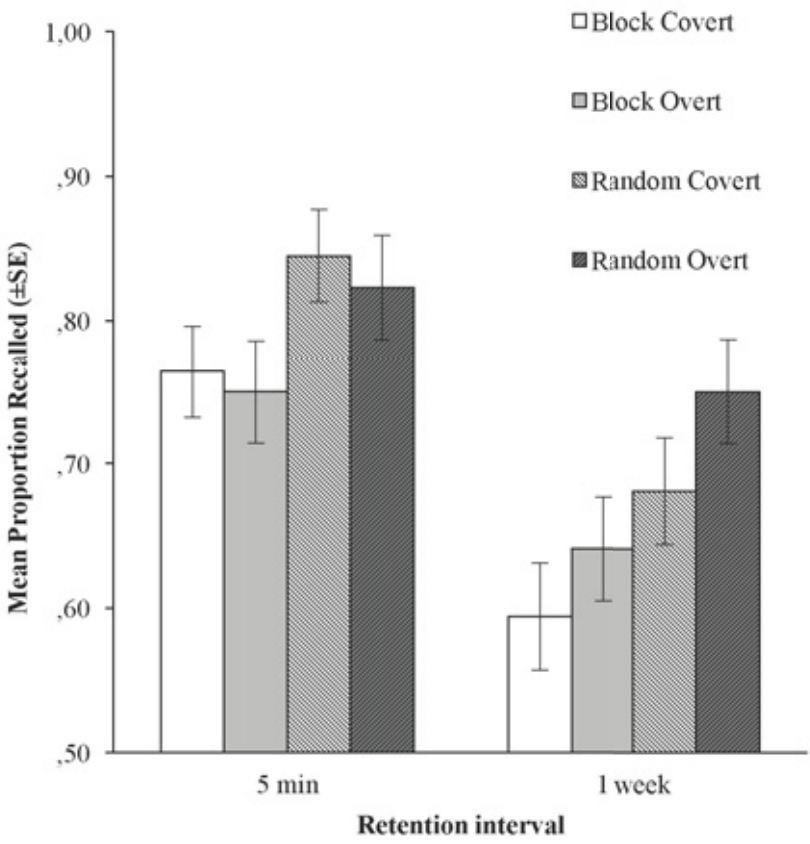


Figure 4. *Mean cued recall performance for items tested, covertly and overtly, in blocks and random order, at the two RIs. Error bars represent the standard errors of the means. Note that the Y axis begins at .50.*

Conclusion

The four experiments of Study III have yielded mixed support for the notion that overt retrieval leads to better retention than covert retrieval. In Experiments 1 and 2, there was no difference in magnitude of the testing effect

produced by the two response formats, whereas Experiments 3 and 4 did find evidence for such a difference. So, on the one hand, I am encouraged to regard the testing effect as driven exclusively by retrieval processes, and on the other, I have indications that articulation may add to the testing effect beyond what is already produced by retrieval alone. By looking at the effect sizes of these experiments, and others like them, it would seem that this added benefit of articulation is too small to have any appreciable impact on the testing effect. Moreover, it still needs to be established whether the relative efficacy of covert versus overt retrieval (whenever such a difference is found) is due to methodological differences (such as the nature of the learning material, its complexity, and the RIs used between tests) rather than processes that pertain specifically to retrieval and articulation, as has been previously assumed.

General discussion

In this thesis, I have presented findings and theories that have different implications for the testing effect, its drivers and its boundary conditions. In the three studies that comprise this thesis, I have demonstrated testing effects produced by covert retrieval, compared covert and overt retrieval with respect to later retention, compared degrees of embodied cognition for the overt retrieval conditions, and finally explored boundary conditions of the testing effect pertaining to experimental designs. The results of these comparisons have practical implications for the ways in which the testing effect can be expected to occur in learning environments, for instance when using tests as a learning aid, or by using flashcards to study for an upcoming exam.

In regards to the first aim (i.e., does covert retrieval produce a testing effect?), there are two experiments in this thesis (i.e., Study I and Exp. 1 of Study II) that support the notion that covert retrieval produces a testing effect, as evidenced by a response format × RI interaction. The findings of Study I are interesting from point of view of the meta-analysis on delayed JOL effects by Rhodes and Tauber (2011), who also found that covert retrieval (i.e., delayed JOLs) produces a testing effect. However, the effect size was very small (Hedge's g = .08), especially when compared to the more robust effects on memory monitoring (i.e, JOL accuracy; g = .93), meaning that the main benefits of delayed JOLs pertain to memory monitoring rather than memory performance. Nevertheless, these comparisons pertain to effects of immediate versus delayed JOLs, meaning that the memorial benefit of retrieval itself (regardless of what prompted it) remains to be ascertained.

Pertaining to the second aim (i.e., is there a relative efficacy of covert versus overt retrieval?), Study I did not include an overt retrieval condition, which prevents us from ascertaining the magnitude of a testing effect produced under more "typical" conditions, for comparative reasons. In other words, there is no way of knowing if repeated (overt) testing would have produced a similar testing effect under the same circumstances. Moreover, it is uncertain whether delayed JOLs and instructed covert retrieval entail similar retrieval efforts (e.g., Tauber, Dunlosky & Rawson, 2015; Jönsson, Olsson & Hedner, 2012), meaning that covert retrieval may be more beneficial to retention than was indicated by the findings of Study I, depending on how it is applied in experimental settings. However, I would argue that while the results of

Study I will not, perhaps, have far-reaching practical consequences for the learning strategies employed by students everywhere (i.e., the JOL condition did not outperform the study-only condition at the long RI), they do demonstrate that covert retrieval (to the extent that delayed JOLs entail it) is indeed associated with a testing effect similar to that which is found when using overt testing formats, which is in line with what is posited by the SFP hypothesis.

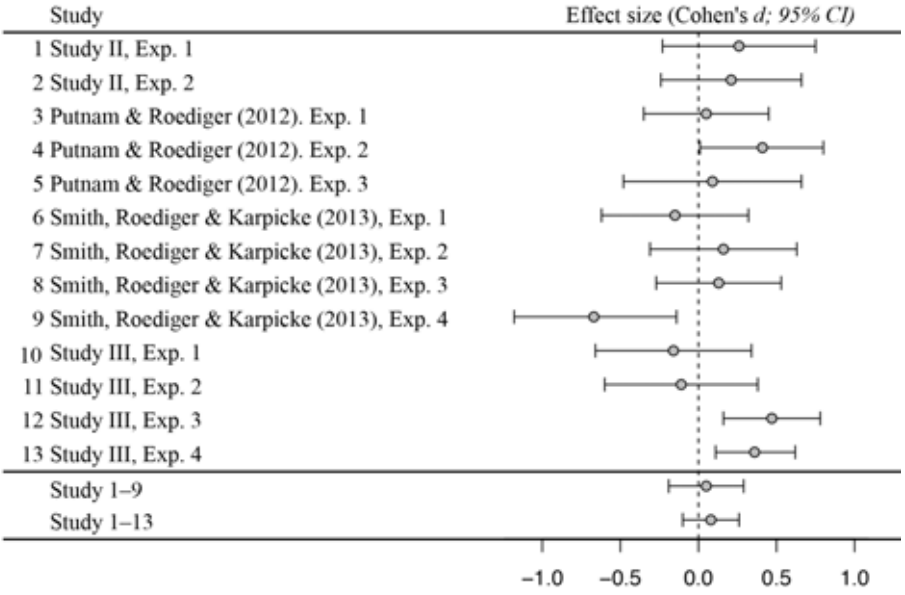
It is important to recall the findings of Tauber, Dunlosky and Rowland (2015), namely that delayed JOLs do not improve memory to the same extent as does overt memory testing, specifically because delayed JOLs are associated with smaller retrieval efforts. Given the design of Study I, it remains uncertain whether its results are consistent with this assertion, simply because there was no overt retrieval condition with which to compare the response latencies. In Study II, however, response latencies were compared for covert and overt retrieval. Specifically, in Experiment 2, response latencies were significantly smaller for the overt condition, but only at the short RI. At the long RI, there was no such difference, indicating that participants are likely to have made equally effortful retrieval attempts when tested covertly and overtly after a week. It should be noted, however, that retrieval latencies during learning were smaller for the overt condition, indicating better learning. For Experiment 1, a typical testing effect was observed, but there was no significant difference in memory performance between overt and covert retrieval. In Experiment 2, such a difference did appear, albeit small. Taken together, these findings suggest that, while it is clear that covert retrieval produces a testing effect (and perhaps more so in the case of instructed covert retrieval than when making JOLs), it is not clear that overt retrieval reliably improves memory more than does covert retrieval. Granted, I have not found any evidence that suggests negative effects of articulation, but that does not prove that articulation should improve memory in cued-recall paradigms. That would essentially be using the absence of negative effects to claim the presence of positive ones. In Study II, Experiment 2 showed that overt retrieval outperformed covert retrieval, however the reported effect size was small (Cohen's $d = .18$)[1]. So, if there are no negative effects of articulation, and the positive effects are so small, what is the reason for promoting retrieval practice that entails articulation? I would say there is no such reason. However, I also see no reason for learners to refrain from articulation altogether, as it does not appear to affect memory adversely.

[1] Because of the within-subjects design of Exp. 2 in Study II, the effect size should be corrected for correlations between measurements (see Morris & DeShon, 2002), which yields a true effect size of $d = .35$. Nonetheless, this effect size is also considered small (Cohen, 1988).

Study III yielded mixed results, such that a relative efficacy in favor of overt retrieval was partially supported by Experiments 3 and 4 (but not Experiments 1 and 2). Similar to the findings of Study I, this effect was demonstrated in the form of a response format × RI interaction, meaning that overt retrieval outperformed covert retrieval only after a long RI (i.e., a week). This finding is interesting when linked to the argument by Toppino and Cohen (2009) that the testing effect itself is typically more pronounced after a longer RI. In Experiments 1 and 2, this is exactly what was found (i.e., a testing effect was replicated), but in Experiments 3 and 4, no study-only condition was included. There, the response format × RI interaction pertains only to the relative efficacy of covert versus overt retrieval, and such a relative efficacy was only observed at the longer RI. In other words, just as the effects of studying and testing are dissociated only after some time has passed, so the effects of retrieval and articulation are dissociated only after a longer RI. This indicates that the overt retrieval condition may benefi later retention, relative to covert retrieval, very similarly to how the testing effect itself is manifested. This line of reasoning is akin to Bjork and Bjork's (1992) distinction between *retrieval strength* (i.e., the current possibility to access certain information) and *storage strength* (i.e., the degree to which something has been learned), where retrieval is reliant on retrieval strength and storage strength determines the rate at which information is forgotten. Thus, because differences were only observed after the longer RI, it would seem that the added element of articulation does not affect the retrieval strength of learned information, but rather the storage strength. In Experiment 2 of Study II, the main effect of response format contradicts this interpretation, however it should be noted that the short RI was approximately 19 minutes, compared to the 5-minute RI used in Study III. Taken together, these observations are all the more reason to thoroughly investigate the importance of the chosen retention intervals in typical testing effect experiments.

Study III includes a limited review of the effect sizes reported in studies that directly compare test effects produced by covert and overt retrieval. This review is by no means exhaustive, but it does provide a reasonable estimate of the true effect size of this comparison. Granted, there are still very few studies that compare covert and overt retrieval. Unlike the testing effect itself, which is very robust (e.g., Roediger & Karpicke, 2006a), the covert/overt relative efficacy comparison yields effect sizes with a weighted average of d = .07 (see Table 4). By contrast, Adesope et al. (2017) compared 118 testing-effect studies in a meta-analysis, and arrived at an overall (testing) effect size of Hedge's g = .61. Moreover, the overall effect size for the testing effect, as reported by Smith, Roediger and Karpicke (2013), was Cohen's d = 1.10, which is considered a large effect.

Table 4. *The relative efficacy of overt versus covert retrieval as reported in 13 experiments (from Study III). A positive effect size denotes a relative efficacy in favor of overt retrieval.*



In regards to the third aim (i.e., does the overt response format matter?), the results of Study III could not confirm any advantages that would arise from processes related to embodied cognition – such as more retrieval routes for handwritten information than for typed information (cf., Mangen & Velay, 2010) – because the two overt response formats did not contribute differentially to the testing effect. However, and as discussed in Study III, this may be the result of difficulties inherent to the task of responding on both keyboard and with pen and paper, such that it was difficult for participants to switch between these two overt response formats, effectively making them preoccupied by the task (or switching thereof). A more thorough investigation of the effects of various ways of articulating responses, (such as typing and writing) will need to be undertaken before they can be definitively ruled out as possible moderators of the testing effect.

Another caveat of Study III pertains to the number of to-be-remembered items (i.e., 48) relative to the number of manipulations (e.g., 2: RI × 2: response format × 2: list order), which leaves only six items to be tested for each condition. Means calculated from so few measurements are prone to measurement errors and unusually large variances, and it appears the standard deviations and standard errors in the four experiments of Study III are larger than those reported in Study II, despite their similar designs. However, the values for skewness and kurtosis are well within ±2 (see Gravetter &

Wallnau, 2014) for all conditions, meaning that data are close to normally distributed, which in turn justifies the chosen statistical analyses.

The fourth aim (i.e., to examine the testing effect magnitude as a function of experimental designs) pertains to some of the issues described above, namely that there may be boundary conditions to the relative efficacy of covert versus overt retrieval. How do we reconcile the disparate findings within this field? For instance, whereas Putnam and Roediger (2013), along with Smith, Roediger and Karpicke (2013), provide evidence that there is no relative efficacy of covert versus overt retrieval, both Studies II and III of this thesis are in partial disagreement with such a conclusion. Study III has specifically investigated factors that may explain these differences from a theoretical perspective.

Firstly, the relative efficacy of covert versus overt retrieval was examined as a function of list order. As evidenced by Rowland, Littrell-Baez, Sensenig and DeLosh (2014), the testing effect itself does not appear to be differentially affected by list order, but that does not definitively prove that covert retrieval is equally effortful as overt retrieval, or even based on the same premises, meaning that list order may indeed contribute to selective effects of response format. The rationale was that if items were tested in blocks, participants could anticipate the way in which subsequent items would be tested and therefore adjust their retrieval efforts or strategies accordingly. For random covert and overt tests, however, this would not be an option (cf., Jonker, Levene & MacLeod, 2014). In Experiments 3 and 4 of Study III, list order had no appreciable impact on covert and overt retrieval, as memory performance did not differ for random or blocked lists. Granted, there was a response format by list order interaction, but only at the short RI. I interpret this as a good ability, on the part of the participants, to follow experimental instructions, that is, they did not seem to adjust their retrieval efforts or strategies based on how (i.e., in what order) the items were tested.

The retrieval efforts associated with covert and overt retrieval were inferred by the response latencies during learning and final testing. Generally, the response latency differences corresponded very well to differences in memory performance, so that better memory performance was associated with smaller response latencies at final recall. In Experiments 1 and 2 of Study III, memory performance was equal for all testing conditions, whereas the study-only condition differed from all three. Similarly, the response latencies did not differ between the testing conditions, but the study-only condition differed significantly from all three. In Experiment 3, the response format × RI interaction was reflected in larger response latencies for covert retrieval than for overt retrieval (albeit for different list orders, depending on retention interval). Finally in Experiment 4, the advantage for overt retrieval

at the long RI was echoed by larger response latencies for covert retrieval at the long RI. This suggests that the advantage for overt versus covert retrieval does not appear to be the result of differences in retrieval effort.

Secondly, the findings of Huff, McNabb and Hutchison (2015) testify to the possible consequences of within- or between-subject manipulations of list order. For this reason, Experiment 4 of Study III compared random and blocked list orders manipulated both within (cf., Exp. 3) and between groups, however results showed that, similarly to Experiment 3, there were no effects of list order whatsoever. Putnam and Roediger (2013) tested the TAP hypothesis by manipulating response format (during learning) within subjects, and final test format between subjects, and found no effects of either response format or final test format, thus refuting the TAP prediction. So, to the extent that Experiment 4 includes conditions that are beneficial from point of view of the TAP prediction, it corroborates the findings of Putnam and Roediger (2013). It should be noted, though, that in their experiments, the covert retrieval condition may have been confounded by JOLs during restudy (effectively making the covert and restudy conditions very similar), but nonetheless, it serves as an example where covert and overt retrieval appear largely unaffected by both within- and between subject manipulations.

In several experiments of this thesis, I find support for the notion that overt retrieval produces a more pronounced testing effect than covert retrieval. However, because of small effect sizes, I am encouraged to regard retrieval as the primary driver behind the test effect, and to conclude that articulation does not boost the testing effect to an appreciable extent. Nonetheless, this assertion is still confined to the typical designs used in testing effect experiments, which are often limited to rather simple materials and tasks (e.g., word pairs and cued recall tests). Therefore, in order to apply these findings in a broader educational setting, I will discuss designs, materials and tasks that better reflect classroom and other real-world learning environments.

Future directions and implications for education

Thus far, this thesis has investigated a number of methodological aspects of typical testing-effect experiments where covert and overt retrieval are compared. The evidence provided here do not definitively rule out the possibility that such aspects may play a role in determining the memorial benefit associated with different response formats. For example, there are still a number of viable candidates, such as the length of the chosen RI, within- and between-subjects manipulation of the response format itself (as opposed to list order, as in Study III), as well as the nature of the learning material itself, for

instance its difficulty and complexity. Given the inconclusive and disparate findings from Studies II and III, along with others that directly compare covert and overt retrieval (e.g. Putnam & Roediger, 2013; Smith, Roediger & Karpicke, 2013), it is essential to understand why and how designs that use similar materials, similar designs, and similar procedures, still produce different results. Such methodological differences may, then, inform us about the boundary conditions of not only the testing effect itself, but also the relative efficacy of covert and overt retrieval.

As mentioned earlier, the testing effect is typically more pronounced after longer RIs (Toppino & Cohen, 2009), and it remains to be established whether the effects of response format are also affected by the length of the RI in a similar way. This position can be explained, in part, by Kornell, Bjork and Garcia's (2011) distribution-based bifurcation model. In this model, testing is assumed to decrease the rate of forgetting relative to information that is not tested (or only studied). This difference in the forgetting rate can be explained in terms of Bjork and Bjork's (1992) storage and retrieval strength hypothesis. So, to the extent that articulation increases storage strength (which, in turn decreases the rate of forgetting), I should expect an advantage for overt versus covert retrieval over increasingly long RIs. Relatedly, Adesope, et al. (2017) report that testing effect studies in classroom settings tend to have longer retention intervals, with 67% of them having RIs between 7 and 42 days, compared to laboratory studies where 59% used RIs shorter than 24 hours. Given the nature and purpose of tests in educational settings, I would argue that studies that use longer RIs better reflect the ongoing learning efforts among students everywhere, compared to shorter RIs in laboratory environments, meaning that the validity of studies that replicate classroom settings is likely higher than laboratory studies. This is a possible venue for future research efforts.

The manipulation of response formats within or between subjects is another viable candidate, insofar as previous studies have used both types of manipulations. However, the findings Putnam and Roediger (2013; Exp. 1 & 2), along with the disparate results Study III, suggest that within- or between-subject manipulations do not selectively affect they way participants perceive the task, and thus engage in similar retrieval processes (as evidenced also by response latencies in Study II). If anything, I would argue that specific instructions pertaining to the covert and overt retrieval conditions (i.e., the extent to which they instruct participants to engage in the same retrieval efforts for both conditions) are just as important in this regard.

Perhaps more so than any other aspect of the testing effect, I would argue that the properties of the learning material itself should be considered, as well as how memory is tested. The testing effect is mainly demonstrated

within the domain of recall tests (either free or cued; but only cued recall tests were used in this thesis) of paired associates, that is, word pairs that are either semantically related to some degree, or translations from a native to a foreign language. The question of whether the effects, that are observed when learning and testing paired associates, are also observed for more complex materials is crucial to our understanding of the testing effect and how it applies to situations outside of the laboratories (cf., Kang, McDermott & Roediger, 2007, who observed testing effects for multiple-choice and short-answer question materials). For instance, I would argue that in order to do well on a final exam, students will need to memorize materials that are more challenging (in terms of their complexity and what is needed to understand them) than paired associates. This means that even if they adopt learning strategies known to produce testing effects, they will still only benefit from those testing effects to the extent that they apply to the learning material itself. Thus, if testing effects are observed for paired associates, but perhaps not for more complex materials, this severely limits the implications of the testing effect in an educational setting.

Recently, Tauber and colleagues (2016) provided very promising evidence in this regard. Specifically, they compared covert and overt retrieval for key-term definitions, which can be considered a more complex learning material. Final recall took place after 48 hours and in the first of two experiments, the overt retrieval group had retained more key-term definitions than the covert group. Interestingly, in the second experiment, the instructions for the covert retrieval condition were altered to more explicitly instruct participants to silently retrieve but not articulate the answer (i.e., the definition of a term). With this change, the relative efficacy in favor of overt retrieval disappeared. These results indicate that the testing effect may indeed be observed for more complex materials, and they raise an interesting question in regards to what covert retrieval actually is. As argued by Putnam and Roediger (2013), the benefit of overt retrieval, relative to covert retrieval, appears to diminish as the two response formats are made to procedurally resemble each other. In other words, what Tauber and colleagues (2016) refer to as "enhanced covert retrieval" is an effortful retrieval attempt because participants were explicitly instructed *not* to rely on the level of familiarity with the cue, but rather retrieve the entire information (albeit silently). This is substantial evidence for the notion that as long as the retrieval processes involved during learning are the same (i.e., retrieval is effortful), so too are the testing effects for different response formats (cf., Hyde & Jenkins, 1973). This notion is also consistent with the previously presented ideas on subvocal articulation (cf., Baddeley, 2001), as the "enhanced covert retrieval" condition clearly encourages subvocal articulation to a greater extent than does ordinary covert retrieval, meaning that it becomes more similar to overt retrieval. So, in addition to understanding effects of the learning material itself, it seems future research

should also focus on separating articulatory processes from retrieval processes, possibly by suppressing subvocal rehearsal during retrieval. In light of the findings presented above, I would call for a shift from typical testing effect designs that feature simple materials and tasks, measured over short periods of time, to designs that feature complex learning materials and tasks, as well as longer RIs that better reflect how learning occurs in real-life settings.

Concluding remarks

Based on the findings of the three studies of this thesis, and in light of the much larger effect sizes reported for testing effect manipulations other than response format, it should be seriously considered whether the comparison between covert and overt retrieval, and their effects on subsequent retention, is nothing more than a laboratory phenomenon with no appreciable implications for real-world educational settings. From here on, focus should instead be turned towards the learning materials themselves, and the extent to which they reflect the challenges posed to learners everywhere on a daily basis.

Most testing effect studies have, until fairly recently, almost exclusively used overt response formats, and the results have directly impacted the development of learning techniques and strategies among students and teachers. Indeed, Dunlosky and colleagues (2013) reviewed the effectiveness of different learning strategies, and concluded that retrieval practice was one of few strategies that reliably improved later retention. However, as many students likely engage in covert retrieval during learning, it has remained uncertain whether students are, in fact, enjoying the full benefits of the testing effect. As this thesis has shown, it appears they are – as long as they engage in retrieval practice.

With respect to underlying processes, the findings reported here suggest that, within a paired-associates, cued-recall paradigm, the testing effect is almost exclusively driven by retrieval, and that articulation adds very little to its magnitude under the circumstances of a typical testing effect experiment. That is not to say that articulation cannot be beneficial for retention altogether, however from point of view of the testing effect, it seems that retrieval is paramount.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. *Review of Educational Research*. http://doi.org/10.3102/0034654316689306.

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*(7), 861–876. http://doi.org/10.1002/acp.1391

Ausubel, D. P., & Youssef, M. (1965). The effect of spaced repetition on meaningful retention. *The Journal of General Psychology*, *73*(1), 147–150. http://doi.org/10.1080/00221309.1965.9711263

Baddeley, A. (2000). The episodic buffer: a new component of working memory?. *Trends in cognitive sciences, 4*(11), 417–423. http://doi.org/10.1016/s1364-6613(00)01538-2

Baddeley, A. D. (2001). Is working memory still working?. *American Psychologist*, *56*(11), 851–864. http://doi.org/10.1037/0003-066x.56.11.851

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of learning and motivation, 8*, 47–89. http://doi.org/10.1016/s0079-7421(08)60452-1

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & cognition*, *35*(2), 201–210. http://doi.org/10.3758/bf03193441

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. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *From learning processes to cognitive processes: Essays in honor of William K. Estes*, *2*, 35–67.

Bransford, J. D., Franks, J. J., Morris, C. D., & Stein, B. S. (1979). Some general constraints on learning and memory research. In L. Cermak & F. Craik (Eds.), *Levels of processing in human memory* (pp. 331–354). Hillsdale, NJ: Lawrence Erlbaum.

Carpenter, S. K., & DeLosh, E. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*, 268–276. http://doi.org/10.3758/bf03193405

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14*(3), 474–478. http://doi.org/10.3758/bf03194092

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test?. *Psychonomic Bulletin & Review, 13*(5), 826–830. http://doi.org/10.3758/bf03194004

Carrier, M. & Pashler, H. (1992). The influence of retrieval on attention. *Memory & Cognition, 20*, 633–642. http://doi.org/10.3758/bf03202713

Cohen, J. (1988). Statistical power analysis for the behavioural sciences. Hillside. *NJ: Lawrence Earlbaum Associates*. http://doi.org/10.4324/9780203771587

Cohen, R. L. (1981). On the generality of some memory laws. *Scandinavian Journal of Psychology*, *22*(1), 267–281. http://doi.org/10.1111/j.1467-9450.1981.tb00402.x

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory?. *Progress in brain research*, *169*, 323–338. http://doi.org/10.1016/s0079-6123(07)00020-9

Craik, F. I. (2002). Levels of processing: Past, present... and future?. *Memory*, *10*(5-6), 305–318. http://doi.org/10.1080/09658210244000135

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior, 11*(6), 671–684. http://doi.org/10.1016/s0022-5371(72)80001-x

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294. http://doi.org/10.1037//0096-3445.104.3.268

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. *Memory, 10*, 317–344. http://doi.org/10.1016/b978-012102570-0/50011-2

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition, 20*(4), 374–380. http://doi.org/10.3758/bf03210921

Ebbinghaus, H. (1885). *Über das gedächtnis: untersuchungen zur experimentellen psychologie*. Duncker & Humblot.

Elwood, R. W. (1997). Episodic and semantic memory components of verbal paired-associate learning. *Assessment*, *4*(1), 73–77. http://doi.org/ 10.1177/107319119700400110

Engelkamp, J., & Krumnacker, H. (1980). Image-and motor-processes in the retention of verbal materials. *Zeitschrift für experimentelle und angewandte Psychologie*, *27,* 511–533.

Foster, J. K., & Jelicic, M. E. (1999). *Memory: Systems, process, or function?*. Oxford University Press. http://doi.org/10.1093/acprof:oso/9780198524069.001.0001

Gardiner, J. M., Passmore, C., Herriot, P., & Klee, H. (1977). Memory for remembered events: Effects of response mode and response-produced feedback. *Journal of verbal learning and verbal behavior, 16*(1), 45–54. http://doi.org/10.1016/s0022-5371(77)80006-6

Gravetter, F., & Wallnau, L. (2014). *Essentials of statistics for the behavioral sciences* (8th ed.). Belmont, CA: Wadsworth.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392–399. http://doi.org/10.1037//0022-0663.81.3.392

Godden, D. R., & Baddeley, A. D. (1975). Context‐dependent memory in two natural environments: On land and underwater. *British Journal of psychology*, *66*(3), 325–331. http://doi.org/10.1111/j.2044-8295.1975.tb01468.x

Hirshman, E., & Bjork, R. A. (1988). The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 484–494. http://doi.org/10.1037//0278-7393.14.3.484

Huff, M. J., McNabb, J., & Hutchison, K. A. (2015). List blocking and longer retention intervals reveal an influence of gist processing for lexically ambiguous critical lures. *Memory & cognition*, *43*(8), 1193–1207. http://doi.org/10.3758/s13421-015-0533-3

Hyde, T. S., & Jenkins, J. J. (1973). Recall for words as a function of semantic, graphic, and syntactic orienting tasks. *Journal of Verbal Learning and Verbal Behavior*, *12*(5), 471–480. http://doi.org/10.1016/s0022-5371(73)80027-1

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of verbal learning and verbal behavior, 17*(6), 649–667. http://doi.org/10.1016/s0022-5371(78)90393-6

Jang, Y., & Nelson, T. O. (2005). How many dimensions underlie judgments of learning and recall? Evidence from state–trace methodology. *Journal of Experimental Psychology: General, 134*, 308–326. http://doi.org/10.1037/0096-3445.134.3.308

Jonker, T. R., Levene, M., & MacLeod, C. M. (2014). Testing the item-order account of design effects using the production effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 441–448. http://doi.org/10.1037/a0034977

Jönsson, F. U., Hedner, M., & Olsson, M. J. (2012). The Testing Effect as a Function of Explicit Testing Instructions and Judgments of Learning. *Experimental Psychology*, *59*(5), 251–257. http://doi.org/10.1027/1618-3169/a000150

Kang, S. H. K. (2010). Enhancing visuospatial learning: the benefit of retrieval practice. *Memory & Cognition, 38*, 1009–1017. http://doi.org/10.3758/mc.38.8.1009

Kang, S. H., 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*(4-5), 528–558. http://doi.org/10.1080/09541440601056620

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. http://doi.org/10.1016/j.jml.2006.09.004

Karpicke, J. D. & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*, 966–968. http://doi.org/10.1126/science.1152408

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language*, *67*(1), 17–29. http://doi.org/10.1016/j.jml.2012.02.004

Kato, C., Isoda, H., Takehar, Y., Matsuo, K., Moriya, T., & Nakai, T. (1999). Involvement of motor cortices in retrieval of kanji studied by functional MRI. *Neuroreport, 10*, 1335–1339. http://doi.org/10.1097/00001756-199904260-00033

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*(2), 85–97. http://doi.org/10.1016/j.jml.2011.04.002

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62*(3), 227–239. http://doi.org/10.1016/j.jml.2009.11.010

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*(5), 493–501. http://doi.org/10.1080/09658210902832915

Kubik, V., Olofsson, J. K., Nilsson, L. G., & Jönsson, F. U. (2016). Putting action memory to the test: testing affects subsequent restudy but not long-term forgetting of action events. *Journal of Cognitive Psychology*, *28*(2), 209–219. http://doi.org/10.1080/20445911.2015.1111378

Kuo, T. M., & Hirshman, E. (1996). Investigations of the testing effect. *The American Journal of Psychology, 109*(3), 451–464. http://doi.org/10.2307/1423016

Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory?. *Journal of experimental psychology: Learning, Memory, and Cognition, 16*(3), 464–467. http://doi.org/10.1037//0278-7393.16.3.464

Longcamp, M., Boucard, C., Gilhodes, J. C., & Velay, J. L. (2006). Remembering the orientation of newly learned characters depends on the associated writing knowledge: A comparison between handwriting and typing. *Human Movement Science*, *25*(4), 646–656. http://doi.org/10.1016/j.humov.2006.07.007

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 671–685. http://doi.org/10.1037/a0018785

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological review, 87*(3), 252–271. http://doi.org/10.1037//0033-295x.87.3.252

Mangen, A., & Velay, J. L. (2010). Digitizing literacy: reflections on the haptics of writing. *INTECH Open Access Publisher*. http://doi.org/10.5772/8710

Mangen, A., Anda, L. G., Oxborough, G. H., & Brønnick, K. (2015). Handwriting versus Keyboard Writing: Effect on Word Recall. *Journal of Writing Research, 7*(2), 227–247. http://doi.org/10.17239/jowr-2015.07.02.1

Matsuo, K., Kato, C., Okada, T., Moriya, T., Glover, G. H., & Nakai, T. (2003). Finger movements lighten neural loads in the recognition of ideographic characters. Cognitive Brain Research, 17(2), 263–272. http://doi.org/10.1016/s0926-6410(03)00114-9

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. http://doi.org/10.1080/09541440701326154

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, *16*(2), 192–201. http://doi.org/10.1016/0361-476x(91)90037-l

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(2), 371–385. http://doi.org/10.1037//0278-7393.11.2.371

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological methods*, *7*(1), 105–125. http://doi.org/10.1037//1082-989x.7.1.105

Naka, M., & Naoi, H. (1995). The effect of repeated writing on memory. *Memory & cognition*, *23*(2), 201–212. http://doi.org/10.3758/bf03197222

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109–133. http://doi.org/10.1037//0033-2909.95.1.109

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science, 2*, 267–270. http://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory*, *2*(3), 325–335. http://doi.org/10.1080/09658219408258951

Nelson, T. O. & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). Cambridge, MA: Bradford Books.

Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A revised methodology for research on metamemory: Pre-judgment recall and monitoring (PRAM). Psychological Methods, 9, 53–69. http://doi.org/10.1037/1082-989x.9.1.53

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

Ozubko, J. D., & MacLeod, C. M. (2010). The production effect in memory: Evidence that distinctiveness underlies the benefit. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1543–1547. http://doi.org/10.1037/a0020604

Putnam, A. L., & Roediger III, H. L. (2013). Does response mode affect amount recalled or the magnitude of the testing effect?. *Memory & Cognition, 41*(1), 36–48. http://doi.org/10.3758/s13421-012-0245-x

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: a meta-analytic review. *Psychological bulletin, 137*(1), 131–148. http://doi.org/10.1037/a0021705

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

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

Roediger, H.L. III & Karpicke, J.D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255. http://doi.org/10.1111/j.1467-9280.2006.01693.x

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. http://doi.org/10.1037/a0037559

Rowland, C. A., Littrell-Baez, M. K., Sensenig, A. E., & DeLosh, E. L. (2014). Testing effects in mixed-versus pure-list designs. *Memory & cognition*, *42*(6), 912–921. http://doi.org/10.3758/s13421-014-0404-3

Schacter, D.L., Wagner, A.D., & Buckner, R.L. (2000). Memory systems of 1999. In E. Tulving & F. Craik (Eds.) *Handbook of memory*. New York: Oxford University Press.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional science, 26*(1-2), 113–125. http://doi.org/10.1023/a:1003044231033

Shaps, L. P., Johansson, B. S., & Nilsson, L. G. (1976). Swedish Association Norms. (*Report No. 196*). Uppsala: Department of Psychology, Uppsala University.

Slamecka, N. J., & Graf, P. (1978). The generation effect: delineation of a phenomenon. *Journal of experimental Psychology: Human learning and Memory, 4*(6), 592–604. http://doi.org/10.1037//0278-7393.4.6.592

Slamecka, N. J., & Katsaiti, L. T. (1988). Normal forgetting of verbal lists as a function of prior testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 716–727. http://doi.org/10.1037//0278-7393.14.4.716

Smith, M. A., Roediger III, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(6), 1712–1725. http://doi.org/10.1037/a0033569

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic bulletin & review*, *8*(2), 203–220. http://doi.org/10.3758/bf03196157

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science, 3*, 315–316. http://doi.org/10.1111/j.1467-9280.1992.tb00680.x

Squire, L. R. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiology of learning and memory*, *82*(3), 171–177. http://doi.org/10.1016/j.nlm.2004.06.005

Surprenant, A.M., & Neath, I. (2009). The 9 lives of short-term memory. In A. Thorn & M. Page (Eds.), *Interactions between short-term and long-term memory in the verbal domain* (pp. 16-43). Hove, England: Psychology Press. http://doi.org/10.4324/9780203938966

Tauber, S. K., Dunlosky, J., Rawson, K. A., Wahlheim, C. N., & Jacoby, L. L. (2013). Self-regulated learning of a natural category: Do people interleave or block exemplars during study?. *Psychonomic bulletin & review*, *20*(2), 356–363. http://doi.org/10.3758/s13423-012-0319-6

Tauber, S. K., Witherby, A. E., Dunlosky, J., Rawson, K. A., Putnam, A. L., & Roediger, H. L. (2017). Does Covert Retrieval Benefit Learning of Key-Term Definitions?. *Journal of Applied Research in Memory and Cognition*. http://doi.org/10.1016/j.jarmac.2016.10.004

Toppino, T. C. & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology, 56*, 252–257. http://doi.org/10.1027/1618-3169.56.4.252

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological review, 80*(5), 352–373. http://doi.org/10.1037/h0020071

Vaughn, S., Schumm, J. S., & Gordon, J. (1992). Early spelling acquisition: Does writing really beat the computer?. *Learning Disability Quarterly, 15*(3), 223–228. http://doi.org/10.2307/1510245

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

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, *9*(4), 625–636. http://doi.org/10.3758/bf03196322

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of memory and language, 46*(3), 441–517. http://doi.org/10.1006/jmla.2002.2864

Zimmer, H. D., & Cohen, R. L. (2001). Remembering Actions: A specific type of memory? In Zimmer, H. D., Cohen, R. L., Guynn, M. J., Engelkamp, J., Kormi-Nouri, R., & Foley, M. A. (2001). *Memory for Action: A Distinct Form of Episodic Memory?* (pp. 3–24). Oxford University Press.

Zimmer, H. D., & Engelkamp, J. (2003). Signing enhances memory like performing actions. *Psychonomic Bulletin & Review*, *10*(2), 450–454. http://doi.org/10.3758/bf03196505