

# Elaboration and the Testing Effect in Cued Recall

Leonora C. Coppens



# Elaboration and the Testing Effect in Cued Recall

Leonora C. Coppens

© 2014 L.C. Coppens  
ISBN: 978-90-5335-833-7

Cover design by L.C. Coppens  
Layout by L.C. Coppens  
Printed by Ridderprint BV, Ridderkerk, the Netherlands

All rights reserved. No part of this dissertation may be reproduced or transmitted in any form, by any means, electronic or mechanical, without the prior permission of the author, or where appropriate, of the publisher of the articles.

# Elaboration and the Testing Effect in Cued Recall

Elaboratie en het testeffect in cued recall

**Proefschrift**

ter verkrijging van de graad van doctor aan de Erasmus Universiteit Rotterdam

op gezag van de rector magnificus

Prof.dr. H.A.P. Pols

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

donderdag 22 mei 2014 om 11.30 uur

door

Leonora Carolina Coppens

geboren te Capelle aan den IJssel.



# Promotiecommissie

## **Promotor**

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

## **Overige leden**

Dr. G. Camp

Dr. K. Dijkstra

Prof.dr. T.A.J.M. van Gog

## **Copromotor**

Dr. P.P.J.L. Verkoeijen

# Contents

Chapter 1	General introduction	7
Chapter 2	Learning Adinkra symbols: The effect of testing	19
Chapter 3	The neurophysiology of retrieval practice versus restudy: ERP correlates of the testing effect	31
Chapter 4	No ERP evidence for increased elaboration during retrieval practice	47
Chapter 5	No evidence for the semantic mediator hypothesis: The testing effect in cued recall is similar for mediator cues and related cues	61
Chapter 6	Summary and Discussion	81
	Samenvatting (summary in Dutch)	95
	References	103
	Dankwoord (acknowledgements in Dutch)	111
	Curriculum Vitae	115





# 1

## General introduction

1

You will most likely remember learning lists of foreign vocabulary words (such as French) in high school. Most of my classmates and I did this by reading the entire list of words once and then covering the column with French words with a piece of paper. Reading one Dutch word at a time, I tried to remember the French equivalent and checked whether my answer was correct. If it was, I put a small note next to the word to signify that I did not have to study that word anymore. And if I was wrong, I tried to imprint the right answer and tested myself again on that word in the next cycle. By doing this, the cycles became shorter and shorter until I was convinced that I would remember all the words during the exam. Without knowing it, I was using the effect of testing to my advantage. I then thought testing myself during learning was just a means of assessing which words deserved more attention. However, I now know that the act of retrieving information strengthens memory. This is called the testing effect: information that is tested is remembered better than information that is only restudied. Research into the testing effect often uses cue-target pairs, such as word pairs. The testing effect in learning cue-target pairs is the topic of this dissertation.

In sum, the testing effect refers to the result of taking an intervening test (often called retrieval practice) during learning on later performance on a final memory test. In the short term, for example on a final test just a few minutes after learning, the effect of intervening testing is negative: people perform worse on a short term final test after intervening testing than after repeatedly studying the materials. However, after several hours or days this pattern reverses: people perform better on a final test after intervening testing than after studying. Thus, testing after an initial study episode improves long-term memory (e.g., Hogan & Kintsch, 1971; Roediger & Karpicke, 2006; Toppino & Cohen, 2009).

### **History of testing effect research**

The first large-scale study into the testing effect was conducted by Gates (1917). He had children in the first, fourth, sixth, and eighth grade study nonsense syllables such as *pib*, *bah*, *rem*, and *lor*. The total study time was kept constant while the percentage of study time spent on recitation of the syllables varied from 0 to 80 percent. This recitation involved trying to recall as much of the syllables as possible while looking away from the list. However, the children were allowed to look at the list during recitation when they could not remember a syllable. Three to four hours after studying, the children wrote down as much of the syllables as they could remember. The children in the first grade performed poorly overall, and their performance was worse when a large portion of learning time was devoted to recitation. Gates explains

this by noting that these children were tested in larger (possibly noisier) groups than the older children, and that they were not used to this kind of learning tasks. The children in the fourth, sixth, and eighth grade, however, showed a clear testing effect: the more time was spent on testing, the more they recalled on the final test.

After Gates and another large-scale testing study by Spitzer (1939), decades followed in which there was little interest in the testing effect. After a publication by Tulving (1967), in which he showed that a test trial produces as much learning as a restudy trial, a small revival of testing effect research followed in the 1970s (e.g., Birnbaum & Eichner, 1971; Donaldson, 1971; Gardiner, Craik, & Bleasdale, 1973). For instance, Hogan and Kintsch (1971) tested the effect of two types of intervening tests on retention. In Experiment 1, participants either studied forty words once and then took three recall or recognition tests (Study-Test-Test-Test, STTT) or studied the words three times and took one recall or recognition test (SSST). Two days later, participants took a final test that involved either recall or recognition. Performance on a recall final test was better after testing through recall or recognition, while performance on a recognition final test was better after restudy or recognition tests. In Experiment 2 these results were confirmed using more extreme conditions: participants learned 40 words through either four study trials (SSSS) or one study trial and three recall trials (STTT). Again, when the final test involved recall, performance was better after testing than after restudying. By contrast, when the final test involved recognition, performance was better after restudying than after testing. The authors conclude that final retrieval performance is improved more by retrieval practice than by restudying.

After the early 1970s, again a few decades went by without much testing effect research. Fifty years after Spitzer's publication, a paper by Glover (1989) was published on the topic with the fitting subtitle 'Not gone but nearly forgotten', which again caused some renewed interest in the effect (e.g., Bangert-Drowns, Kulik, & Kulik, 1991; Carrier & Pashler, 1992; McDaniel & Fisher, 1991; McDaniel, Kowitz, & Dunay, 1989). However, a real boom in testing effect research followed the publication of two papers by the same authors: a review of literature about the testing effect (Roediger & Karpicke, 2006a) and an empirical study (Roediger & Karpicke, 2006b). The empirical study confirmed the testing effect in learning prose passages. In Experiment 1, participants studied prose passages and either restudied them once or retrieved them once through free recall. The final free recall test was administered after five minutes, two days or one week. After five minutes, participants recalled more idea units of the repeatedly studied passage, whereas after two days and after one week, participants recalled more idea units of the tested passage. In Experiment 2

this result was confirmed using a design similar to that of Hogan and Kintsch (1971). Participants learned prose passages through four study trials (SSSS), through three study trials and one test trial (SSST) or through one study trial and three test trials (STTT). The final test was administered either after five minutes or after one week. Results showed that after five minutes, the passages were better recalled when they had been studied more often. By contrast, after a week the passages were better recalled when they had been tested more often. The authors concluded on the basis of their findings that testing is a powerful way to improve long-term memory.

### **Mechanisms underlying the testing effect**

Looking at the testing effect research conducted over the years, it becomes clear that the advantage of testing over restudying is ubiquitous. The testing effect has been found in experiments using various to be learned materials, such as words or word pairs (e.g., Allen, Mahler, & Estes, 1969; Carpenter, Pashler, Wixted, & Vul, 2008; Darley & Murdock, 1971; Pyc & Rawson, 2009), texts (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Chan, McDermott, & Roediger, 2006; Duchastel, 1981; Kintsch, 1994; Roediger & Karpicke, 2006a), sentences or facts (McDaniel & Fisher, 1991), videotaped lectures (Butler & Roediger, 2007), maps (Carpenter & Pashler, 2007), resuscitation skills (Kromann, Jensen, & Ringsted, 2009), pictures of faces (Carpenter & DeLosh, 2005), and it has been established with educationally relevant materials (Glover, 1989; McDaniel, Anderson, Derbish, & Morrisette, 2007; McDaniel & Fisher, 1991). Various types of intervening tests have been investigated (Duchastel, 1981; McDaniel et al., 2007) and the testing effect has been found when the final test involves free recall (e.g., McDermott, 2006; Whitten & Bjork, 1977), cued recall (e.g., Carrier & Pashler, 1992; Toppino & Cohen, 2009), and recognition (Carpenter, 2011; Jacoby, Wahlheim, & Coane, 2010). Furthermore, the testing effect has been established in learners of different ages such as children (e.g., Bouwmeester & Verkoeijen, 2011; Gates, 1917), college students (e.g., Pyc & Rawson, 2010; Tulving, 1967), and older adults (e.g., Coane, 2013).

As apparent from the paragraph above, the practical aspects of the testing effect have been well investigated and the testing effect has proven to be a very robust empirical phenomenon. However, the cognitive mechanisms underlying the testing effect are less clear. A first attempt to understand the emergence of the testing effect was made by proposing that testing was simply a form of additional exposure to the materials (e.g., Slamecka & Katsaiti, 1988). In most of the early studies, there was no restudy control condition to compare the testing condition to: study plus testing was

compared to study only (e.g., Erdelyi & Kleinbard, 1978; Roediger & Thorpe, 1978; Wheeler & Roediger, 1992). This setup makes the results difficult to interpret, because compared to doing nothing a test provides additional exposure to the to be learned materials. Thus the 'testing effect' could be caused not by testing, but by additional exposure during the test. However, in subsequent research testing was compared to restudying for an equal amount of time. Results showed that there is an advantage of testing over restudying even when the testing and the restudying condition are equal in terms of total learning time (e.g., Carrier & Pashler, 1992). Moreover, the fact that repeated study yields a memory advantage after a short retention interval cannot be explained by additional exposure. Thus the additional exposure explanation of the testing effect does not seem to hold.

By now, two major classes of theories have been put forward to explain the testing effect in cue-target pairs: transfer appropriate processing and retrieval effort. Transfer appropriate processing will be briefly discussed, and then retrieval effort and more specific retrieval effort theories based on elaboration will be described. Elaboration is often supposed to play a role in the emergence of the testing effect, but there is little direct evidence of enhanced elaboration during testing. Therefore, theories of the testing effect based on elaboration will be the focus of this dissertation.

### **Transfer appropriate processing**

According to this class of theories, a test is not beneficial for memory, but for retrieval. Taking a test during learning is a way of practicing retrieval; a way of training for the test. If information has been retrieved earlier, then this retrieval practice makes it easier to retrieve the information again on a later test. Following this line of reasoning, performance on a final test should be best when the final test highly resembles the intervening tests taken during learning (Thomas & McDaniel, 2007). Thus, performance on a final test that involves free recall (i.e., no retrieval cue is provided and participants simply write down all that they can recall) should be better when the intervening test also involves free recall. And vice versa, performance on a recognition final test (i.e., in the case of words, the word is given and the participant indicates whether the word was presented before or not) should be better after a recognition intervening test. However, this is not the case. Intervening tests are most effective when they involve recall (such as in word stem completion tasks, free recall or cued recall) instead of recognition. This is true regardless of the type of final test that is administered. For instance, in a follow-up of Glover (1989), Carpenter and DeLosh (2006, Experiment 1) used three different types of test as intervening and final tests. Participants learned word lists through restudy, recognition testing (i.e., identify

which items were presented before), cued recall testing (i.e., given the first letter of every word, fill in the rest of the words) and free recall testing (i.e., write down all items on a blank sheet of paper). After a retention interval of five minutes, the participants took one of three kinds of final test, corresponding with the intervening test forms. The transfer-appropriate processing account predicts that final test performance would be best when the intervening test matches the final test. However, final test performance was best after a free recall intervening test, regardless of final test type. Even when the final test involved recognition, performance on that final test was better when the intervening tests relied on recall. These results go against the transfer-appropriate processing account of the testing effect.

Another problem with the transfer-appropriate processing account is that it cannot explain the fact that the testing effect is larger after a long retention interval. According to the transfer-appropriate processing account, retrieval of information is easier when the processes during learning match those during the final test. This should be true regardless of the retention interval. However, as discussed above, the testing effect is larger after a long retention interval; generally there is even an advantage of restudying on a short-term test.

### **Retrieval effort**

Bjork (1994) proposed the desirable difficulties framework. According to this framework, techniques that slow down initial learning but improve long-term memory create desirable difficulties. Testing is a desirable difficulty, because compared to restudying it appears to slow down learning (directly after learning, restudied items are recalled better than tested items) but improves memory measured after a long retention interval. An example of a testing effect account based on the desirable difficulties framework is the effortful retrieval hypothesis (Pyc & Rawson, 2009). This hypothesis states that retrieval of information is beneficial for memory, because retrieval requires more effort than restudying. To test this idea, researchers used varying intervals between the first study episode and a subsequent intervening test in order to manipulate retrieval effort (e.g., Jacoby, 1978; Karpicke & Roediger, 2007; Modigliani, 1976). The assumption here is that after such a delay the learned information is more difficult to retrieve and therefore it is remembered better upon successful retrieval. The studies show that the testing effect is larger when the study-test interval is longer, thus providing support for the effortful retrieval hypothesis and the desirable difficulties framework. For instance, Karpicke and Roediger (2007, Experiment 3) manipulated the schedule of repeated tests as well as the spacing between learning and the first test. The spacing between studying and testing was

either 0 (i.e., each item was tested immediately after the first presentation) or 5 (i.e., five other trials occurred between the first presentation of an item and the first intervening test of that item). On the final test, retention was better with a spacing of 5 items than with no spacing. The authors explain this in terms of effortful retrieval: difficult but successful retrieval is more beneficial for memory than easy successful retrieval.

Another result that is in line with the desirable difficulties approach is the finding that the testing effect is larger when people need more time to answer questions on an intervening test (Gardiner et al., 1973). When answering a question requires more time, this is presumably caused by retrieval difficulty: the more difficult it is to find the answer to a question, the more time it takes to answer the question. Thus the longer response times on the intervening test that Gardiner et al. (1973) found, suggest that more effort is being put into retrieving the answer. Because final test performance was better after longer response times on the intervening test, this suggests that the more effort is required during the intervening test, the better the information is retained.

There are two main versions of the effortful retrieval hypothesis. The first version proposes that it is the difficulty of retrieval that counts. According to this hypothesis, when people are retrieving information they get into a 'retrieval mode' and hence they are better able to remember the information (Dempster, 1996). So, according to this version of the effortful retrieval hypothesis, retrieval effort is not item-dependent: non-tested items are also remembered better, as long as some items are tested so the learner gets into the 'retrieval mode.' The second version of the effortful retrieval hypothesis, however, states that the increased effort during testing effect is caused by item-specific elaboration of memory traces and retrieval routes (Bjork, 1994; McDaniel et al., 1989; McDaniel & Masson, 1985). This version of the effortful retrieval hypothesis inspired testing effect theories based on elaboration.

### **Elaboration and spreading activation**

Elaboration theories are based on spreading-activation theories of semantic memory (Collins & Loftus, 1975). Put briefly, spreading-activation theories of memory state that information in memory is stored in a network of nodes with relations between them. When a given node (e.g., *red*) is activated, related concepts (e.g., *orange*, *fire*, and *apple*) are activated through spreading activation in the network. According to testing effect theories based on elaboration, testing benefits memory because during testing more elaborative processing takes place than during restudying. Through this elaborative processing concepts that are related to the learned information are activated. These activated concepts are then stored with the

to-be-remembered information and serve as additional retrieval cues at the final test (Carpenter & DeLosh, 2006). As a result, tested items will be better remembered on a final test than restudied items.

An example of a testing effect theory based on elaboration is the elaborative retrieval hypothesis (Carpenter, 2009). According to this hypothesis, during testing an elaborative structure is generated based on the information that is retrieved. For instance, during learning of the word pair *orange* – *shampoo* one might activate the semantically related words *red*, *hair*, and *redhead*. All activated information is stored with the learned information. At the final test, when the cue *orange* is presented the words activated during learning help to retrieve the target *shampoo*.

A related theory is the mediator effectiveness hypothesis (Pyc & Rawson, 2010, 2012). Compared to the elaborative retrieval hypothesis, the mediator effectiveness hypothesis is somewhat more specific about what information is activated during learning. The mediator effectiveness hypothesis states that during retrieval practice, a mediator (i.e., information related to the cue) is activated and coupled with the target to form an additional retrieval cue. The link that is established between mediator and target facilitates retrieval at the final test. For example, suppose someone is learning the Swahili – English word pair *wingu* - *cloud*. For this word pair, *wing* might be a good mediator, because it is easy to recall using the cue *wingu* and it is easy to couple *wing* with the target *cloud*. At the final test, the cue *wingu* might then activate the mediator *wing*, which facilitates retrieval of the target *cloud* because of the link between *wing* and *cloud* that was strengthened during retrieval practice in the learning phase.

The mediator effectiveness hypothesis is also able to explain the fact that the effect of testing is larger when there is more time between learning and the final test (i.e., there is an interaction between learning condition and retention interval). According to the mediator effectiveness hypothesis, tested information is remembered better because of semantically related information that is activated during testing. Because of the semantic organization of long-term memory, semantic information is more likely to be used as a retrieval cue in the long term than perceptual information. Therefore, if tested information is remembered by activating semantically related concepts, it is more likely than restudied information to be remembered in the long term (Carpenter, 2011).

Carpenter (2011) tested the mediator effectiveness hypothesis. She had participants learn related word pairs (e.g., *prescription* – *doctor*) through restudying or testing. At the final test, she used three kinds of cues: the original cue (e.g., *prescription*), a word associated with the target (a ‘related’ word, e.g., *hospital*), or a



word associated with the cue (a mediator, e.g., *drug*). The mediator effectiveness hypothesis would predict that because mediators are more often activated during testing than during restudying, these mediators would be more often falsely recognized after testing than after restudying. In addition, according to the mediator effectiveness hypothesis mediators are coupled with the target during testing and thus the association between the mediator and the target is strengthened. Therefore, after testing a mediator should be a better retrieval cue than a 'related' word that is only weakly associated with the target and was not activated during learning. Indeed, at the final test mediators were more often falsely recognized after testing than after restudying. Moreover, mediators were a more effective final test cue after testing than after restudying and the testing effect was larger when the final test cue was a mediator than when the final test cue was a 'related' word. This suggests that the mediators were indeed more often activated during testing than during restudying, and thus provides support for the mediator effectiveness hypothesis.

In sum, several empirical findings are in line with the mediator effectiveness hypothesis. Consequently it currently is an important account of the testing effect. A crucial assumption of the mediator effectiveness hypothesis is that semantically related information is activated during testing, more than during restudying. In other words, the mediator effectiveness hypothesis assumes that more elaboration takes place during testing than during restudying. However, there is little direct evidence in favor of this assumption. Therefore, to investigate the mediator effectiveness hypothesis of the testing effect it is important to establish whether testing indeed elicits more elaboration than restudying. The current dissertation will provide information on this question.

### Dissertation outline

This dissertation describes a series of four studies. The general aim of the studies was to investigate the role of elaboration in the emergence of the testing effect. In all of the studies, we used cue-target pairs as to-be-learned materials. We used two innovating techniques. In the studies described in Chapters 3 and 4, we used event-related potentials (ERPs), a technique that had not been used before in testing effect research (but see Pastötter, Schicker, Niedernhuber, & Bäuml, 2011, for an oscillatory EEG study on retrieval effects). Through recording ERPs, we were able to investigate cognitive processes during study and retrieval more directly than through behavioral measures such as reaction times and test performance. Furthermore, in the study in Chapter 5 we conducted three experiments of which two were replications of an

earlier study. Subsequently, we performed a small-scale meta-analysis to combine the findings of all three experiments and the original study. The confidence intervals we used in the meta-analysis indicated the limited precision of the findings from the individual experiments. Moreover, in addition to providing an overview of the results, the analysis gave a very precise combined estimate of the effect we found. The aims of each study reported in this dissertation will be described below.

In Chapter 2, we investigated the testing effect in learning symbol-word pairs. Because learning of symbol-word pairs is common in everyday life (e.g., learning musical or mathematical symbols), we examined whether intervening testing could improve learning of symbols. We had participants learn Adinkra symbols coupled with words through testing or restudying and complete a final test after five minutes or seven days. Because the Adinkra symbols allowed no verbal elaboration, this study provided some information about the role of elaboration in the emergence of the testing effect; if we would find a testing effect with these materials, this suggests that elaboration is not necessary in order to obtain a testing effect.

In Chapter 3, we recorded event-related potentials (ERPs) during learning of word pairs such as *orange* – *shampoo* through testing and restudying, in order to investigate the cognitive processes that occur during testing and restudying. We analyzed ERP components that give an indication of the amount of semantic processing that occurs during learning.

We started to investigate the elaboration accounts of the testing effect in more detail in Chapter 4. In this study we tested elaboration, again using ERPs. We had participants restudy or retrieve pairs of words that were each related to a third word. The third word was a homonym (i.e., a word with multiple meanings). For instance, for the homonym *bank* (i.e., a financial institution or a shoreline), the word pair was *teller* – *account*. If testing causes more elaboration than restudying, one would expect that during testing the related homonym *bank* is more often activated than during restudying. Immediately after the word pair learning phase, participants read sentences that contained the homonym in the meaning that was not activated during learning of the word pairs (e.g., *he lay down in the grass on the bank*). Using ERPs, we measured the amplitude of the N400, which indicates the amount of cognitive conflict at the time of reading the word *bank*. If the word *bank* had been activated during learning of the word pairs, it would be in the meaning of a financial institution, thus resulting in a conflict upon reading the word *bank* in the meaning of a shoreline. We hypothesized that the homonym would be more often activated during testing than during restudying. Therefore, we predicted that there would be a stronger conflict

during reading of the homonyms after testing than after restudying and thus a larger N400 amplitude.

In Chapter 5, we again tested the activation of mediators during testing, using an experiment developed by Carpenter (2011). Participants were tested using Mechanical Turk, an online marketplace for work. We had participants learn word pairs through restudying or testing and administered a final test in which the final test cue was either a word associated with the target (related cue) or a word associated with the cue (mediator). In Experiment 1, we set out to investigate an alternative explanation of Carpenter's results (i.e., a larger testing effect when the final test cues were mediators compared to when the final test cues were related cues). Our alternative explanation involved the activation of the original cue. That is, in some of Carpenter's stimuli sets there was a strong association from the mediator to the cue (mediator-cue association). We hypothesized that through this strong association, participants were able to retrieve the original cue when given a mediator during the final test. This would result in a larger testing effect for mediator final test cues, without activation of the mediator during learning. To test our hypothesis, we created two lists of word pairs. In one of the lists, there was no pre-existing mediator-cue association. In the other list, there was a strong mediator-cue association. Based on our alternative explanation of Carpenter's results, we predicted that in the list with strong mediator-cue associations we would find the effect Carpenter found (i.e., a larger testing effect for mediator final test cues than for related final test cues). By contrast, in the list without mediator-cue associations we predicted similar testing effects for mediator and related final test cues. In Experiments 2 and 3 we performed two direct replications of Carpenter (2011). To combine the findings of all three experiments and the original study by Carpenter, we performed a small-scale meta-analysis.

In Chapter 6, the findings are summarized and discussed in relation to existing literature. Furthermore, theoretical implications are discussed and suggestions for future research are provided.



# 2

## Learning Adinkra symbols: The effect of testing

2

This chapter has been published as:

Coppens, L. C., Verkoeijen, P. P. J. L., & Rikers, R. M. J. P. (2011). Learning Adinkra symbols: The effect of testing. *Journal of Cognitive Psychology*, 23, 351–357. doi:10.1080/20445911.2011.507188

### **Abstract**

The testing effect (i.e., long-term memory is improved more by intermediate testing than by restudying the information) has been studied using a variety of materials. However, almost all testing effect studies to date have used purely verbal materials such as word pairs, facts and prose passages. The testing effect has not yet been established using symbol-word pairs. In the present study symbol-word pairs were used as to-be-learned materials to demonstrate the generalizability of the testing effect to symbol learning. The results showed that there was no difference in final memory-test performance after a retention interval of 5 minutes, but after a retention interval of a week tested pairs were retained better than repeatedly studied pairs. Hence, the present research suggests that the testing effect can also be obtained in symbol learning.

When people are tested on the material they have just learned, they show better memory for that information in the long term as compared to a situation in which the same material is restudied (Gates, 1917; for a recent review see Roediger & Karpicke, 2006a). This pattern even holds when the total duration of exposure to the materials is equal in both conditions (Roediger & Karpicke, 2006a). This phenomenon is called the testing effect. The testing effect has been demonstrated using different stimulus materials, such as word pairs (e.g., Carpenter et al., 2008; Pyc & Rawson, 2009), prose (Chan et al., 2006; Roediger & Karpicke, 2006a), and general knowledge facts (e.g., Carpenter et al., 2008).

Almost all testing effect studies have used verbal materials (Pashler, Rohrer, Cepeda, & Carpenter, 2007). However, recently a number of studies have examined the testing effect with other materials. For instance, Butler and Roediger (2007) used videos of lectures about art history as stimulus material. Participants watched three video lectures and after each video engaged in a different type of learning activity: studying a summary of the lecture (restudy), taking a multiple choice test or taking a short answer test. A strong positive effect of taking a short answer test over restudying on long-term (one month) retention was found. Similarly, Johnson and Mayer (2009) used an educational video about lightning formation and they as well found a testing effect.

In addition to these multimedia materials, the testing effect has been demonstrated with visuospatial tests. For instance, Carpenter and Pashler (2007) required participants to study maps with several features (e.g., roads, rivers, a school). Subsequently, participants in the restudy condition studied the same map again. Alternatively, participants in the testing condition were presented the same map with one of the features deleted and they were instructed to covertly retrieve the missing item (intervening test). After 30 minutes, participants in the testing condition performed significantly better than participants in the restudy condition on the final test that consisted of redrawing the map from memory.

A type of nonverbal material that is used relatively often in testing effect studies is face-name pairs. Landauer and Bjork (1978) already used face-name pairs to investigate the effect of repeated testing on memory and found that tested pairs were remembered better than restudied pairs after a retention interval of 30 minutes. Recently, Carpenter and DeLosh (2005) found a testing effect comparing memory for restudied and tested name-face pairs after a retention interval of five minutes.

As described above, some research on the testing effect has been done using not purely verbal materials. However, research on a specific type of non-verbal materials, namely symbol-word pairs, has not been described in the published literature (but see

Kang, 2010). In our view, this dearth of symbol-word pair testing studies represents an important caveat in the literature because learning of symbol-word pairs is common in educational settings. For example, a pupil taking piano lessons has to learn how to read notes from a score and a student of mathematics has to learn the meaning of mathematical and logic symbols. Hence, the purpose of the present study is to examine whether the testing effect generalizes to symbol learning. If our study, in which symbol-word pairs were used as to-be-learned materials, would demonstrate the same mnemonic testing benefit that is generally observed with purely verbal materials, then this could serve to improve educational practice.

## Method

### Participants

Fifty students (17-30 years old, mean age 20.3 years, 29 females) of the Erasmus University Rotterdam participated in partial fulfillment of a course requirement. None of the participants were familiar with the symbols that were used in the present study.

### Materials

Forty Adinkra symbols (MacDonald, 2009) paired with 40 concrete five-letter words (see Appendix) were used. These symbols were developed by the Ghanaian Asante tribe and can be found everywhere in Ghana: On cloth, pottery, walls, etc. Complex, abstract symbols were selected to ensure participants could not verbally label the symbols. In a normative study, the symbols were rated on verbal labeling difficulty by 15 undergraduate students who did not participate in the main experiment. The procedure of this normative study was based on the procedure of Vanderplas and Garvin (1959). Twenty pictures of objects and 20 pictures of simple shapes in addition to the 40 Adinkra symbols were presented during eight seconds per picture. Participants were asked to 'try to describe the pictures as best as you can'. After participants typed their description of the picture, they were asked to indicate on a scale ranging from 1 to 9 (1 meaning 'very easy' and 9 meaning 'very difficult') how difficult it was for them to name the picture. Mean naming difficulty (*SD*) was 1.48 (0.48) for pictures of objects, 2.05 (0.57) for simple shapes and 6.84 (0.94) for Adinkra symbols. A repeated measures ANOVA yielded a significant effect of Picture Type,  $F(2,28) = 283.03$ ,  $MSE = 0.46$ ,  $p < .001$ ,  $partial \eta^2 = .953$ . Adinkra symbols were judged to be more difficult to recode verbally compared to simple shapes,  $t(14) = 16.59$ ,  $p < .001$ ,  $d = 4.29$ , and compared to pictures of objects,  $t(14) = 19.43$ ,  $p < .001$ ,  $d = 5.02$ .



Words were selected from the CELEX database (Baayen, Piepenbrock, & Van Rijn, 1993). The number of words starting with the same letter was limited to three and words starting with the same letter had a different second letter to reduce confusion. Four additional pairs were used as practice pairs.

Symbols and words were paired randomly. There was no connection between the word paired with a symbol and the actual meaning of the symbol. The pairing was the same for all participants. Two lists of 20 randomly selected pairs were created. To control for item difficulty, two versions of the experiment were created, so that a given list of 20 pairs was learned through restudy trials by half the participants and through test trials by the other half of the participants. Tested and restudied lists were not mixed; all items in one condition (study or test) were presented in sequence. Lists were 'pure' to prevent selective displaced rehearsal of tested items during study trials (Slamecka & Katsaiti, 1988). To control for order effects, half of the participants started with the restudy trials and half started with the test trials.

### Design

A mixed design was used. All participants studied all 40 symbol-word pairs once, studied one list of 20 pairs three additional times (SSSS: restudy) and were tested by means of a cued-recall test on the other list of 20 pairs three times (STTT: test). A final cued-recall test on all 40 symbol-word pairs was administered after 5 minutes or after 7 days. Trial type (restudy vs. test) was varied within participants, whereas retention interval (five minutes vs. a week) was varied between participants.

### Procedure

Participants were informed that they would be presented with a series of symbol-word pairs, that they had to type the word appearing in each symbol-word pair, and that they had to remember as many pairs as possible.

In a study trial the symbol-word pair was shown in the center of the computer screen at a presentation rate of 8s per pair, with the word always displayed above the symbol (this was done to ensure that participants would adequately process the word; see Carpenter & DeLosh, 2005, for a similar procedure). Furthermore, during a study trial, participants typed the word presented with a symbol. In a test trial the symbol was presented without the word and participants had 8 seconds (this corresponded to the presentation rate of the study trials) to type the word paired with the symbol. During both study and test trials, typed words appeared directly below the symbol on the screen. After participants had entered their response, no feedback was provided.

Participants first studied a subset of four pairs from either the restudy list or the tested list. Then they restudied these pairs twice or were tested on these pairs twice.

During testing or restudying, the four items of the subset appeared in a new random order. This procedure was repeated until all 20 pairs of a list had been restudied or tested. When all items in one condition (restudy or intermediate testing) had been processed, participants restudied or were tested one additional time on all 20 symbol-pairs of the first list. Following the first list, the pairs of the second list were presented either according to the testing or the restudy procedure. Thus, depending on the condition, each pair was studied four times, or it was studied once and tested thrice. This procedure was used to maximize learning and to prevent item selection effects (cf. Toppino & Cohen, 2009).

After the restudy and testing trials, half of the participants performed a filler task, a self-efficacy questionnaire, for five minutes (short retention interval), after which the final cued recall test began. The other half of the participants also completed the self-efficacy questionnaire immediately after the study phase, but returned after seven days (long retention interval) for the delayed final test. In the final cued recall test, symbols were presented in a random order one at a time without the word and participants typed in the word, as they did in an initial test trial during the learning phase. The final test was self-paced, and no feedback was provided.

At the end of the experiment, participants answered a number of open-ended questions on the employed study strategy ("how did you try to remember the symbols?") and prior knowledge of the symbols ("did you know any of the symbols presented in this experiment?"). To motivate participants, there was a lottery among the 50% of participants in each retention interval condition who remembered most pairs at the final test (one prize of 25 Euro for each condition). Participants were informed about this lottery before starting the learning phase of the experiment. Because the stimuli were nonsense pairs, there was no risk of participants studying at home to improve their performance on the final test.

## Results

### Learning phase

An alpha level of .05 was used for all statistical tests reported in this article. Subject responses were checked for typing errors. Minor errors, in which one letter was incorrect or missing, were corrected.

Mean accuracy and median response latency (first key press) on the three immediate tests in the learning phase was computed for each trial type (restudy or test) and each practice trial (see Tables 1 and 2). A 3 (Trial: first/second/third)  $\times$  2 (Trial Type: Study/Test) repeated measures ANOVA on the accuracy data yielded a

significant main effect of Trial Type,  $F(1,49) = 52.44$ ,  $MSE = 0.021$ ,  $p < .001$ ,  $partial \eta^2 = .517$ , a significant main effect of Trial,  $F(2,98) = 37.68$ ,  $MSE = 0.004$ ,  $p < .001$ ,  $partial \eta^2 = .435$ , and a significant Trial Type \* Trial interaction,  $F(2,98) = 39.77$ ,  $MSE = 0.004$ ,  $p < .001$ ,  $partial \eta^2 = .448$ . Similarly, a  $3 \text{ (Trial: first/second/third)} \times 2 \text{ (Trial Type: Study/Test)}$  repeated measures ANOVA on the response latency data yielded a significant main effect of Trial Type,  $F(1,49) = 25.55$ ,  $MSE = 110063$ ,  $p < .001$ ,  $partial \eta^2 = .343$ , a significant main effect of Trial,  $F(2,98) = 64.07$ ,  $MSE = 80978$ ,  $p < .001$ ,  $partial \eta^2 = .567$ , and a significant Trial Type \* Trial interaction,  $F(2,98) = 12.69$ ,  $MSE = 71464$ ,  $p < .001$ ,  $partial \eta^2 = .206$ . The results of pairwise comparisons are shown in Tables 1 and 2.

Table 1

*Mean accuracy scores (SD) on each practice trial in the learning phase*

Practice trial	Trial type		Restudy-test comparison
	Restudy	Test	
First	0.995 (0.290)	0.918 (0.093)	$t(49)=5.47$ , $p<.001$ , $d=0.77$
Second	0.998 (0.010)	0.929 (0.092)	$t(49)=5.24$ , $p<.001$ , $d=0.74$
Third	0.999 (0.007)	0.783 (0.200)	$t(49)=7.61$ , $p<.001$ , $d=1.08$

Table 2

*Mean median response latency (SD) in ms on each practice trial in the learning phase*

Practice trial	Trial type		Restudy-test comparison
	Restudy	Test	
First	1192 (327)	1305 (249)	$t(49)=-2.657$ , $p=.011$ , $d=0.38$
Second	1122 (253)	1178 (186)	$t(49)=-2.189$ , $p=.033$ , $d=0.31$
Third	1379 (508)	1790 (573)	$t(49)=-4.703$ , $p<.001$ , $d=0.67$

### Final test

The proportion of correctly entered words on the final test was computed for each trial type and retention interval. Accuracy scores for the different conditions are given in Table 3. A  $2 \text{ (Trial Type: Restudy/Test)} \times 2 \text{ (Retention Interval: short/long)}$  mixed measures ANOVA yielded a significant main effect of Retention Interval,  $F(1,48) = 112.11$ ,  $MSE = 0.057$ ,  $p < .001$ ,  $partial \eta^2 = .700$ , no significant main effect of Trial Type,  $F(1,48) = 2.022$ ,  $MSE = 0.034$ ,  $p = .162$ ,  $partial \eta^2 = .040$ , and a significant Trial Type \* Retention Interval interaction,  $F(1,48) = 12.230$ ,  $MSE = 0.207$ ,  $p = .001$ ,

*partial*  $\eta^2 = .203$ . After a retention interval of five minutes, restudied items were remembered as well as initially tested items,  $t(49) = 1.391$ ,  $p = .177$ ,  $d = 0.278$ . However, after a retention interval of seven days, initially tested items were recalled significantly better than restudied items,  $t(49) = 3.694$ ,  $p = .001$ ,  $d = 0.723$ .

For completeness, the data were checked for recency and primacy effects. There was no relation between the position of the symbol-word pair in the initial test and restudy trials of the study phase and the proportion of correct responses on the final test.

Table 3

*Proportion correct (SD) on final test for each trial type and retention interval*

Retention interval	Trial type	
	Restudy	Test
Short (5 minutes)	0.834 (0.178)	0.780 (0.232)
Long (7 days)	0.236 (0.187)	0.364 (0.167)

## Strategy

At the end of the learning phase, participants were asked to indicate how they had tried to remember the symbols. Forty-three participants (86%) reported they had tried to find a connection between the symbol and the word, usually by thinking of a story or a situation to connect the two. Four participants (8%) reported they had simply repeated the word and looked at the symbol during the eight seconds of study time. The remaining three participants (6%) could not describe their learning strategy. Strategy did not interact with response latency or number of pairs correctly recalled, neither during the learning phase nor during the final test.

## Extra test

Of the 25 participants who took the final test after a short retention interval, 20 participants took an additional test after a retention interval of seven days. The remaining five participants in the short retention interval condition did not return after seven days for practical reasons, for example illness of the participant. Test scores for the 20 participants who took both tests are given in Table 4.

Table 4

*Proportion Correct (SD) at Immediate and Delayed Test for Participants in the Short Retention Interval Condition*

Test Time	Trial type	
	Restudy	Test
5 minutes	0.825 (0.187)	0.803 (0.239)
7 days	0.480 (0.266)	0.565 (0.284)

By comparing performance on the test after a week between participants who did take an immediate final test (the 20 participants in the short retention interval condition who took an extra test after seven days; these scores are in the bottom row of Table 4) and those who only took a delayed final test (all 25 participants in the long retention interval condition; these scores are in the bottom row of Table 3), we were able to examine the effect of one test directly after the learning phase on long term retention of both restudied and initially tested items. To this end, a 2 (Initial Test/No Initial Test)  $\times$  2 (Trial Type: Restudy/Test) mixed measures ANOVA was performed. It yielded a significant main effect of Initial Test,  $F(1,43) = 13.507$ ,  $MSE = 0.081$ ,  $p = .001$ ,  $partial \eta^2 = .239$ , a significant main effect of Trial Type,  $F(1,43) = 12.172$ ,  $MSE = 0.21$ ,  $p = .001$ ,  $partial \eta^2 = .221$ , and no significant Initial Test  $\times$  Trial Type interaction effect,  $F < 1$ ,  $p = .485$ ,  $partial \eta^2 = .011$ . Thus, the immediate final test that the participants in the short retention interval condition took improved long term memory for the symbol pairs compared to no immediate final test, regardless of Trial Type.

## Discussion

The goal of the present study was to determine whether the testing effect, which has been established using words, facts and prose as to-be-learned materials, is also present in learning symbol meanings. Our results demonstrated that there was no difference between studied and initially tested symbols after a short retention interval of five minutes, but initially tested symbols were remembered significantly better than restudied symbols after a long retention interval of seven days. Thus, a clear testing effect was demonstrated. It should be noted that these results cannot be attributed to an item selection effect (see Toppino & Cohen, 2009 for a discussion of this topic), because performance after five minutes (immediate recall) did not differ between studied and tested symbols.

The results of the extra test are also worth noting. Twenty participants who took an immediate final test of all the symbols directly after the learning phase were tested

again after seven days. Long-term memory performance of these participants was significantly better than that of participants in the long retention interval condition, who only received a delayed final test after seven days. Interestingly, performance on initially tested symbols improved to the same degree as performance on restudied symbols. According to previous studies (e.g., Bangert-Drowns et al., 1991), performance improvement is a negatively accelerated function of the number of tests. This implies that the improvement of performance from the first test (in the restudy condition) should be greater than the improvement from the fourth test (in the testing condition). However, the present results suggest that an additional final test after three initial test trials improved memory as much as a final test after three restudy trials.










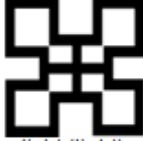












There could be three reasons for this difference in effects of testing. First, the data that Bangert-Drowns et al. (1991) used were collected in the classroom and concerned tests of different material in each test; to-be-learned information was not repeatedly tested. Instead, each test examined memory for different information. In the present study, information was repeatedly tested. This difference in experimental design could explain the difference between the results of the experiments. Second, the average lag between the tests of one item in the learning phase of the present study was shorter than the average lag between the last presentation of an item in the learning phase and the immediate final test of that item. Although Bangert-Drowns et al. did not give information about the spacing of repeated tests, it seems that increasing the lag before the last test could enlarge the effect of this test. Third, the final test that participants took in the present experiment comprised all of the symbols participants had learned, both through restudy and through testing. Therefore, the immediate final test may have served as an overview of the to-be-learned symbols. Participants that did not take an immediate final test did not receive this overview, because tested and restudied stimuli were presented in separate parts of the learning phase and never mixed during learning. Thus, it could be that the overview of all stimuli that participants receive during the immediate final test is the critical factor in improving memory in this situation, instead of the extra test trial. Future research could clarify this question.

As the present study was a laboratory experiment, we can not be sure whether the results copy to the classroom. However, because the testing effect generalizes to classroom applications with verbal materials (e.g., Bangert-Drowns et al., 1991; McDaniel, Anderson, Derbish, & Morrisette, 2007), it seems likely that in the classroom, testing enhances learning of symbols as well. Therefore, it may be good

advice to for instance music and math teachers to regularly test their students on knowledge of the to-be-learned symbols.

In conclusion, similar to learning other materials, symbol-word pair learning seems to be greatly improved by incorporating tests into the study regime. Providing an overview test at the end of the learning phase may improve learning even further. These results are promising for educators teaching disciplines in which symbols have to be memorized; testing students appears to be a powerful way to improve symbol learning.

Appendix: Used words and symbols

				
boete (fine)	draak (dragon)	water (water)	kleed (rug)	tafel (table)
				
fruit (fruit)	nagel (nail)	regen (rain)	asbak	bezem (broom)
				
forel (trout)	toren (tower)	chips (chips)	veter (shoestring)	gebit (teeth)
				
licht (light)	kraan (faucet)	emmer (bucket)	azijn (vinegar)	nylon (nylon)
				
kogel (bullet)	garen (yarn)	zomer (summer)	oever (bank)	ijzer (iron)
				
einde (end)	jacht (hunt)	pizza (pizza)	peper (pepper)	haven (harbor)
				
dwerg (dwarf)	graan (cereal)	vloer (floor)	plein (plaza)	radio (radio)
				
maand (month)	orgel (organ)	hotel (hotel)	snoep (candy)	steen (stone)



# 3

## The neurophysiology of retrieval practice versus restudy: ERP correlates of the testing effect

3

This chapter is in preparation for publication as:

Coppens, L. C., Verkoeijen, P. P. J. L., Van Strien, J. W., & Rikers, R. M. J. P. (in preparation). No evidence for the semantic mediator hypothesis: The testing effect in cued recall is similar for mediator cues and related cues.

### **Abstract**

The testing effect, the finding that retrieved items are remembered better in the long run than repeatedly studied items, is a well-documented effect. However, the neurophysiological underpinnings of the effect are less well-studied. In the present study, we used event-related potentials based on 28 participants to investigate processes during restudy and retrieval trials in a standard testing effect paradigm. The behavioral data showed a clear testing effect at the final test after two days. In regard to the event-related potentials, we found a larger P300 amplitude during retrieval practice trials than during restudy trials, which suggests that retrieval trials required more effort than restudy trials. Furthermore, we found N400 and late positive component repetition effects when comparing first study trials to subsequent restudy trials, indicating recognition of previously presented items. When retrieval practice trials were compared with first study trials, a larger N400 repetition effect was present and no late positive component repetition effect was found. Taken together, these results suggest that retrieval effort and semantic retrieval play a role in the emergence of the testing effect.

When learners study information and afterwards have to retrieve the information, the act of retrieving has an impact on memory. Specifically, long-term retention is better for retrieved information than for restudied information. This phenomenon is referred to as the testing effect, and it has generated interest lately as a useful tool for learning (Roediger & Karpicke, 2006b). The advantage of retrieval practice over restudying has been established with different learners, using various kinds of materials, both in the laboratory and in the classroom (e.g., Carpenter & Pashler, 2007; Glass, Brill, & Ingate, 2008; Johnson & Mayer, 2009; for a review, see Roediger & Butler, 2011). The testing effect can therefore be considered a robust empirical phenomenon. However, the underlying mechanisms that contribute to the testing effect are less clear. To foreshadow, in the current paper we will investigate the frequently proposed retrieval effort mechanism as well as the (related) elaborative retrieval mechanism in a standard cue-target testing effect paradigm using event-related potentials (ERPs).

Recently, many researchers have proposed that the testing effect can be explained in terms of Bjork's (1994) desirable difficulties framework. The central tenet in this framework is that difficult and successful processing of stimulus materials will slow down initial learning, but improve long-term memory. Eventually, difficult and successful learning will benefit memory more than easy and successful processing. Reasoning from the desirable difficulties approach, retrieval after an initial study episode will require more processing effort from a learner than restudying the same material. As a result, retrieval (testing) will lead to slower acquisition than restudying. In the long run, however, it will be more effective, in the sense that the memory traces formed through retrieval will be less susceptible to decay than those formed through restudying. This account of the testing effect is dubbed the retrieval effort hypothesis (Toppino & Cohen, 2009). Consistent with this hypothesis, testing studies that include both a short (i.e., minutes) and a long retention interval (i.e., hours or days) often show a restudy advantage on the short term final test and an advantage of retrieval practice on the long term final test (e.g., Toppino & Cohen, 2009; Wheeler, Ewers, & Buonanno, 2003).

The results of several studies (e.g., Carpenter & DeLosh, 2006; Glover, 1989; Karpicke & Roediger, 2007) are in line with the retrieval effort hypothesis. For instance, Carpenter (2009, Experiment 3) had participants learn words through retrieval practice under different levels of 'cue support': During retrieval practice in the learning phase, varying numbers of letters were provided as a retrieval cue. Carpenter assumed that the more letters are provided, the less effort is needed to retrieve a particular word. If the retrieval effort hypothesis is correct, the *final test* performance

should be better for words retrieved on the basis of a weak cue (i.e., few letters) than for words retrieved on the basis of a strong cue (i.e., more letters). It turned out that the final test results were entirely in line with this prediction, hence providing support for the retrieval effort hypothesis.

Direct empirical evidence of the influence of retrieval effort on memory performance was provided by Pyc and Rawson (2009). They had participants learn Swahili-English word pairs through one study trial and a number of test-restudy trials. Pyc and Rawson manipulated retrieval difficulty by varying the time between subsequent retrieval practice trials (interstimulus interval; ISI) and requiring a varying number of successful retrievals (criterion) per item. The authors predicted that performance on the final cued recall test should be higher when items were learned under the supposedly difficult condition of a long ISI. In regard to the criterion level, their hypothesis was that each successive successful retrieval would be less difficult than the previous one, so the benefit of an additional retrieval trial would diminish as the criterion increased. Indeed, in both experiments these predictions were confirmed: performance was higher for items learned with a long ISI compared to a short ISI and the advantage of an additional retrieval practice trial became smaller as the criterion increased. Additionally, in Experiment 2 response latencies suggested that the long ISI condition was more difficult (longer response latencies) than the short ISI condition and that each successive retrieval was less difficult than the last one. This suggests that the retrieval trials that benefited final test memory most were the most effortful trials.

So, previous studies have provided support for the retrieval effort hypothesis by demonstrating that difficult successful retrieval is more beneficial to memory performance than easy successful retrieval. These studies focused on retrieval effort differences within retrieval practice trials. However, it is reasonable to assume that retrieval effort differences also exist between restudy and retrieval practice trials. Studies on the spacing effect (i.e., long-term retention is better for repeatedly studied items that are studied with a delay or presentation of other items between repetitions, compared to 'massed' study without an interval or other items between successive presentations; for a review, see Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008) have shown that during a repetition of a stimulus (i.e., restudy), study phase retrieval takes place (i.e., the implicit retrieval of a previous presentation of a stimulus during a repetition of the stimulus; Greene, 1989; Hintzman, Summers, & Block, 1975). Testing experiments involve repetitions of to be learned items that are essentially spaced repetitions because other items are presented between repetitions. Therefore, we can assume that during testing experiments study phase retrieval takes place.

Consequently, retrieval takes place not only during retrieval trials but also during restudy trials. However, a retrieval trial involves explicit retrieval on the basis of a cue with only a part of the information (i.e., only the cue without the target for cued recall), whereas a restudy trial involves implicit study phase retrieval with the entire cue-target pair provided. Therefore, retrieval effort should be higher during a retrieval trial than during a restudy trial. In the current experiment, we tested this hitherto untested prediction using ERPs.

Our ERP analysis focused on the P300 component, which is a positive going component that can be found between 250-500 ms after stimulus onset, depending on task and participant characteristics (Polich, 2007). P300 is associated with attention and memory processes. More specifically, the P300 seems to stem from frontal inhibitory activity that focuses attention on the task at hand and subsequent temporal and parietal activity that promotes memory operations (Polich & Criado, 2006; Polich, 2007). Because we expected participants to put more effort into retrieval trials than restudy trials, we expected a larger P300 amplitude during retrieval trials than during restudy trials. A larger P300 amplitude during retrieval trials would indicate more attention being allocated to the processing of these trials, which would be in accordance with the retrieval effort hypothesis.

To further investigate the processes that underlie the supposedly increased effort during retrieval, we looked at processing of the restudy and retrieval trials. As discussed above, both during retrieval and during restudying study-phase retrieval takes place: a previous occurrence of the item is retrieved. Based on previous findings in the literature discussed below, we predicted that during retrieval practice, study-phase retrieval focuses more on semantic properties of the word than during restudying.

Most theories of the testing effect imply that the processing that distinguishes retrieval practice from restudying is focused on the meaning of the word instead of on (contextual) details such as font or background color. For instance, a class of explanations is based on elaborative processing during retrieval, such as the elaborative retrieval hypothesis (Carpenter, 2009). According to this hypothesis, retrieval practice causes activation of concepts in memory semantically related to the learned materials. The activated information is then coupled with the to be remembered information and aids later recall. During restudying, this activation of related concepts is less elaborate. For instance, when learning the word pair *orange – shampoo* through testing one might think of the word *hair* and perhaps the word *redhead*. At the final test, when given the cue *orange* the target *shampoo* easily comes to mind through re-activation of *hair* and *redhead*. To activate semantically related

information, the meaning of a word has to be activated. So, study-phase retrieval during retrieval practice trials should be more focused on the meaning of the word than on the contextual details of the previous presentations. Indeed, the results of several studies show that more semantically related information is activated during testing than during restudying (e.g., Carpenter, 2009; Carpenter & DeLosh, 2006; Pyc & Rawson, 2010; Verkoeijen, Bouwmeester, & Camp, 2012).

Bouwmeester and Verkoeijen (2011) provided direct evidence for the hypothesis that during retrieval, processing is more semantic or meaning-directed. The authors used Deese-Roediger-McDermott (DRM)-lists (lists of words that are related to a central, not-presented lure; e.g., *rest, bed, nap, peace, drowsy, blanket, doze, tired, awake, snooze, yawn, slumber, snore, wake, and dream* are all related to the lure *sleep*) to investigate the testing effect in children. Children learned the words in the DRM-lists through restudy or retrieval practice and subsequently completed a recognition test including studied words, the lures and unrelated distractors. The results showed that the lures of retrieved lists were more often falsely recognized than the lures of restudied lists. This finding suggests that the children activated the central theme of the DRM-list more often during retrieval than during restudying. Because the theme of DRM-lists can only be activated when the meaning of the words is processed, this provides support for our hypothesis that semantic processing is stronger during retrieval practice than during restudying.

To test our prediction that during retrieval practice semantic processing will be stronger than during restudy trials, we looked at familiarity and recollection. These concepts are part of the repetition effect. Repetition effects have been extensively studied in the ERP literature (e.g., Curran, Tepe, & Piatt, 2006; Curran, 1999, Van Strien, Verkoeijen, Van der Meer, & Franken, 2007). Studies on the ERP repetition effect often use continuous recognition paradigms, in which participants see a continuous stream of stimuli and have to indicate whether the stimulus was presented before or not (old/new task). The ERP differences found when participants compare correctly recognized old versus correctly rejected new stimuli typically indicate the more positive going of two components: a frontal early negative component (300-500 ms, often described as an attenuation of the N400 component; see Rugg, 1995) which we will label the N400 repetition effect, and a parietal late positive component (500-800 ms, interpreted as an enhanced LPC, P600 or late P3), which we will label the LPC repetition effect. The N400 repetition effect is thought to indicate recognition based on familiarity, i.e., knowing that the stimulus occurred earlier, but not remembering the specific occurrence. Familiarity is a 'feeling of knowing' based on perceptual or conceptual fluency (Jacoby, 1991). The LPC repetition effect is thought to indicate

recognition based on recollection, i.e., remembering specific details (such as physical attributes) or the context of the previous presentation of the 'old' stimulus (Curran, 2000; Mecklinger, 2000; Rugg, Schloerscheidt, & Mark, 1998; Schnyer, 1997; Voss & Paller, 2009).

In the current study, participants studied and repeated word pairs in retrieval practice or restudy trials. Therefore, the paradigm differs somewhat from the previously mentioned continuous recognition paradigm (participants did not indicate whether the presented item is old or new). However, because repetition effects are also found for repeated stimuli tasks other than old/new tasks, such as lexical decision (e.g., Curran, 1999) we expected to see the standard N400 and LPC repetition effects during restudy and retrieval practice trials. In addition, as described above we hypothesized that retrieval practice is focused on meaning/semantics instead of on the physical or contextual details of the words. Therefore, we expected stronger familiarity processing and less recollection processing during retrieval than during restudying. Thus, we expected a larger N400 repetition effect and a smaller LPC repetition effect during retrieval vs. during restudying.

Thus, the goals of the present experiment were to test the assumption that retrieval is more effortful than restudying, and to investigate the processes underlying the increased effort during retrieval trials. We expected ERPs recorded during retrieval of word pairs to show an increased P300 amplitude compared to ERPs recorded during restudying of word pairs, which would indicate increased effort during retrieval. Furthermore, we expected LPC repetition effects during restudy trials and during retrieval trials and a larger N400 repetition effect during retrieval trials than during restudy trials. This would indicate stronger familiarity processing during restudy trials, which would fit with the elaborative retrieval hypothesis.

## Method

### Participants

Forty undergraduate Psychology students participated in partial fulfillment of a course requirement. Data from twelve participants were excluded from analysis because too little segments remained after artifact rejection (see below), leaving 28 participants (23 females and 5 males, mean age 20.9, age range 18-36). All were native Dutch speakers with normal or corrected-to-normal vision and were able to touch type.

### Ethics statement

Prior to the experiment, all participants gave written informed consent. The study was approved by the Ethical Committee Psychology of the Erasmus University Rotterdam.

### Stimuli

Eighty Dutch 6- or 7-letter words from the CELEX database (Baayen et al., 1993) with an average Institute for Dutch Lexicology (INL) frequency of 31 per million were used to form 40 unrelated word pairs (e.g., filter – island). Two additional words were used as a practice pair. To prevent confounds caused by different word frequencies, item difficulty and the order of items and conditions, four different counterbalancing sequences were created (see Table 1).

Table 1

*Block sequence in the 4 counterbalance lists.*

Block	Activity in CB list 1	Activity in CB list 2	Activity in CB list 3	Activity in CB list 4
1	Study items 1-20	Study items 1-20	Study items 21-40	Study items 21-40
2	Restudy items 1-10 Retrieve items 11-20	Retrieve items 1-10 Restudy items 11-20	Restudy items 21-30 Retrieve items 31-40	Retrieve items 21-30 Restudy items 31-40
3	Restudy items 1-10 Retrieve items 11-20	Retrieve items 1-10 Restudy items 11-20	Restudy items 21-30 Retrieve items 31-40	Retrieve items 21-30 Restudy items 31-40
4	Study items 21-40	Study items 21-40	Study items 1-20	Study items 1-20
5	Restudy items 21-30 Retrieve items 31-40	Retrieve items 21-30 Restudy items 31-40	Restudy items 1-10 Retrieve items 11-20	Retrieve items 1-10 Restudy items 11-20
6	Restudy items 21-30 Retrieve items 31-40	Retrieve items 21-30 Restudy items 31-40	Restudy items 1-10 Retrieve items 11-20	Retrieve items 1-10 Restudy items 11-20
7	Restudy/retrieve (according to condition) items 1-40	Restudy/retrieve (according to condition) items 1-40	Restudy/retrieve (according to condition) items 1-40	Restudy/retrieve (according to condition) items 1-40

*Note.* Item order was random within blocks. CB = counterbalancing.

### Procedure

The experiment consisted of two phases: a learning phase and a final test two days later. EEG was recorded only during the learning phase.

At the beginning of the learning phase, participants were informed that they would study word pairs and that they should try to remember as many of these pairs as



possible for an unspecified test. After this instruction, two practice trials were presented (one study and one retrieval trial), after which the learning phase of the experiment started. Word pairs were presented in black lower case font (Arial, 30 points) on a light gray background, on a 17-inch computer screen (resolution 1280 x 1024 pixels) placed 1.25 m from the participant. Presentation rate was 6 s per pair, plus 1 s feedback and 1 s intertrial interval, both for (re)study and retrieval trials. During feedback, the word 'correct' or 'incorrect' was displayed together with the correct response for 1 s. In Figure 1, the sequence for each trial type is shown.

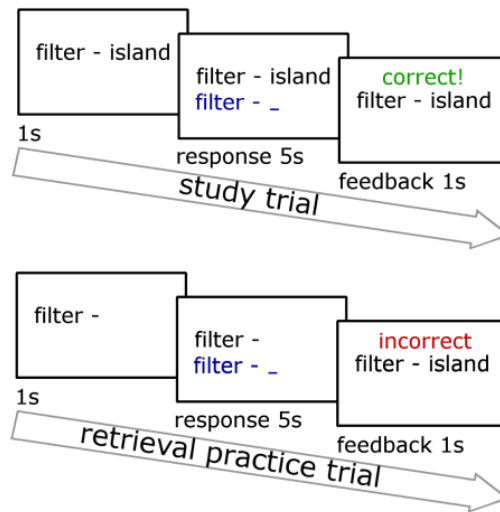


Figure 1. Trial sequence for study (first study and restudy) and retrieval practice trials.

In a (re)study trial, both the cue word and the target word were presented and after 1 s, participants were required to type the target word. This typing task was inserted to ensure participants would accurately process the word pairs, and to be able to assess learning and retrieval during the learning phase (cf. Coppens, Verkoeijen, & Rikers, 2011). To diminish eye movements that could cause EEG artifacts, we used a blank keyboard placed on the participants' lap. As participants were all able to touch type, they responded with both hands and did not look at the keyboard during the experiment. In a retrieval trial, only the cue word was given and participants typed the target word as they did in a study trial. The participants had to type during both conditions in order to avoid confounds due to typing. Feedback was provided after each trial, because a pilot study indicated that without feedback, memory for retrieved items was poor. All word pairs were learned through one study

trial (first study, 40 trials) and either three additional study trials (restudy, 3x20 trials) or three retrieval trials (retrieval, 3x20 trials). Trials were presented in blocks of 20 items. In Table 1, the block sequences are shown. One block of 20 first study trials was followed by two blocks of 20 restudy/retrieval trials. Within each restudy/retrieval block, one restudy or retrieval trial of each of the items in the block was presented, depending on the condition. Items were presented in a random order within blocks. When all 40 items were learned through one study trial and two restudy or retrieval trials in 20-item blocks, all items were presented for one additional restudy or retrieval trial (according to condition) in a block containing all 40 items.

Two days after the learning phase participants returned to take the final test. In this test, all cue words were presented one by one in a random order and participants were instructed to type the target word in response to the cue word. The final test was self-paced.

### **EEG Recording**

Participants sat in a comfortable chair in a dimly lit room, separate from the experimenter. EEG was recorded from 32 active Ag/AgCl electrodes (BioSemi, Amsterdam, the Netherlands) placed in an elastic cap according to the international 10/20 system, positioned at Fz, Cz, Pz, Oz, FP1/2, AF3/4, F3/4, F7/8, FC1/2, FC5/6, C3/4, T7/8, CP1/2, Cp5/6, P3/4, P7/8, PO3/4, and O1/2. Recordings were amplified using an ActiveTwo amplifier system and sampled at 2048 Hz. An additional active electrode (CMS, common mode sense) and a passive electrode (DRL, driven right leg) were used as a feedback loop for amplifier reference. An electrode was placed on each mastoid, and eye movements and blinks were monitored by four additional electrodes.

### **EEG Analysis**

Data were processed and analyzed with Brain Vision Analyzer software (Brain Products, Gilching, Germany). Signals were re-referenced offline to the averaged mastoids. Eye-movement artifacts were removed using the algorithm of Gratton, Coles, and Donchin (1983).

Data were filtered with a band-pass filter of 0.01 to 40 Hz, segmented into epochs of 1200 ms (-200 – 1000 ms post stimulus onset) and baseline-corrected relative to the 200 ms before stimulus onset. Epochs including a signal exceeding  $\pm 100 \mu\text{V}$  were excluded from further analysis. Only trials with a correct response were analyzed. Participants with less than 25 (of 40) valid (i.e., after a correct response and without artifacts) first study segments or less than 40 (of 60) valid restudy or retrieval segments were excluded from analysis, leaving 28 participants. The mean number of valid epochs per remaining participant was 36.7 for study trials, 55.6 for restudy trials

and 46.6 for retrieval trials. Mean ERPs were calculated by averaging trials for each participant, electrode and trial type separately.

The ERP analysis focused on the N400, LPC, and P300 components. In previous studies, N400 repetition effects have been found in the 300-500 ms time window, most pronounced at frontal sites, whereas LPC repetition effects are found in the 500-800 ms time window, mainly at parietal electrode sites (e.g., Curran & Cleary, 2003; Curran, 2000). To avoid overlapping time intervals, we used mean ERP amplitude in the time window from 250 to 350 ms after stimulus onset as a measure of P300, and 350 to 450 ms after stimulus onset as a measure of N400. For the LPC repetition effect, a time window from 500 to 800 ms after stimulus onset was used.

### Statistical Analyses

Analyses of variance (ANOVAs) were conducted for each component, with trial type (first study, restudy, retrieval) and electrode (Fz, Cz, Pz) as within-subjects variables. Because sphericity could not be assumed, we report the multivariate tests. An alpha level of .05 was used for all statistical tests reported in this paper.

## Results

Minor typing errors in participants' responses, in which one letter was added, missing or in the wrong place, were corrected before analysis.

### Behavioral Data

Mean accuracy (proportion of targets correctly reported) during the learning phase was .999 for first study trials, 1.000 for restudy trials and .823 for retrieval trials.

On the final test two days after the learning phase, mean accuracy for restudied items was .532 ( $SD = .229$ ), whereas accuracy for retrieved items was .846 ( $SD = .142$ ). Retrieved items were remembered better than repeatedly studied items,  $t(27)=8.800$ ,  $p<.001$ ,  $d=1.663$ ; a testing effect occurred.

### ERPs

Grand average waveforms at selected midline electrodes Fz, Cz and Pz are shown in Figure 2. This figure shows that the P300 amplitude was larger for restudy trials than for first study trials, and larger for retrieval trials than for restudy trials, especially at electrodes Cz and Pz. Compared to first study trials, the restudy trials elicited a smaller N400 amplitude that was followed by a late positivity (LPC), especially at Cz and Pz. Furthermore, the retrieval trials show a smaller N400 amplitude than the first study trials, but no difference in LPC amplitude.

**P300.** Scalp distributions of the restudy-first study and retrieval-first study difference waves in the 250-350 ms time window are shown in Figure 3. There was a significant effect of trial type,  $F(2,26)=24.177$ ,  $p<.001$ ,  $\eta^2_p=.650$  and a Trial Type  $\times$  Electrode interaction effect,  $F(4,24)=3.452$ ,  $p=.023$ ,  $\eta^2_p=.365$ . Subsequent Bonferroni-corrected pairwise comparisons indicated that P300 amplitude was larger during restudy trials than during first study trials,  $F(1,27)=25.389$ ,  $p<.001$ ,  $\eta^2_p=.485$ , and larger during retrieval practice trials than during restudy trials,  $F(1,27)=14.037$ ,  $p=.003$ ,  $\eta^2_p=.342$ . Consequently, P300 amplitude was larger during retrieval practice trials than during first study trials,  $F(1,27)=47.828$ ,  $p<.001$ ,  $\eta^2_p=.639$ .

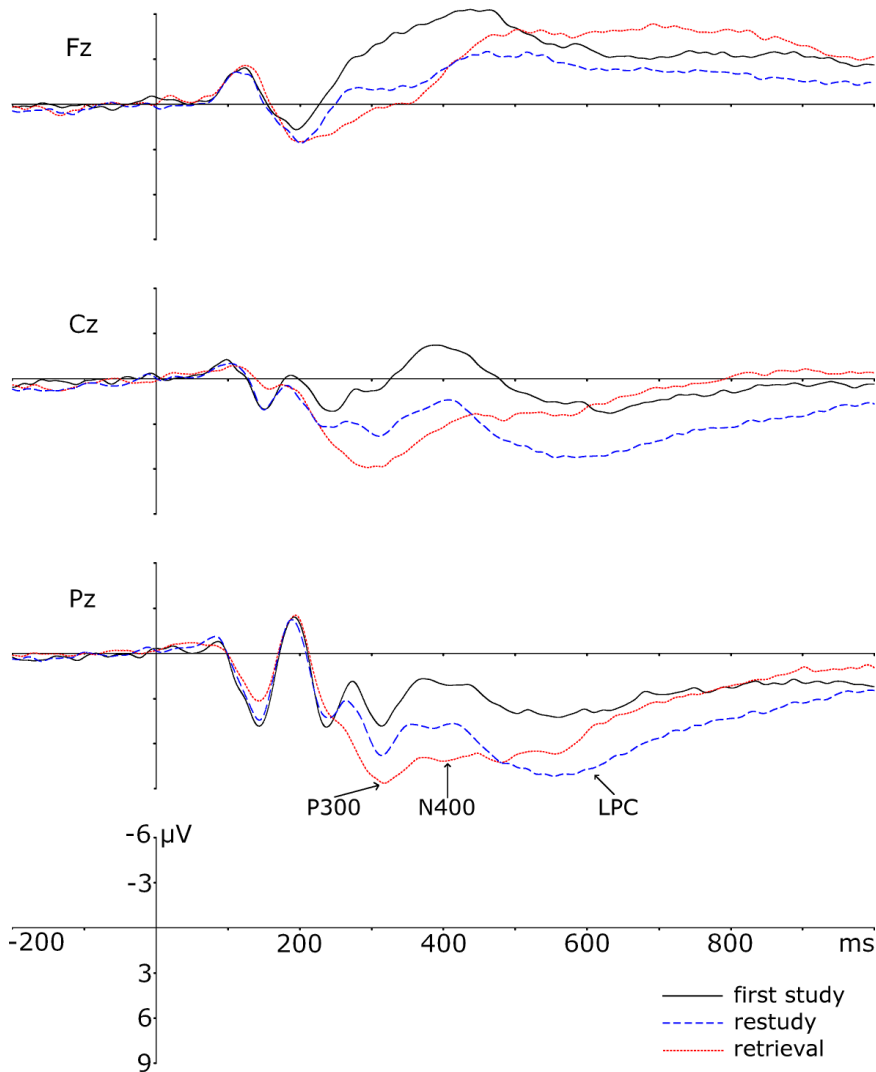


Figure 2. Grand average ERPs during first study, restudy and retrieval trials at the analyzed electrodes.

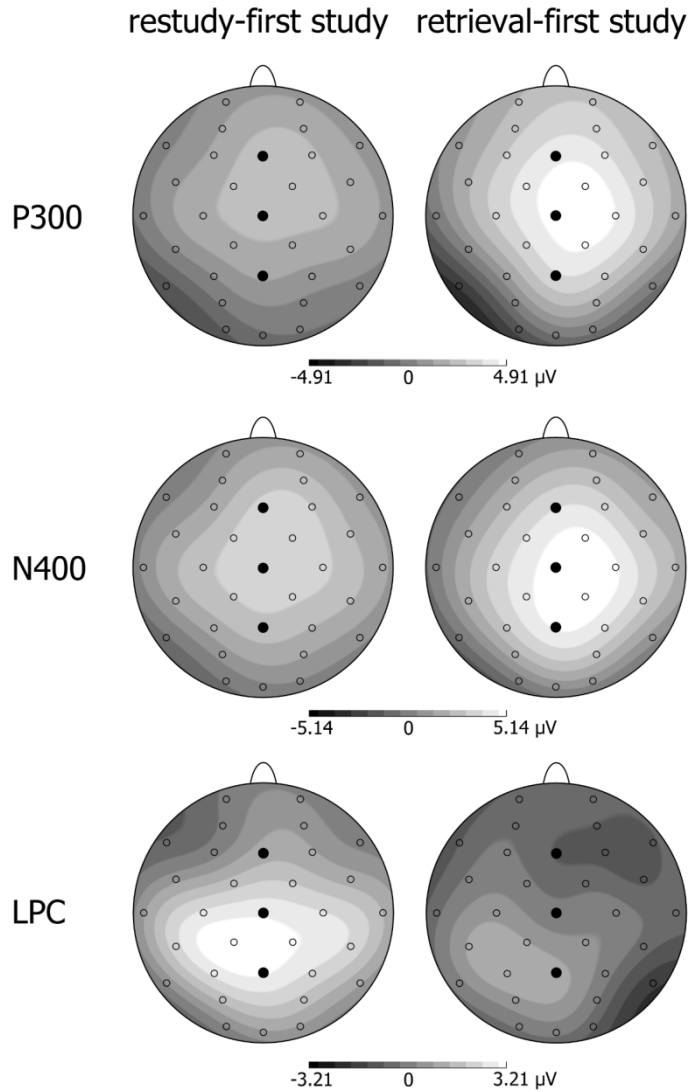


Figure 3. Scalp distributions of the restudy-first study and retrieval-first study difference waves for P300 (250-350 ms), N400 (350-450 ms), and LPC (500-800 ms). The analyzed electrodes are shown in black.

**N400.** Scalp distributions of the restudy-first study and retrieval-first study difference waves in the 350-450 ms time window are shown in Figure 3. This figure shows that, for both the restudy and the retrieval condition, the N400 repetition effect was widely distributed across the scalp. The 3 (trial type)  $\times$  3 (electrode) ANOVA on the 350-450 ms timeframe yielded a significant main effect of trial type,

$F(2,26)=31.022$ ,  $p<.001$ ,  $\eta^2_p=.705$ . Subsequent Bonferroni-corrected pairwise comparisons revealed that the N400 was less negative during restudy trials than during first study trials,  $F(1,27)=46.140$ ,  $p<.001$ ,  $\eta^2_p=.631$ , and less negative during retrieval trials than during first study trials,  $F(1,27)=46.472$ ,  $p<.001$ ,  $\eta^2_p=.633$ . There was no difference in N400 amplitude between restudy and retrieval trials,  $F(1,27)=3.654$ ,  $p=.200$ ,  $\eta^2_p=.119$ .

**LPC.** Scalp distributions of the restudy-first study and retrieval-first study difference waves in the 500-800 ms time window are shown in Figure 3. This figure shows that the repetition effect for the restudy trials was located at the typical parietal region. The 3 (trial type)  $\times$  3 (electrode) ANOVA on the 500-800 ms time window yielded a significant effect of trial type,  $F(2,26)=12.557$ ,  $p<.001$ ,  $\eta^2_p=.491$ , and a Trial Type  $\times$  Electrode interaction effect,  $F(4,24)=3.647$ ,  $p=.019$ ,  $\eta^2_p=.378$ . Bonferroni-corrected pairwise comparisons indicated that the LPC was more positive during restudy than during first study trials,  $F(1,27)=22.264$ ,  $p<.001$ ,  $\eta^2_p=.452$ , and more positive during restudy than during retrieval trials,  $F(1,27)=10.610$ ,  $p=.009$ ,  $\eta^2_p=.282$ . There was no difference in LPC amplitude between first study and retrieval trials,  $F(1,27)=0.019$ ,  $p=1$ ,  $\eta^2_p=.001$ .

## Discussion

The goals of the present experiment were to test the hypothesis that the testing effect arises from differences in processing effort, and to investigate the processes that underlie this increased effort. We used ERPs to assess effort (P300), familiarity (N400) and recollection (LPC). Our behavioral data showed that retrieved items were more often correctly reported than repeatedly studied items at the final test after two days. Given that the testing effect is usually observed after a relatively long retention interval, this result indicates a classic testing effect (Roediger & Butler, 2011).

According to the retrieval effort hypothesis, effortful successful retrieval is more beneficial to retention than less effortful successful retrieval. Our hypothesis was that this difference in retrieval effort should also be present when comparing restudy and retrieval practice, because a restudy trial also entails retrieval of a previous study trial. Based on this hypothesis, we predicted that the P300 amplitude would be larger during retrieval practice trials than during restudy trials. This would indicate higher retrieval effort during retrieval practice than restudying. In line with our expectations, the amplitude of P300 was larger for retrieval trials than for restudy trials. This finding is in accordance with the desirable difficulties framework. An assumption of this framework is that increased processing effort during retrieval drives the testing effect.

In the present experiment, retrieval trials elicited a larger P300 amplitude, indicating more processing effort. This suggests that effort during learning indeed plays a role in the testing effect and therefore supports theories based on the desirable difficulties framework, such as the retrieval effort hypothesis (Pyc & Rawson, 2009).

In addition to the differences in retrieval effort indexed by P300 amplitude, we aimed to investigate the processes underlying this increased effort by looking at the N400 and LPC. We hypothesized that the retrieval of previous study episodes during retrieval trials focuses more on semantic aspects of the item and less on contextual details, compared to restudy trials. Based on this hypothesis, we predicted that the N400 (familiarity) repetition effect should be more pronounced during retrieval trials than during restudy trials. This would suggest that study-phase retrieval during test trials is more focused on the meaning of the items than study-phase retrieval during restudy trials. In regard to the LPC repetition effect that indicates retrieval based on contextual details, we predicted that the LPC amplitude would be smaller during retrieval practice than during restudying. This would suggest less detail-focused processing during retrieval trials.

Indeed, we observed a stronger N400 repetition effect during retrieval trials than during restudy trials, and a stronger LPC repetition effect during restudy trials than during retrieval trials. The LPC repetition effect was even absent during retrieval trials. Taken together, these findings support our hypothesis that compared to restudy trials, processing during retrieval trials focuses more on semantic aspects of the words and less on the contextual details of the previous presentation. These findings are in line with testing effect explanations that rely on elaboration (e.g., elaborative retrieval, Carpenter, 2009; mediator effectiveness, Pyc & Rawson, 2010) and suggest that the increased processing effort during retrieval is at least partly caused by the retrieval of semantic information.

In sum, we found a clear effect of testing and the ERP patterns during learning suggest that more attention is allocated to retrieval items than to restudy items. This provides support for the effortful retrieval hypothesis. In addition, our results suggest that during retrieval, processing of the items was focused on semantic properties of the words instead of on the contextual details.

## Appendix: Used Dutch words with English translations

cue	target
afgrond (abyss)	horloge (watch)
antenne (antenna)	klimaat (climate)
atleet (athlete)	gesprek (conversation)
bedrijf (company)	lengte (length)
brommer (moped)	slinger (pendulum)
cowboy (cowboy)	emotie (emotion)
daalder (thaler)	vulkaan (volcano)
drempel (threshold)	zwabber (mop)
element (element)	buffel (buffalo)
ernstig (serious)	fregat (frigate)
fabriek (factory)	lollie (lollipop)
filter (filter)	eiland (island)
fontein (fountain)	citroen (lemon)
garnaal (shrimp)	advies (advice)
gieter (watering can)	rijkdom (wealth)
halter (halter)	woning (home)
hippie (hippie)	karton (card board)
keuken (kitchen)	piloot (pilot)
knecht (servant)	papaver (poppy)
kruipen (crawl)	cursist (student)
lucifer (match)	olifant (elephant)
metaal (metal)	armband (bracelet)
monnik (monk)	twijfel (doubt)
notitie (note)	gulden (guilder)
omvang (size)	zakdoek (handkerchief)
plafond (ceiling)	etiket (label)
proces (process)	balkon (balcony)
rozijn (raisin)	bitter (bitter)
schort (apron)	nikkel (nickel)
sigaar (cigar)	rugzak (backpack)
snavel (beak)	nummer (number)
suiker (sugar)	minuut (minute)
toneel (stage)	amandel (almond)
trommel (drum)	visser (fisherman)
typist (typist)	fluweel (velvet)
vergiet (strainer)	muziek (music)
vlieger (kite)	gorilla (gorilla)
vrucht (fruit)	dochter (daughter)
wrijven (rub)	opgave (task)
zuster (nurse)	deksel (lid)



# 4

## No ERP evidence for increased elaboration during retrieval practice

This chapter is in preparation as:

Coppens, L. C., Verkoeijen, P. P. J. L., & Rikers, R. M. J. P. (in preparation). No ERP evidence for increased elaboration during retrieval practice.

### Abstract

A common explanation of the testing effect (i.e., long-term retention is better for materials learned through retrieval practice than through restudying) involves elaboration during retrieval. In the current experiment, we aimed to investigate this elaboration hypothesis using homonyms in an event-related potential study. Participants learned word pairs, consisting of two words related to one meaning of a homonym (e.g., *river* – *coast*, both words are related to the homonym *bank*) through restudying or retrieval practice. Based on the elaborative retrieval hypothesis, we hypothesized that retrieval practice would lead to activation of the homonym *bank* more often than restudying. Subsequently they read sentences containing the homonym in the other meaning (e.g., *he needed money so he went to the bank*). We predicted a higher semantic conflict-indicating N400 amplitude on the homonym after retrieval practice of the related word pairs than after restudying the word pairs. We did not find such a difference, which is problematic for the elaboration hypothesis of the testing effect.

When information is retrieved from memory during a test, it is remembered better in the long term than information that is only restudied. This phenomenon is known as the testing effect (for a review, see Roediger & Butler, 2011). In a typical testing effect experiment, participants study word pairs and subsequently restudy the word pairs or try to retrieve them from memory in test trials. After a retention interval of several days, a final test of the word pairs is administered and participants remember more tested word pairs than restudied word pairs. The testing effect in learning word pairs is often explained in terms of elaboration (Carpenter, 2009; Carpenter & DeLosh, 2006; Pyc & Rawson, 2010, 2012). An example of an elaboration-based account of the testing effect is the elaborative retrieval hypothesis (Carpenter, 2009; Carpenter & DeLosh, 2006). According to this hypothesis, which is based on spreading activation theories of memory (e.g., Collins & Loftus, 1975), information that is semantically related to the cue-target pair is activated during learning and coupled with the word pair. This activation of related information is assumed to happen more often during testing than during restudying. Subsequently, on the final test, the semantically related information helps retrieval by providing extra retrieval cues. As a consequence, targets from previously tested word pairs are more likely to be retrieved than targets from restudied word pairs.

Carpenter and DeLosh (2006) tested the elaborative retrieval hypothesis in a series of three experiments. In Experiment 3, participants learned words through testing under different levels of 'cue support': During testing in the learning phase, varying numbers of letters were provided as a retrieval cue. The more letters were provided on the test during the learning phase, the worse participants performed on the final test. The authors explained this by proposing that giving more letters constrains the number of possible associations. Thus, they assumed that when less letters of the word were given (less cue support), there was more opportunity to elaborate. Hence, Carpenter and DeLosh took their findings as evidence in support of the idea that testing improves memory through elaboration.

In a more direct test of the elaborative retrieval hypothesis, Carpenter (2011) had participants learn related word pairs (e.g., *mother* – *child*) through restudying or testing. At the final test, participants indicated for three types of items whether they had been previously presented or not: the original cue (*mother*), a word related to the target (e.g., *birth*), or a 'semantic mediator', a word related to the original cue (e.g., *father*) that presumably is activated during testing. Semantic mediators were more often falsely recognized after testing than after restudying. This suggests that the mediators were indeed more often activated during testing than during restudying and that these activated mediators remained active, at least until the final test.

However, there might be an alternative explanation of Carpenter's results. We will discuss this explanation and present a way to investigate the activation of related words during testing that excludes the alternative explanation.

In the false memory literature, there is a discussion on the explanation of these memories. One explanation is the activation-monitoring theory (Roediger, Watson, McDermott, & Gallo, 2001), which is also used by Carpenter (2011). This theory states that false recognition of the critical lure of DRM-lists (lists of words highly related to a critical lure, which is not presented) is caused by spreading activation *during* learning. Thus, according to the activation-monitoring theory, the critical lure is activated during learning of the related words. On the final test, when the critical lure is presented, the lure will be falsely recognized because it was activated during learning and the source of the activation is no longer clear. However, according to global-matching models of false memory (e.g., Hintzman, 2001) false memories emerge during the final test. The global-matching explanation proposes that during the final test, presented words are compared to the memory traces of the words learned earlier. If a presented word matches the memory traces of previously learned words closely enough, it is judged as 'old'. The critical lures are highly related to the learned words, and thus highly likely to be incorrectly judged as 'old'. The global-matching account could explain Carpenter's (2011) results in the following way: during learning of the words, mediators are not activated. Testing of words strengthens the memory trace more than restudying. On the final test, the presented mediators are compared to the memory traces of the learned words. The mediators are highly related to the learned words. Therefore, they match the memory traces and are incorrectly judged as 'old'. Because memory traces for tested words are stronger than those for restudied words, tested words are more often incorrectly recognized than restudied words. Thus, the global-matching account is able to explain Carpenter's results without the assumption of mediator activation during learning.

In the current study we tested the activation of related concepts during testing not by presenting words that are related to supposedly activated concepts, but by presenting the concepts themselves during the final test. Therefore, any effect we find cannot be caused by global matching, because during the final test no concept-related (matching) words were presented. In particular, in the current experiment, we had participants learn word pairs through restudying or testing. The word pairs were related to homonyms (words with multiple meanings). An English example of a homonym is *bank*, which can denote a financial institution or a shoreline. In the learning phase, we presented word pairs that were strongly associated to the homonyms, such as *teller - account*. The elaborative retrieval hypothesis would predict

that upon testing of *teller - account*, more semantically related words are activated than upon restudying of this word pair. A word that is likely to be activated in this case is the homonym *bank*. We tested this activation directly after the learning phase, using sentences that featured the homonyms in the other meaning. If the word *bank* is indeed activated during testing of *teller - account*, then it is the financial institution meaning that is activated, not the shoreline. An encounter with the word *bank* in the meaning of a shoreline, such as in the sentence *He lay down in the grass on the bank* would thus result in a conflict between the two meanings of *bank*. We measured this conflict by recording electroencephalograms (EEG) during reading. More specifically, we measured N400 amplitude.

The N400 is a negative ERP component that occurs around 400 ms after stimulus onset (Kutas & Hillyard, 1980). N400 varies in amplitude with different levels of semantic integration difficulty (Chwilla, Kolk, & Vissers, 2007; Hagoort, Hald, Bastiaansen, & Petersson, 2004; Kutas & Federmeier, 2000). During sentence comprehension, the more difficult it is to integrate a word into the sentence the higher the N400 amplitude will be. This integration difficulty can either be caused by a semantic or a syntactical violation. For instance, in the first study that reported the N400 (Kutas & Hillyard, 1980) a larger N400 amplitude was found on the last word of sentences like *He spread the warm bread with socks*, compared to *He spread the warm bread with butter*. The word *socks* is harder to integrate into the sentence than the word *butter*, because one usually does not spread bread with socks. This integration difficulty is associated with a larger N400 amplitude.

In research on the N400 and homonyms, N400 is usually not measured on the homonym itself but on a disambiguating word that follows the homonym. For instance, Gunter, Wagner, and Friederici (2003), and Wagner and Gunter (2004) had participants read German sentences that contained a homonym followed by disambiguating context, for instance *The clay was baked by the potter* (the German word *ton* can mean either clay or tone, but a tone can obviously not be baked). The homonym was either used in the dominant meaning (e.g., *tone*), which presumably is activated when there is no disambiguating information present, or in the subordinate meaning (e.g., *clay*). Wagner and Gunter (2004; and Gunter et al., 2003, for high-span readers) found a larger N400 amplitude for the disambiguating cue related to the subordinate meaning of the homonym (e.g., *baked*) compared to the cue related to the dominant meaning. What we can take from these studies is that the activation of one meaning of a homonym (the dominant meaning *tone*, in this case) causes a conflict when the homonym is encountered in the other (subordinate, *clay*) meaning, and that this conflict can be measured using N400 amplitude.

In the current study, N<sub>400</sub> was measured on the homonym itself. Instead of disambiguating context following the homonym, the context was presented before the homonym so that the meaning of the homonym was clear upon reading it. All participants read the same sentences (e.g., *he lay down in the grass on the bank*). We expected differences in N<sub>400</sub> amplitude based on whether word pairs related to the homonyms were restudied or tested during the learning phase. If the elaboration hypothesis is correct, participants should activate the homonym *bank* more often during learning of *teller* - *account* through retrieval than during learning through restudy. Because the elaboration hypothesis assumes that activated information is semantically related, if the homonym *bank* is indeed activated during learning of *river* and *coast*, it is in the meaning of a shoreline. When a homonym is activated in one meaning (shoreline) during learning and is then encountered in another meaning (financial institution), it is more difficult to integrate the word into the sentence. N<sub>400</sub> amplitude varies with integration difficulty, so the N<sub>400</sub> amplitude on the homonym should be larger after testing than after restudying. Therefore, we predicted a larger N<sub>400</sub> amplitude on homonyms of which the related word pairs were learned through testing, compared to the homonyms of which the related word pairs were learned through restudying.

## Method

### Participants

Forty-one healthy native Dutch university students (16 female and 25 male, mean age 22.2 years, age range 18-29) with normal or corrected-to-normal vision participated for course credits or a reward of €15. Prior to the experiment, all participants gave written informed consent and the study was approved by the Ethical Committee Psychology of the Erasmus University Rotterdam.

### Materials

Thirty Dutch homonyms were selected (a list of the used materials is included in the Appendix). For each homonym, two words were selected to create word pairs that were associated with one meaning of the homonym. The associated words had an average word-homonym association value of 0.116 according to the CELEX database (Baayen et al., 1993). To control for word frequency, item difficulty and order effects, four different counterbalancing sequences were created (see Table 1). In addition to the word pairs pertaining to one meaning of the homonyms, 30 sentences were created that featured the homonym in the other meaning. To avoid EEG artifacts, the

sentences were constructed so that the homonym was the last word. Ten extra sentences were used as fillers.

### Procedure

The experiment consisted of three phases, presented to the participants as two independent experiments. In the first phase participants learned word pairs that were related to the homonyms. The second phase (presented to the participants as the second experiment) was the critical sentence reading task in which participants read sentences that contained the homonyms. The third phase (presented as the second part of the first experiment) consisted of a test of the word pairs learned in the first phase.

At the beginning of the learning phase, participants were informed that they would study word pairs and that they should try to remember as many of these pairs as possible for a later test. Word pairs were presented in black lower case font (Arial, 24 points) on a white background, on a 17-inch computer screen (resolution 1280 x 1024 pixels) placed 1.25 m from the participant. In a (re)study trial, the word pair was presented for 5 s with an interstimulus interval of 1 s. In a test trial, only the cue word was presented and participants had to retrieve and type the target word. Presentation duration was the same as for (re)study trials. Retrieved and restudied items were presented in blocks; a participant first learned 15 word pairs through 3 study trials and then learned the other 15 word pairs through one study trial and two practice trials, or vice versa. Word pairs were presented in a random order within blocks. Table 1 shows the block sequences.

Table 1

*Block sequence in the learning phase in the 4 counterbalance lists.*

Block	Activity in CB list 1	Activity in CB list 2	Activity in CB list 3	Activity in CB list 4
1	Study items 1-15	Study items 1-15	Study items 16-30	Study items 16-30
2	Restudy items 1-15	Retrieve items 1-15	Restudy items 16-30	Retrieve items 16-30
3	Restudy items 1-15	Retrieve items 1-15	Restudy items 16-30	Retrieve items 16-30
4	Study items 16-30	Study items 16-30	Study items 1-15	Study items 1-15
5	Retrieve items 16-30	Restudy items 16-30	Retrieve items 1-15	Restudy items 1-15
6	Retrieve items 16-30	Restudy items 16-30	Retrieve items 1-15	Restudy items 1-15

*Note.* Item order was random within blocks. CB = counterbalancing.

After the learning phase, participants received an instruction stating that they would see sentences and that after some of the sentences a yes/no question would appear. Questions were asked only about the filler sentences. Participants were instructed to carefully read the sentences and answer the questions using a keyboard. If there was no question, a '+' appeared and participants pressed any key to continue to the next sentence. Sentences were presented word by word in light gray lower case font (Arial, 22 points) on a black background, using the Variable Serial Visual Presentation procedure described by Otten and Van Berkum (2008). Words were presented for 290 ms plus 30 ms per letter, with a maximum of 590 ms. For words just before a comma or period extra time was added: 200 ms for a comma and 300 ms for a period. To ensure word length differences did not have an effect on the elicited ERPs because of different presentation durations, the homonyms had a fixed duration (650 ms) based on the average homonym length.

In the final phase the word pairs were tested. The cue words were presented on the screen one by one in a random order and participants typed the target words. The final test was self-paced.

### **EEG Recording**

Participants sat in a comfortable chair in a dimly lit room, separate from the experimenter. EEG was recorded from 64 active Ag/AgCl electrodes (BioSemi, Amsterdam, the Netherlands) placed in an elastic cap according to the international 10/20 system, at Fp1/z/3, AF7/3/z/4/8, F7/5/3/1/z/2/4/6/8, FT7/8, FC5/3/1/z/2/4/6, T7/8, C5/3/1/z/2/4/6, TP7/8, CP5/3/1/z/2/4/6, P9/7/5/3/1/z/2/4/6/8/10, PO7/3/z/4/8, O1/z/2, and Iz. Recordings were amplified using an ActiveTwo amplifier system and sampled at 512 Hz. An additional active electrode (CMS, common mode sense) and a passive electrode (DRL, driven right leg) were used as a feedback loop for amplifier reference. Two electrodes were placed on the mastoids, and eye movements and blinks were monitored by four additional electrodes.

### **EEG Analysis**

Data were processed and analyzed with Brain Vision Analyzer software (Brain Products, Gilching, Germany). Signals were re-referenced offline to the averaged mastoids. Eye-movement artifacts were removed using the algorithm of Gratton et al. (1983).

Data were filtered with a band-pass filter of 0.01 to 40 Hz, segmented into epochs of 1000 ms (-100 – 900 ms post stimulus onset) and baseline-corrected relative to the 100 ms before stimulus onset. Epochs including a signal exceeding  $\pm 100 \mu\text{V}$  were excluded from further analysis. Participants with less than 10 (of 15) valid (i.e.,



without artifacts) segments per condition were excluded from analysis, leaving 33 participants. The mean number of valid epochs per remaining participant was 12.8 in the restudy condition and 12.9 in the test condition. Mean ERPs were calculated by averaging trials for each participant, electrode and trial type separately.

The ERP analysis focused on the N400 component. N400 effects are usually found in the 300-500 ms time window (Chwilla et al., 2007; Hagoort et al., 2004; Kutas & Federmeier, 2000). Therefore, amplitude in this time window will be used as a measure for the N400 repetition effect.

### Statistical Analyses

To investigate the scalp topography of the N400, a repeated-measures analysis of variance (ANOVA) was conducted on electrodes F3/1/z/2/4, FC3/1/z/2/4, C3/1/z/2/4, CP3/1/z/2/4, P3/1/z/2/4, and PO7/3/z/4/8, with caudality (F, FC, C, CP, P, PO), laterality (left, left-midline, midline, right-midline, right) and trial type (restudy, test) as within-subjects variables. An alpha level of .05 was used for all statistical tests reported in this paper.

## Results

Minor errors in participants' typed responses, in which one letter was missing, added or in the wrong place, were corrected before analysis.

### Behavioral Data

Mean accuracy (proportion of targets correctly reported) on the test trials during the learning phase was .823. On the final test, mean accuracy for restudied items was .774 ( $SD = .233$ ), whereas accuracy for retrieved items was .675 ( $SD = .234$ ). Restudied items were correctly recalled more often than retrieved items,  $t(31) = 2.655$ ,  $p = .012$ ,  $d = 0.469$ .

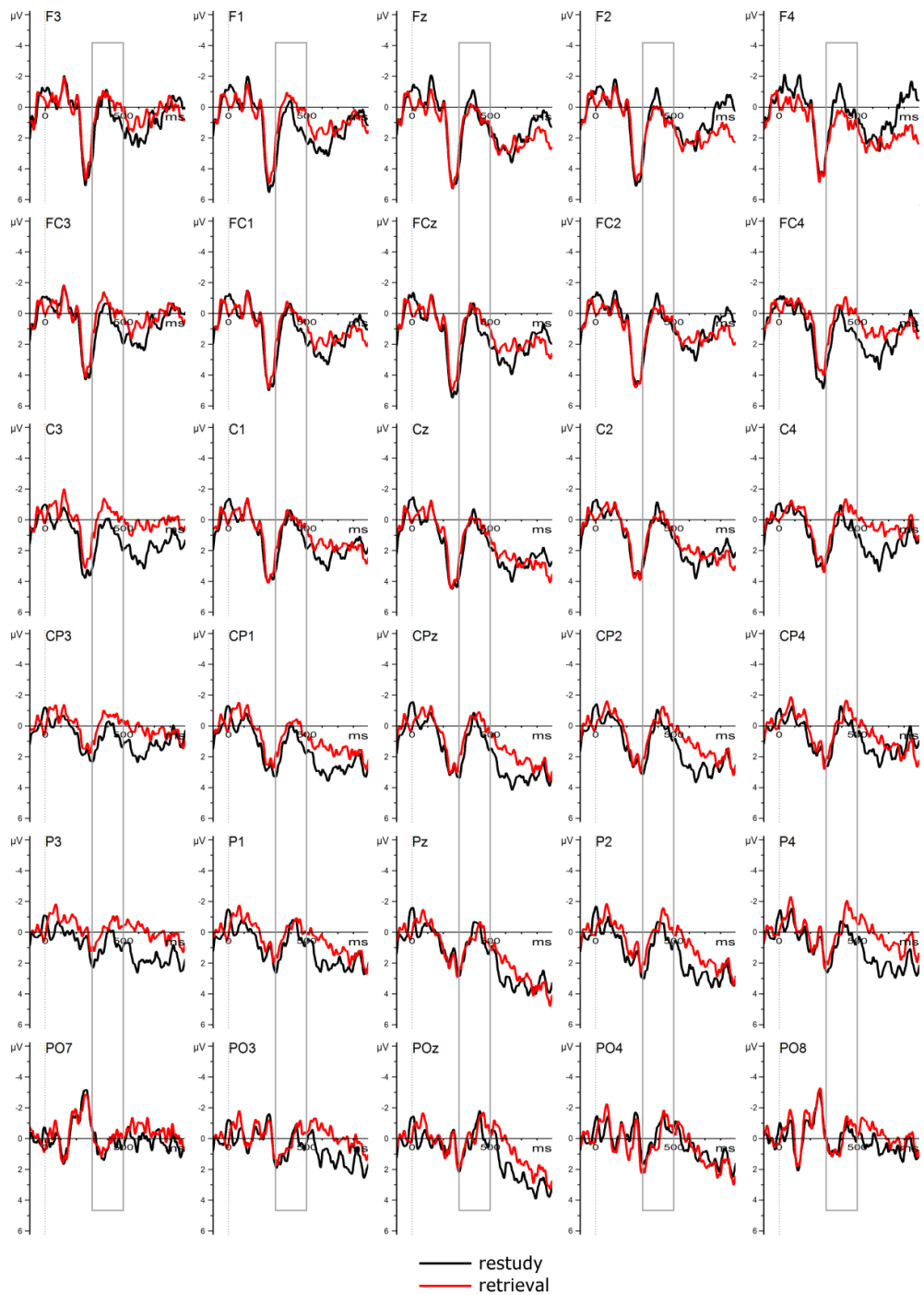


Figure 1. Grand average waveforms at the analyzed electrodes. Negative is plotted upwards. The grey boxes indicate the analyzed time interval (300-500 ms after stimulus onset).

## N400

Grand average waveforms at the analyzed electrodes are shown in Figure 1. This figure shows that there are only minimal differences between the two conditions. Because sphericity could not be assumed, we report the multivariate test. The  $6 \times 5 \times 2$  ANOVA yielded no significant results (all  $F_s < 2$ , all  $\eta_p^2 < .05$ ).

## Discussion

The goal of the present experiment was to test the elaborative retrieval hypothesis. Participants restudied or retrieved word pairs that were related to a homonym in one meaning and subsequently read sentences that contained the homonym in the other meaning. In regard to behavioral measures, restudied words were more often correctly recalled on the final test than retrieved words. This is not surprising, since the benefit of testing typically emerges only after a long retention interval. After a short retention interval, either no difference between restudy and testing is found or there is an advantage of restudying (see Toppino & Cohen, 2009). Because we administered the final test after about 20 minutes (i.e., a short retention interval), the restudy advantage we found is in line with previous findings in the literature.

In regard to the ERPs, we hypothesized based on elaboration accounts of the testing effect that during testing of the word pairs, the homonym would be activated more strongly than during restudying. During the reading phase the homonyms were presented in the other meaning. If the homonym had been activated during learning, then this presentation would result in a conflict between the two meanings of the homonym upon reading of the homonym in the sentence. Therefore, we predicted a higher (semantic conflict-indicating) N400 amplitude on the homonyms in the sentences after testing than after restudying. However, we did not find such an effect. That is, we found no difference in N400 amplitude between the two learning conditions. There are several possible explanations for this null effect.

First, one might argue that the interval between learning and reading was too long to find an effect of activation, and that we might have found an effect if we had presented the sentences directly after each learning trial. However, as discussed above, the effect of testing on memory is largest after a retention interval of several days. Therefore, if the activation of related concepts plays a role in the emergence of the testing effect, as the elaboration hypothesis assumes, the effects should be measurable after 20 minutes. Moreover, Coppens, Gootjes, and Zwaan (2012) found N400 modulations in a sentence reading task based on pictures presented 15 minutes

beforehand. Therefore, we do not think the current time interval of maximally 20 minutes between the learning and the reading phase can explain our lack of effect.

Second, it might seem far-fetched to suppose that one particular related word is activated during learning. Even if people elaborate during learning, one could argue that they might activate different related words than the ones we tested. However, we selected word pairs so that the related homonym was very likely to be activated according to the norms of Baayen et al. (1993). Moreover, a similar procedure testing only one related word was used successfully in a study by Carpenter (2011). Carpenter presented related word pairs and administered a final test using 'semantic mediators', words that were related to the original cues. Semantic mediators were more often falsely recognized after testing than after restudying. Moreover, semantic mediators were a more effective final test cue after testing than after restudying. These results show that it is possible to predict the words participants will activate during retrieval. We used an approach similar to that of Carpenter: We chose words that were related to the homonym that participants supposedly activated. Therefore, we think that it was highly likely that the participants activated the homonyms during learning.

Third, it is possible that our lack of N400 effect is due to an ERP noise issue. Because of the limited number of suitable homonyms and the limits of how many word pairs a participant is able to remember, we presented only fifteen reading trials per condition. Thus, after excluding segments containing artifacts, maximally fifteen segments were included in the ERP of a participant. The more segments are included in the average ERP, the better the signal-to-noise ratio becomes and the more reliable the ERP waveform is (Luck, 2005). A bad signal-to-noise ratio makes it difficult to find significant effects. Therefore, our low number of segments could make it difficult to find an effect. However, the equipment we used has an excellent signal-to-noise ratio because it uses active electrodes with a very low input impedance and a driven right leg circuit (see MettingVanRijn, Peper, & Grimbergen, 1990, 1991). With a good signal-to-noise ratio of individual segments, less segments are needed to obtain a reliable average. Moreover, there are studies in which N400 effects are found with low numbers of trials (Arzouan, Goldstein, & Faust, 2007; Franklin, Dien, Neely, Huber, & Waterson, 2007). Therefore, we think that our lack of effect is not due to a bad signal-to-noise ratio.

The last possible explanation of the fact that we did not find an effect is that the elaboration hypothesis does not provide a complete account of the testing effect. Specifically, it might be possible that the magnitude of elaboration does not constitute the crucial difference between restudying and testing. Indeed, there is existing evidence against the elaboration account of the testing effect. For instance,

Karpicke and Smith (2012) compared testing to repeated studying with two elaborative encoding strategies (an imagery-based strategy that uses keywords and a verbal elaboration method). The authors predicted that, if the testing effect is indeed caused by elaboration, the elaborative encoding strategies would have the same effect on learning as testing. However, testing led to better final test performance than elaborative restudying. Furthermore, when the to-be-learned stimuli included not only unrelated word pairs (e.g., *mountain – hammer*) but also identical word pairs (e.g., *castle – castle*) that required no elaboration, there was still a testing effect for the identical word pairs. The authors conclude from these results that the testing effect is not likely to be caused by elaboration and propose an alternative account: cue diagnosticity. According to this account testing enhances memory by improving the effectiveness of retrieval cues generated during learning. This improvement could be in the form of targets being easier to find on the final test or competing target words becoming less likely to be activated.

In sum, we did not find evidence for the activation of related concepts during testing of cue-target pairs. Although this finding does not provide conclusive evidence, it does suggest that the role of elaboration in the emergence of the testing effect is limited. Additional research is required to investigate the underlying mechanisms that contribute to the testing effect.

## Appendix: Critical stimuli with English translations

Homonym	Word pair in part 1	Sentence in part 2
arm <i>arm / poor</i>	horloge – schouder <i>watch – shoulder</i>	Omdat hij geen werk meer had, was hij erg arm. <i>Because he was unemployed, he was very poor.</i>
baan <i>track / job</i>	atletiek – racen <i>athletics – racing</i>	Hij was erg blij met zijn nieuwe baan. <i>He was very happy with his new job.</i>
bal <i>ball / ball</i>	Assepoester – kostuum <i>Cinderella – costume</i>	Hij schopte tegen de bal. <i>He kicked the ball.</i>
bank <i>bench / bank</i>	park – zitten <i>park – sit</i>	Ze bracht haar geld naar de bank. <i>She brought her money to the bank.</i>
blik <i>can / look</i>	bonen – opener <i>beans – opener</i>	In zijn ogen had hij weer die gemene blik. <i>In his eyes, he had again that mean look.</i>
bloem <i>flower / flour</i>	tuin – vlinder <i>garden – butterfly</i>	In de cake ging boter en bloem. <i>Into the cake went butter and flour.</i>
bril <i>glasses / toilet seat</i>	secretaresse – lezen <i>secretary – reading</i>	Op de wc bleek dat hij niet had gedacht aan de bril. <i>On the toilet it appeared he hadn't thought of the seat.</i>
kater <i>hangover / tomcat</i>	alcohol – pijn <i>alcohol – pain</i>	Over de schutting liep een dikke rode kater. <i>On the fence walked a fat red tomcat.</i>
klok <i>clock / bell</i>	wekker – tijd <i>alarm clock – time</i>	Na de bruiloft luidden de klokken. <i>After the wedding rang the bells.</i>
kolen <i>coal / cabbages</i>	schoorsteen – mijn <i>chimney – mine</i>	Eigenlijk zijn spruitjes ook gewoon kolen. <i>Brussels sprouts are actually just cabbages.</i>
kop <i>head / cup</i>	hoofd – staart <i>head – tail</i>	Hij schonk de koffie in de kop. <i>He poured the coffee into the cup.</i>
koper <i>copper / buyer</i>	metaal – trompet <i>metal – trumpet</i>	Gelukkig hadden ze voor het huis toch een koper. <i>Luckily, they had for the house a buyer.</i>
kraan <i>faucet / crane</i>	water – loodgieter <i>water – plumber</i>	De container werd de tuin in gehesen met een kraan. <i>The container was hoisted into the yard using a crane.</i>
lijst <i>frame / list</i>	spiegel – schilderij <i>mirror – painting</i>	Hij zette de boodschappen op de lijst. <i>He put the groceries on the list.</i>
monster <i>monster / sample</i>	draak – schotland <i>dragon – Scotland</i>	Om de vloeistof te onderzoeken, nam hij een monster. <i>To examine the liquid, he took a sample.</i>
muis <i>mouse / mouse</i>	knaagdier – grijs <i>rodent – gray</i>	Toen de computer vastliep, gooide ze met de muis. <i>When the computer crashed, she threw the mouse.</i>
munt <i>mint / coin</i>	peper – tandpasta <i>pepper – toothpaste</i>	Ze kon niet kiezen, dus gooide ze een munt. <i>She couldn't decide, so she tossed a coin.</i>
noot <i>nut / note</i>	pit – eikel <i>seed – acorn</i>	Het publiek schrok, want de pianist speelde een valse noot. <i>The audience was shocked, because the pianist played a false note.</i>
pad <i>toad / path</i>	amfibie – schild <i>amphibian – shell</i>	De fiets moest de tuin in, dus hij liep over het pad. <i>The bike had to go into the garden, so he walked on the path.</i>
pasta <i>pasta / (chocolate) paste</i>	graanproduct – spaghetti <i>cereal product – spaghetti</i>	Hij wilde geen pindakaas op zijn brood, maar pasta. <i>He didn't want peanut butter on his sandwich, but paste.</i>
pony <i>pony / bangs</i>	zadel – paard <i>saddle – horse</i>	Ze ging naar de kapper voor het bijknippen van haar pony. <i>She went to the hair dresser for trimming her bangs.</i>
pop <i>doll / pupa</i>	speelgoed – sneeuwman <i>toys – snowman</i>	De vlinder kroop uit de pop. <i>The butterfly crawled out of the pupa.</i>
roos <i>bullseye / rose</i>	pijl – schot <i>arrow – shot</i>	Met valentijnsdag kreeg hij een rode roos. <i>On Valentine's day, he got a red rose.</i>
schaal <i>platter - scale</i>	magnetron – fruit <i>microwave – fruit</i>	Hij baalde omdat zijn baas hem had ingedeeld in een lagere schaal. <i>He was annoyed because his boss had put him on a lower pay scale.</i>
school <i>school / school</i>	les – rugzak <i>class – backpack</i>	Op vakantie zag ze haaien, een keer zelfs een hele school. <i>On holiday she saw sharks, once even an entire school.</i>
slot <i>lock / castle</i>	kluis – deur <i>safe – door</i>	Ze wilde trouwen in Frankrijk, in een prachtig slot. <i>She wanted a wedding in France, in a beautiful castle.</i>
tocht <i>trip / draught</i>	fakkel – safari <i>torch – safari</i>	Hij had het koud, want hij zat op de tocht. <i>He was cold, because he was sitting in a draught.</i>
veer <i>feather / spring</i>	indiaan – pluimpje <i>indian – plume</i>	Dat de balpen stuk was, kwam door de kapotte veer. <i>That the ballpoint pen was broken was caused by the broken spring.</i>
was <i>wash / wax</i>	strijkijzer – mand <i>iron – basket</i>	Ze maakte kaarsen uit een lont en was. <i>She made candles from a wick and wax.</i>
zin <i>sentence / energy</i>	papier – punt <i>paper – period</i>	Ze moest naar de tandarts, maar ze had totaal geen zin. <i>She had to go to the dentist, but she had no energy.</i>

# 5

No evidence for the semantic mediator hypothesis: The testing effect in cued recall is similar for mediator cues and related cues

5

This chapter is in preparation for publication as:

Coppens, L. C., Verkoeijen, P. P. J. L., Bouwmeester, S., & Rikers, R. M. J. P. (in preparation). No evidence for the semantic mediator hypothesis: The testing effect in cued recall is similar for mediator cues and related cues.

### Abstract

The testing effect refers to the finding that information that is retrieved during learning is more often correctly retrieved on a final test than information that is restudied. According to the elaborative retrieval hypothesis, the testing effect in learning word pairs arises because retrieval practice of cue-target pairs (*mother-child*) activates semantically related mediators (*father*) more than restudying. Hence, the mediator-target (*father-child*) association should be stronger for retrieved than restudied pairs. Indeed, Carpenter (2011) found a larger testing effect when participants received mediators (*father*) as final test cues than when the final test cues were non-mediator words related to the target (*birth*). In Experiment 1, we investigated whether the association from the mediator to the cue could explain these findings by manipulating the mediator-cue association. The mediator-cue association did not influence the effectiveness of mediators as final test cues. Surprisingly, in Experiment 1 we did not find a consistently larger testing effect with mediators as final test cues compared to target-related words. In Experiments 2 and 3, direct replications of Carpenter again did not yield a larger testing effect for mediator final test cues than for related cues. We combined the findings of our three experiments with those of Carpenter in a small-scale meta-analysis and found no evidence for a difference in the magnitude of the testing effect between mediator and related final test cues. These results are at conflict with the semantic mediator hypothesis, but possibly are in line with alternative explanations of the testing effect such as cue diagnosticity.



The testing effect refers to the finding that performance on a final memory test is generally better when previously studied information has been retrieved from memory than when it has been restudied. The widely investigated testing effect has proven to be a robust phenomenon as it has been demonstrated with various final memory tests, materials, and participants (Karpicke, 2012; Roediger & Butler, 2011; Roediger & Karpicke, 2006a).

Although the testing effect has been well established empirically, the underlying cognitive mechanisms that contribute to the emergence of the effect are less clear. Recently, Carpenter (2009) suggested that elaborative processes are underlying the testing effect (see Verhoeijen et al., 2012, for a similar account). According to her elaborative retrieval hypothesis, retrieving the target based on the cue during practice causes more elaboration than restudying the entire pair. This elaboration helps retrieval at a final memory test because the related information that is activated during retrieval practice of the target is coupled with the target, hence creating additional retrieval routes. To exemplify the proposed theoretical mechanism, consider a participant who has to learn the word pair *mother - child*. When the target has to be retrieved from the cue (i.e., *mother*) this is more likely to lead to the activation of information associated with that cue (e.g., *love, father, diapers*) than restudying the entire word pair. As a result, the activated information becomes associated with the target (i.e., *child*) thereby providing additional retrieval routes to the target. Due to these additional retrieval routes, final test performance will be better for targets from word pairs learned through testing/retrieval practice than for targets from restudied word pairs.

However, Carpenter (2011) noted that the elaborative retrieval hypothesis was not specific about exactly what related information is activated during retrieval practice. To address this shortcoming, she turned to the mediator effectiveness hypothesis put forward by Pyc and Rawson (2010, 2012). Based on the mediator effectiveness hypothesis, Carpenter proposed that *semantic mediators* might be more likely to get activated during retrieval practice than during restudying (henceforth denoted as the semantic mediator hypothesis). In her paper, Carpenter defined a semantic mediator as a word that according to the norms of Nelson, McEvoy, and Schreiber (1998) has a strong forward association with the cue (i.e., when given the cue people will often spontaneously activate the mediator) and that is easily coupled with the target. For instance, for the word pair *mother-child*, the cue (*mother*) will elicit - at least for a vast majority of people - the word *father*. The word *father* can easily be coupled with the target *child*. Hence, *father* is the semantic mediator in case of this particular word pair. The semantic mediator hypothesis predicts that the link between

the semantic mediator *father* and the target *child* will be stronger after retrieval practice than after restudying. Also, this link should be stronger than the link between a *new related cue* (henceforth denoted as *related cue*) such as *birth* and the target *child*, because the cue *mother* has no pre-existing forward association with *birth*. Hence, *birth* is unlikely to get activated during retrieval practice.

Carpenter (2011, Experiment 2) tested exactly these predictions using cue-target pairs such as *mother* - *child*. These word pairs were studied and then restudied once or retrieved once. After a 30-minute distractor task, participants received a final test with either one of three cue types: the original cue, semantic mediators or related cues. The latter two are relevant for the present study. Carpenter's results showed a testing effect in the original cue condition. Moreover, at the final test the advantage of retrieval practice over restudying was greater when participants were cued with a mediator (*father*) than when they were cued with a related word (*birth*). Furthermore, targets from the retrieval practice condition were more often correctly produced during the final test when they were cued with mediators than when they were cued with related words. This difference in memory performance between mediator-cues and related-cues was much smaller for restudied items. These results of Carpenter's second experiment are important because they provide the first direct empirical support for a crucial assumption of the semantic mediator hypothesis; the assumption that the link between a mediator and a target is strengthened more during retrieval practice than during restudying.

However, in our view there might be an alternative explanation for the findings of Carpenter's (2011) second experiment. Specifically, we noted that some of the mediators used in this study were quite strongly associated with the cue. For example, one of the word pairs was *mother* – *child* with the mediator *father* and the related word *birth*. In this case, there is a strong cue-mediator association from *mother* to *father* (and no forward association from *mother* to *birth*), but the mediator *father* is also strongly associated with the original cue *mother* (.706 according to the norms of Nelson et al., 1998). Now it might be possible the larger testing effect on a mediator-cued final test (*father* - \_) as opposed to a related word-cued final test (*birth* - \_) was caused by mediators with strong mediator-cue associations. That is, when given the mediator at the final test, participants can easily retrieve the original cue *mother* if the mediator is strongly associated with the original cue. Because it is easier to retrieve the target from the original cue after retrieval practice than after restudying (in Carpenter's Experiment 2, final test performance was better for tested than for restudied items; cf. Carrier & Pashler, 1992; Halamish & Bjork, 2011), activation of the original cue through the mediator will facilitate retrieval of the target more after

retrieval practice than after restudying. By contrast, the related final test cues in Carpenter's experiment did not have an associative relationship with the original cues, and therefore it is harder to retrieve the original cue from a related final test cue than from a mediator final test cue. If the testing effect emerges due to a strengthened cue-target link then related final test cues are less likely to produce a testing effect than mediator final test cues. Thus, strong mediator-cue associations in Carpenter's stimulus materials in combination with a strengthened cue-target link might explain why the testing effect was larger for mediator final test cues than for related final test cues.

To test this alternative explanation of the results of Carpenter's Experiment 2, we repeated the experiment with new stimuli. We created two lists of 16 word sets that consisted of a cue, a target, a mediator, and a word that was related to the target (see Figure 1).

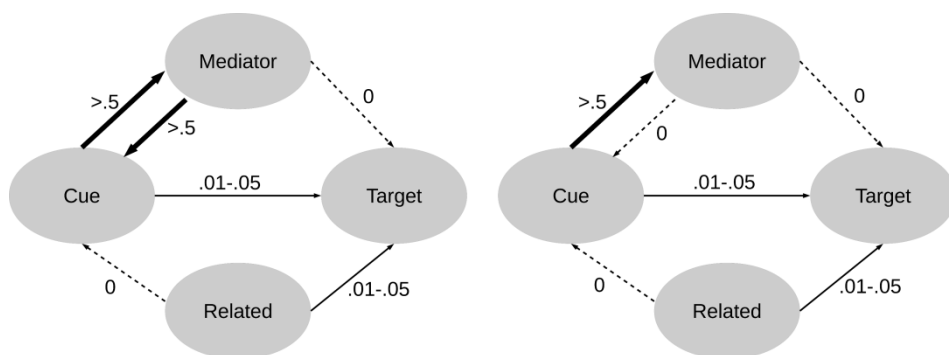


Figure 1. Word associations in Experiment 1. In the strong mediator-cue association condition (left), there was a strong association between the mediator and the cue. In the no mediator-cue association condition (right), there was no association between the mediator and the cue.

In both the stimuli lists, there was a weak cue-target association, a strong cue-mediator association and a weak association between the related word and the target. The difference between the two stimuli lists was the mediator-cue association. In one stimuli list, there was a strong mediator-cue association (as illustrated in the left part of Figure 1). This corresponds with the situation in some of the stimuli of Carpenter (2011), such as *mother* – *child* with the mediator *father*. In the other stimuli list, there was no mediator-cue association (as illustrated in the right part of Figure 1). An example of such a word set is the pair *anatomy* - *science* with the mediator *body*. There is no pre-existing association from *body* to *anatomy*. Therefore, if the proposed

mediator *body* is not activated during learning it will not activate the original cue *anatomy* and the alternative route from the mediator through the original cue to the target is blocked.

If our alternative account is correct and the larger testing effect in the mediator-cue final test condition is caused by a strong mediator-cue association, then the stimuli with a strong mediator-cue association should yield a replication of the pattern Carpenter (2011) found: a larger testing effect on a mediator-cued final test than on a related-word-cued final test. By contrast, for stimuli without a mediator-cue association the magnitude of the testing effect should not differ between mediator final test cues and related final test cues. It should be noted that Carpenter's semantic mediator hypothesis predicts a larger testing effect on a mediator-cued final test than on a related-word-cued final test for both stimuli lists.

## Experiment 1

### Method

**Participants.** 235 United States residents completed the experiment via the online work marketplace Amazon Mechanical Turk (MTurk; <http://www.mturk.com>). We used such a large sample because of two reasons: first, the crucial interaction effect might be small and we wanted to have sufficient power to detect such an effect. Second, MTurk participants usually show more variability in their data than the traditionally used (psychology) undergraduate participants. With that said, participants recruited via MTurk are more diverse than college students, which is beneficial to the external validity of research. Also, the test-retest reliability of the results is at least as high as that from traditional methods of data collection (Buhrmester, Kwang, & Gosling, 2011).

Participants were paid \$1.50 for their participation. The data of 9 participants were not included in the analysis because their native language was not English, leaving 226 participants (142 females, 84 males, age range 19-66, mean age 35.4,  $SD=11.7$ ). Participants were randomly assigned to conditions. In Table 1, the distribution of participants over the cells of the factorial design is shown.

**Materials and design.** A 2 (list: strong mediator-cue association vs. no mediator-cue association)  $\times$  2 (learning condition: restudy vs. retrieval practice)  $\times$  2 (final test cue: mediator vs. related) between-subjects design was used. To investigate the effect of the mediator-cue association, we used the association norms of Nelson et al. (1998) to create two lists of 16 word sets (see Appendix A). Each word set consisted of a cue and a target (weak cue-target association, .01 - .05), a mediator (strong cue-mediator

Table 1

*Number of participants in Experiment 1 as a function of mediator-cue association, learning condition and final test cue*

List	Final test cue	
	Related	Mediator
No M-C association		
Restudy	28	29
Retrieval practice	27	28
Strong M-C association		
Restudy	25	29
Retrieval practice	33	27

association,  $>.5$ ) and a related word (weak related word-target association,  $.01 - .05$ ). The difference between the two lists was in the mediator-cue associations. In one of the lists, the mediator-cue association in each word set was always higher than  $.5$ . In the other list, the mediator-cue association was always 0 (see Figure 1).

The experiment was created and run in Qualtrics (Qualtrics Labs Inc., Provo, UT), which allows timing and randomization of stimuli.

**Procedure.** The procedure was identical to that of Experiment 2 of Carpenter (2011), with the exception of the original cue final test condition, which we did not include because it was not relevant to the current research question. The experiment was placed as a task on MTurk with a short description of the experiment ('this task involves learning word pairs and answering trivia questions'). When a worker was interested in completing the task, she or he could participate in the experiment by clicking on a link and visiting a website.

The welcome screen of the experiment included a description of the task and questions about participants' age, gender, mother tongue, level of education, and working environment at the time of starting the task. After the participant answered these questions, the learning phase began. In the learning phase, all 16 cue-target pairs in one of the lists were shown in a different random order for each participant. The cue was presented on the left side of the screen and the underlined target was presented on the right. The task of the participants was to judge how related the words were, on a scale from 1 to 5 (1 = not at all related – 5 = highly related), and to try to remember the word pairs for a later memory test. The study trials were self-paced. After the study trials, there was a short filler task of 30 seconds, which involved adding single-digit numbers that appeared rapidly on the screen. Then the cue-target pairs

were presented again in a new random order during restudy or retrieval practice trials. Restudy trials were the same as study trials; participants again indicated how related the words were on a scale from 1 to 5. In retrieval practice trials, only the cue word was presented and participants had to type the target in a text box to the right of the cue. Both the restudy and retrieval practice trials were self-paced, as was the case in Carpenter's (2011) Experiment 2.

After a filler task of 30 minutes, in which participants answered multiple-choice trivia questions (e.g., 'What does NASA stand for? A. National Aeronautics and Space Administration; B. National Astronauts and Space Adventures; C. Nebulous Air and Starry Atmosphere; D. New Airways and Spatial Asteroids'), the final test began. Participants were informed that they would see words that were somehow related to the second, underlined word of the word pairs they saw earlier, and that their task was to think of the target word that matched the given word and enter the matching word in a text box. An example, using words that did not occur in the experiment, was included to elucidate the instructions. During the final test, participants were either cued with the mediator or with the related word of each word pair. The cue was presented on the left side of the screen and participants entered a response into a text box on the right side of the screen. The final test was self-paced.

To end the experiment, participants rated some concluding statements about their motivation ('I found the experiment interesting', 'I only participated to earn money', 'The experiment was boring'), effort ('I tried my best to remember the word pairs') and concentration ('I was distracted during the experiment'). Participants answered these questions on a 5-point Likert scale. The duration of the entire experiment was about 45 minutes.

## Results

An alpha level of .05 was used for all statistical tests reported in this paper. Minor typing errors in which one letter was missing, added or in the wrong place were corrected before analysis. For Learning Condition x Final Test Cue interaction effects, we report the *b* weight because it is a good indication of the direction of the effect; a positive *b* weight demonstrates that the pattern of results is in line with Carpenter's (i.e., a larger testing effect on a mediator-cued final test than a related word-cued final test).

**Intervening test.** In the list with no mediator-cue association, the mean proportion of correct targets retrieved on the intervening test was .90 (*SD*=.15, range .31-1) in the mediator final-test condition and .82 (*SD*=.23, range 0-1) in the related final-test condition. In the list with a strong mediator-cue association, the mean proportion of correct targets retrieval on the intervening test was .93 (*SD*=.15, range

.38-1) in the mediator final-test condition and .95 ( $SD=.08$ , range .69-1) in the related final-test condition.

**Final test.** The proportion of correctly recalled targets on the final test for List 1 and List 2 is shown in Table 2.

Table 2

*Proportion of correctly recalled targets (SD) on the final test in Experiment 1 as a function of mediator-cue association, learning condition and final test cue*

List	Final test cue	
	Related	Mediator
No M-C association		
Restudy	.174 (.168)	.121 (.234)
Retrieval practice	.185 (.198)	.272 (.267)
Strong M-C association		
Restudy	.148 (.137)	.416 (.397)
Retrieval practice	.381 (.234)	.500 (.432)

**List 1: no mediator-cue association (No MC).** A 2 (learning condition: restudy vs. retrieval practice)  $\times$  2 (final test cue: related vs. mediator) between-subjects analysis of variance (ANOVA) yielded a marginally significant main effect of learning condition,  $F(1,108)=3.821$ ,  $p=.053$ ,  $\eta^2_p=.034$ . Numerically, mean target retrieval was higher for cue-target pairs learned through retrieval practice than through restudying (i.e., a marginally significant testing effect). The effect of final test cue was not significant,  $F(1,108)=0.164$ ,  $p=.686$ ,  $\eta^2_p=.002$ . This suggests that mean target retrieval did not differ between related final test cues and mediator final test cues. Furthermore, there was a marginally significant Learning Condition  $\times$  Final Test Cue interaction,  $F(1,108)=2.852$ ,  $p=.094$ ,  $\eta^2_p=.026$ ,  $b=.141$ . Although the interaction effect is small and non-significant despite the large sample size, the pattern of results is consistent with the pattern demonstrated by Carpenter (2011).

**List 2: strong mediator-cue association (Strong MC).** A 2 (learning condition: restudy vs. retrieval practice)  $\times$  2 (final test cue: related vs. mediator) between-subjects ANOVA revealed a significant medium sized main effect of learning condition,  $F(1,110)=6.813$ ,  $p=.010$ ,  $\eta^2_p=.058$ : mean target retrieval was higher for cue-target pairs learned through retrieval practice than through restudying (i.e., a testing effect). Furthermore, we found a significant main effect of final test cue,  $F(1,110)=10.179$ ,  $p=.002$ ,  $\eta^2_p=.085$ . The mean final test performance was better for

mediator final test cues than for related final test cues. The Learning Condition x Final Test Cue interaction was not significant,  $F(1,110)=1.506$ ,  $p=.222$ ,  $\eta^2_p=.014$ ,  $b=-.149$ . This finding is inconsistent with Carpenter's (2011) result; the effect is not significant and the pattern of results is not consistent with that of Carpenter (hence the negative  $b$  weight).

## Discussion

The results of Experiment 1 revealed no significant interaction effect between final test cue and learning condition in either of the two lists. The effect is small and the non-significance can hardly be attributed to a lack of power since our sample was sufficiently large. The pattern of sample means shows, however, a larger testing effect for mediator final test cues than for related final test cues in the list with no mediator-cue associations. This pattern of results is similar to the one observed by Carpenter (2011) in her second experiment. By contrast, in the list with strong mediator-cue associations, the testing effect was larger for related final test cues than for mediator final test cues. Taken together, these findings are not in line with the predictions based on our alternative account of the findings from Carpenter's second experiment. Reasoning from this account, we expected to replicate Carpenter's finding in the list with the strong mediator-cue associations. In addition, with respect to the list with no mediator-cue associations, we predicted similar testing effects for the mediator final test cues and the related final test cues. However, the findings from Experiment 1 are also inconsistent with the semantic mediator hypothesis. According to this hypothesis mediator final test cues ought to produce a larger testing effect than related final test cues both in the strong mediator-cue association list and in the no mediator-cue association list.

The outcomes of Experiment 1, which failed to corroborate the semantic mediator hypothesis, casts some doubt on the reliability of Carpenter's (2011) results. This doubt was amplified because Carpenter's second experiment had a 2 x 2 between subjects design with only 10 participants per cell. Such a small sample size is accompanied by imprecise parameter estimates that are prone to be overestimations once they appear in journals that favor significant results (i.e., a publication bias; Ferguson & Heene, 2012).

Given the outlined problem associated with small sample experiments, we reasoned it would be good to attempt to replicate Carpenter's (2011) second experiment with considerably larger samples. Doing so will yield a more precise estimate of the interaction between final test cue and the testing effect. In Experiment 2 of the present study, we conducted a replication of Carpenter's experiment, using the same procedure and learning materials that Carpenter used.



There were two differences between Carpenter's and our experiment: our participants were adult US residents tested via Mechanical Turk, whereas Carpenter tested college students in the lab.

## Experiment 2

### Method

**Participants.** 173 United States residents who had not participated in Experiment 1 completed the experiment via MTurk (<http://www.mturk.com>). Participants were randomly assigned to conditions of the factorial design mentioned below. They were paid \$1.60 for their participation. Eight participants were excluded from further analysis because their native language was not English, leaving 165 participants (99 females, 66 males, age 18-67, mean age 34.6,  $SD = 12.2$ ). Of these participants, 82 learned the word pairs through restudy and 83 learned the word pairs through retrieval practice. Forty-four participants in the restudy condition and 47 participants in the retrieval practice condition completed the final test with mediator cues. Thirty-eight participants in the restudy condition and 36 participants in the retrieval practice condition completed the final test with related cues.

**Materials and design.** We used a 2 (learning condition: restudy vs. retrieval practice)  $\times$  2 (final test condition: mediator vs. related) between-subjects design. Participants studied the same word pairs Carpenter (2011) used (see Appendix B). The experiment was programmed and run in Qualtrics (Qualtrics Labs Inc., Provo, UT).

**Procedure.** The procedure was identical to that of Experiment 1.

### Results

**Intervening test.** On the intervening test, participants correctly retrieved .91 ( $SD=.19$ , range .06-1.00) of the targets on average in the related final test cue condition, and .95 ( $SD=.09$ , range .50-1.00) in the mediator final test condition.

**Final test.** Table 3 shows the proportion correctly recalled targets on the final test per condition. Note that the average proportion correct was comparable to the average proportion correct in Carpenter's second Experiment. A 2 (learning condition: restudy vs. retrieval practice)  $\times$  2 (final test cue: mediator vs. related) between-subjects ANOVA yielded a significant main effect of learning condition,  $F(1,161)=10.038$ ,  $p=.002$ ,  $\eta^2_p=.059$ , indicating that final test performance was better for retrieved than restudied word pairs (i.e., a testing effect), and a main effect of final test cue,  $F(1,161)=11.914$ ,  $p<.001$ ,  $\eta^2_p=.069$ , indicating better final test performance with related cues than with mediator cues. There was no Learning Condition  $\times$  Final Test Cue interaction,  $F(1,161)=0.241$ ,  $p=.624$ ,  $\eta^2_p=.001$ ,  $b=-.041$ , indicating that the

effect of learning condition did not differ between final test cue conditions. Moreover, the pattern of sample means is contrary to Carpenter's results.

Table 3

*Proportion of correctly recalled targets (SD) on the final test in Experiment 2 as a function of learning condition and final test cue*

Learning condition	Final test cue	
	Related	Mediator
Restudy	.375 (.264)	.250 (.244)
Retrieval practice	.530 (.269)	.363 (.296)

## Discussion

The results from Experiment 2 are inconsistent with the results of Carpenter's (2011) second experiment, and with the mediator effectiveness hypothesis for that matter. This hypothesis predicts that the testing effect is larger for mediator cues than for related cues. Contrary to this prediction, we observed that the magnitude of the testing effect was comparable for mediator cues and for related cues. Also, the effect of final test cue was comparable for retrieved and restudied items.

In sum, the results from Experiment 2 deviated from Carpenter's (2011) original experiment even though our Experiment 2 was a replication of Carpenter's experiment. To investigate the reliability and reproducibility of these rather remarkable results, we attempted another replication of Carpenter's (2011) experiment in our third experiment.

## Experiment 3

### Method

**Participants.** 118 United States residents who had not participated in Experiment 1 or Experiment 2 completed the experiment via MTurk (<http://www.mturk.com>). Participants were randomly assigned to conditions. They were paid \$1.33 for their participation. Two participants were excluded from further analysis because their native language was not English, leaving 116 participants (78 females, 38 males, age 19-67, mean age 33.4,  $SD = 11.9$ ). Of these participants, 59 learned the word pairs through restudy and 57 learned the word pairs through retrieval practice. Thirty participants in the restudy condition and 26 participants in the retrieval practice condition completed the final test with mediator cues. Twenty-nine participants in the

restudy condition and 31 participants in the retrieval practice condition completed the final test with related cues.

**Materials, design, procedure.** Materials, design, and procedure were the same as in Experiment 2.

## Results

**Intervening test.** On the intervening test, participants correctly retrieved .94 ( $SD=.12$ , range .44-1.00) of the targets in the related final test cue condition and .96 ( $SD=.08$ , range .69-1.00) in the mediator final test cue condition.

**Final test.** Table 4 shows the proportion correctly recalled targets on the final test per condition. A 2 (learning condition: restudy vs. retrieval practice)  $\times$  2 (final test cue: mediator vs. related) between-subjects ANOVA yielded a significant main effect of learning condition,  $F(1,112)=6.679$ ,  $p=.011$ ,  $\eta^2_p=.056$ , indicating that final test performance was better for retrieved than restudied word pairs (i.e., a testing effect). There was no main effect of final test cue,  $F(1,112)=1.304$ ,  $p=.256$ ,  $\eta^2_p=.012$ , indicating that performance did not differ between mediator and related final test cues. Furthermore, there was no Learning Condition  $\times$  Final Test Cue interaction,  $F(1,112)=1.815$ ,  $p=.181$ ,  $\eta^2_p=.016$ ,  $b=.140$ , indicating that the effect of learning condition (i.e., the testing effect) did not differ between final test cue conditions.

Table 4

*Proportion of correctly recalled targets (SD) on the final test in Experiment 3 as a function of learning condition and final test cue*

Learning condition	Final test cue	
	Related	Mediator
Restudy	.319 (.223)	.308 (.275)
Retrieval practice	.383 (.247)	.512 (.361)

## Discussion

The pattern of results in Experiment 3 is similar to the one observed in Carpenter's (2011) second experiment with a larger testing effect for mediator final test cues than for related test cues. However, the crucial interaction effect between the final test cue and the testing effect is much smaller than in Carpenter's experiment (and not statistically significant). Therefore, we consider the results from Experiment 3 to be inconsistent with Carpenter's results.

**Small-scale meta-analysis.** The present study resulted in four estimates of the interaction effect between learning condition (retrieval practice vs. restudy) and final

test cue (mediator vs. related): two in Experiment 1, and one each in Experiments 2 and 3. The estimates of the interaction effect revealed a larger testing effect for mediator cues than for related cues in two cases (i.e., in the no-mediator-cue list of Experiment 1, and in Experiment 2), whereas Experiment 2 and the strong mediator cue list in Experiment 1 demonstrated a reversed pattern. Furthermore, the interaction effects appeared to be much weaker than in Carpenter's (2011) second experiment. Taken together, the four interaction effects provide substantial evidence that the testing effect does not differ between mediator final test cues and related final test cues.

To quantify our aforementioned qualitative overview of the results in our study, we performed a small-scale meta-analysis to combine the interaction effects from the present study with the one from Carpenter's (2011) second experiment. By doing so, we were able to provide a very precise estimate of the interaction between learning condition and final test cue. This procedure was inspired by Cumming's (2012) "new statistics" approach, which emphasizes the use of confidence intervals and meta-analysis. In our meta-analysis we used the *b* weights of each interaction effect (the mean difference in proportion points between the testing effect for mediator cues and the testing effect for related cues; a positive value indicates a larger testing effect for mediator cues than for related cues). For each *b* weight of the four Learning Condition x Final Test Cue interaction effects in the present study and for the Learning Condition x Final Test Cue interaction in Carpenter's (2011) second experiment, we calculated the 95% confidence intervals (CIs; see Figure 2). The squares in the forest plot represent the estimate of the interaction parameter (i.e., the *b* weight).

One way to interpret CIs is that they indicate the precision of a parameter estimate: given a particular scale of measurement, wide CIs reflect more uncertainty about the parameter of interest than narrow CIs. A look at our forest plot shows that each of five experiments separately provides a very imprecise estimate of the value of the interaction effect parameter. In addition, the values of the *b* weights vary considerably across the experiments. Consequently, the existence and size of the Learning Condition x Final Test Cue interaction effect is uncertain.

The combined effect in Figure 2 represents the 95% CI of the Learning Condition x Final Test Cue interaction effect based on combining the four interaction effects from our Experiments 1 through 3, and the interaction effect in Carpenter's (2011) Experiment 2. The point estimate in the combined effect CI was obtained by calculating the average interaction effect parameter estimates over the five comparisons by weighing each parameter estimate according to the error degrees of freedom in the general linear model. The combined standard error was calculated in

the same manner. The point estimate of the combined 95% CI shows that the testing effect is somewhat larger for mediator cues than for related cues. However, this difference is much smaller than the estimate from Carpenter's original experiment. Furthermore, the combined 95% CI is considerably narrower than the 95% CIs of the separate experiments, and therefore it provides a more precise estimate of the interaction effect parameter. Also, the combined CI includes the value of 0 indicating that the combined interaction effect is not statistically significant. Hence, when we consider the results of the five experiments together (i.e., the combined effect), there is no evidence that the testing effect differs between mediator final test cues and related final test cues.

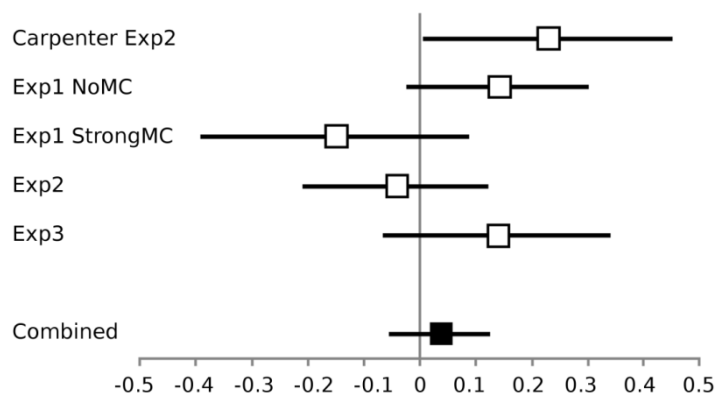


Figure 2. Forest plot of the confidence intervals of the regression  $b$  weights for the interaction effect between learning condition (retrieval practice vs. restudy) and final test cue (mediator vs. related cue) for Carpenter's (2011) Experiment 2, Experiments 1 through 3 of the present study and a combined confidence interval (based on all five interactions) of the regression beta weight for the interaction effect. A positive point estimate indicates a larger testing effect for mediator cues than for related cues. Exp1 No-MC refers to List 1 in Experiment 1, whereas Exp1 Strong-MC refers to List 2 in Experiment 1.

## General Discussion

The present series of experiments started with a question about the validity of the Learning Condition (retrieval practice vs. restudy)  $\times$  Final Test Cue (mediator vs. related) interaction effect found in Carpenter's (2011) second experiment. Carpenter proposed that her findings showed that people are more likely to retrieve semantic mediators during retrieval practice than during restudy. Consequently, the testing effect ought to be larger when people receive mediator final test cues than related final test cues that are presumably not active during retrieval practice. By contrast, we hypothesized that Carpenter's results might be due to a strong mediator-cue

association in some of her stimulus materials. To assess this hypothesis, we replicated Carpenter's design with stimulus materials in which there was no mediator-cue association and with stimulus materials in which there was a strong mediator-cue association. Reasoning from our hypothesis, we expected a larger testing effect for mediator cues compared to related cues only in the strong mediator-cue association condition but not in the no mediator-cue association condition. According to Carpenter's semantic mediator account, the testing effect should be larger for mediator cues than for related cues regardless of the mediator-cue association.

The results of Experiment 1 were not in line with our hypothesis, but we also failed to demonstrate convincing evidence for the semantic mediator account. Intrigued by the latter finding, we conducted two replications (i.e., Experiments 2 and 3) of Carpenter's (2011) second experiment. Consistent with the findings from Experiment 1, the findings from Experiment 2 and 3 indicated there is not enough empirical evidence to conclude that the magnitude of the testing effect is different for mediator final test cues than for related final test cues. Subsequently, we combined the results of all current experiments and those of Carpenter's Experiment 2 in a small-scale meta-analysis. This analysis (which was based on the data of 468 participants) showed that the testing effect might be somewhat larger for mediator cues than for related cues. However, the size of the interaction effect was much smaller than in Carpenter's original experiment. Also, the CI of this combined effect, which was more precise than the CI of the separate experiments, showed that 0 cannot be excluded as the parameter value of the interaction effect between learning condition and final test cue. All in all, the results of the present study indicate that there is no empirical ground to assume that the testing effect differs between mediator cues and test cues. Hence our results fail to support an important assumption of the semantic mediator account, namely that retrieval practice strengthens a mediator-target connection more than restudying.

A difference between the experiments in the present study and Carpenter's (2011) experiment was that Carpenter tested undergraduate students in the lab, whereas we tested Mechanical Turk workers online. One could note that this difference might underlie the inconsistency between our results and those reported by Carpenter. However, we think there are a number of strong arguments against such a point of critique. First, there is nothing in the semantic mediator hypothesis suggesting that the interaction effect between learning condition and final test cue is restricted to undergraduate students tested in the lab. That is, according to the semantic mediator hypothesis Carpenter's results should generalize to Mechanical Turk workers tested online. Second, on the basis of the available empirical data it is

actually very hard to compare the interaction effect in Mechanical Turk studies with the interaction effect in lab studies conducted with the traditional undergraduate participants. This is because Carpenter's experiment is the only lab study we are aware of reporting the interaction between learning condition and final test cue. Moreover, considering the width of the CI of the interaction effect, it is clear that this particular experiment does not provide a precise estimate of the effect. Hence, if we want to compare the crucial interaction effects between Mechanical Turk studies and traditional undergraduate studies, a much more precise estimate of the interaction effect in lab studies is needed.

The current results fit with other empirical evidence that challenges the semantic mediator hypothesis. For instance, Karpicke and Smith (2012) compared two elaborative study strategies (an imagery-based keyword method and verbal elaboration) with repeated retrieval practice. After successful retrieval, repeated retrieval practice enhanced retention, but elaborative restudying did not produce additional learning. This suggests that elaboration is not the key mechanism in the emergence of the testing effect. Furthermore, when the to-be-learned stimuli included not only unrelated word pairs (e.g., *mountain – hammer*) but also identical word pairs (e.g., *castle – castle*) that required no elaboration, there was still a testing effect for the identical word pairs. The authors conclude from these results that the testing effect is not likely to be caused by elaboration. This conclusion resonates with the results of the current study.

In reaction to empirical evidence that seems to be inconsistent with the semantic mediator account, researchers have proposed a different explanation as to why retrieval practice produces a better performance on a memory test than restudying. For instance, Karpicke and Zaromb (2010) suggested that retrieval practice enhances memory by improving the diagnostic value of retrieval cues. According to this view, it is not the elaboration of memory traces that strengthens memory during retrieval practice, but enhanced discrimination of possible targets. Retrieval practice could enhance cue diagnosticity by for instance reducing the match between the cue and competitor targets, or by reducing the number of possible targets (Karpicke & Smith, 2012). In contrast with the activation of mediators, cue diagnosticity is not dependent on elaborative processes and therefore does not conflict with the findings in the current paper. However, as the current paper does not provide direct evidence in favor of the importance of cue diagnosticity this hypothesis could be investigated in future work.

We think that the findings from the present study are of relevance because they have an important implication for theories on the testing effect. However, in our view

the study is also interesting because it can be considered as an example of new research approaches that are likely to improve the quality of psychological research. First, the present study can be seen as a series of conceptual (Experiment 1) and exact (Experiments 2 and 3) replications of Carpenter's (2011) original experiment. Recently the issue of replication of results from psychological research has received a lot of attention. For instance, a special section in a recent issue of *Perspectives on Psychological Science* (Pashler & Wagenmakers, 2012) was devoted to replicability in psychological science. According to several authors in this issue not enough replications of psychological studies are conducted or published (Koole & Lakens, 2012; Makel, Plucker, & Hegarty, 2012). Replication attempts are useful to separate reliable from unreliable findings and allow researchers to optimize their research methods by providing information about effect sizes (Koole & Lakens, 2012). The current paper does exactly that: further assess earlier findings and help correctly estimate the size of the theoretically relevant interaction between learning condition and final test cue type. Thus, the current paper demonstrates the value of replication attempts.

Second, the small-scale meta-analysis demonstrates the virtues of the new statistics approach (Cumming, 2012). For one, presenting CIs for parameter estimates of interest makes it clear how much uncertainty is involved with a single study. When small samples are used there is a very high degree of uncertainty. This is nicely illustrated by the CI of Carpenter's (2011) original Learning Condition x Final Test Cue interaction effect, which has a lower bound of .006 and an upper bound of .456. Because the CI does not include 0, the interaction effect is statistically significant. However, the fact that the interaction effect is significant is not really informative considering the imprecision – as reflected in the wide CI - of the parameter estimation. Cumming argues convincingly that psychologists should shift their focus from asking whether there is a significant effect to what the size of the effect is and how accurate the estimate of the effect size is. CIs provide an answer to the second question. In addition, one way to obtain more accurate estimates of an effect of interest is to combine the results of existing studies in a meta-analytic fashion. The small-scale meta-analysis in the present study is an illustration of how the new statistics can be brought into practice.

In conclusion, the results of the present study are not in line with the semantic mediator account of the testing effect. Therefore, future research is needed to clarify the cognitive processes that underlie the testing effect. Furthermore, the present study is an example of how replication research and the new statistics approach can be implemented in psychological research.



## Appendix A: Stimuli used in Experiment 1

## No mediator-cue association

Cue	Target	Mediator	Related	C-T	C-M	M-C	R-T
BLACKBOARD	CLASS	CHALK	BORED	0,014	0,676	0,000	0,048
RACQUET	SPORT	BALL	COACH	0,020	0,689	0,000	0,047
ARCHITECTURE	DESIGN	BUILDING	DECORATION	0,027	0,510	0,000	0,041
MARE	NIGHT	HORSE	FLASHLIGHT	0,021	0,740	0,000	0,041
ANATOMY	SCIENCE	BODY	GEOLOGY	0,041	0,607	0,000	0,047
SAP	STICKY	TREE	GOO	0,027	0,703	0,000	0,046
PUBLISHER	NEWSPAPER	BOOK	HOROSCOPE	0,020	0,533	0,000	0,035
HERD	GROUP	COW	PEER	0,021	0,562	0,000	0,039
PERCH	STAND	BIRD	POSITION	0,020	0,547	0,000	0,045
OAR	MAN	BOAT	POST	0,014	0,695	0,000	0,041
BUDGET	PLAN	MONEY	PROCEDURE	0,021	0,541	0,000	0,031
LUMBER	YARD	WOOD	RAKE	0,040	0,596	0,000	0,041
CALORIES	BURN	FAT	ROPE	0,040	0,527	0,000	0,039
CORK	STOPPER	WINE	RUBBER	0,020	0,517	0,000	0,014
SKUNK	STRIPE	SMELL	SOLID	0,016	0,559	0,000	0,028
CRADLE	ROCK	BABY	SWAY	0,048	0,678	0,000	0,054

## Strong mediator-cue association

Cue	Target	Mediator	Related	C-T	C-M	M-C	R-T
WEST	WILD	EAST	ADVENTUROUS	0,031	0,780	0,886	0,049
DOG	FRIEND	CAT	ADVICE	0,019	0,667	0,513	0,036
MOTHER	CHILD	FATHER	BIRTH	0,010	0,597	0,706	0,015
NIGHT	MOON	DAY	GRAVITY	0,019	0,686	0,819	0,042
ANSWER	RIGHT	QUESTION	INCORRECT	0,040	0,540	0,767	0,040
QUEEN	BEE	KING	INSECT	0,041	0,730	0,772	0,039
BOTTOM	BARREL	TOP	KEG	0,014	0,507	0,696	0,030
NOUN	THING	VERB	MATERIAL	0,016	0,690	0,642	0,041
FRONT	FACE	BACK	MIRROR	0,014	0,520	0,715	0,047
SUPPER	TIME	DINNER	PLACE	0,049	0,545	0,535	0,035
HAMMER	SAW	NAIL	SANDPAPER	0,028	0,800	0,622	0,021
PEPPER	SNEEZE	SALT	SNIFF	0,041	0,695	0,701	0,026
TODAY	SHOW	TOMORROW	STAGE	0,013	0,503	0,527	0,047
LEG	WALK	ARM	TROT	0,036	0,503	0,673	0,048
LOSER	SORE	WINNER	ULCER	0,030	0,508	0,600	0,040
VOLCANO	MOUNTAIN	ERUPT	WATERFALL	0,022	0,525	0,641	0,047

Note. C-T indicates cue-to-target association strength, C-M indicates cue-to-mediator association strength, M-C indicates mediator-to-cue association strength, and R-T indicates related-to-target association strength.

Mediator-to-target association strength and related-to-cue association strength was always 0.

## Appendix B: Stimuli used in Experiments 2 and 3

Cue	Target	Mediator	Related	C-T	C-M	M-C	R-T
WEAPON	KNIFE	GUN	AX	0,075	0,592	0,024	0,046
COFFEE	TABLE	TEA	BANQUET	0,020	0,442	0,369	0,020
MOTHER	CHILD	FATHER	BIRTH	0,010	0,597	0,706	0,015
SOIL	EARTH	DIRT	CONTINENT	0,040	0,717	0,055	0,041
SONNET	MUSIC	POEM	DANCER	0,059	0,471	0,020	0,052
SEA	RIVER	OCEAN	FLOOD	0,017	0,456	0,291	0,020
EMPLOYMENT	OFFICE	JOB	GOVERNMENT	0,020	0,605	0,016	0,024
JACKET	SHIRT	COAT	HANGER	0,013	0,564	0,176	0,014
PRESCRIPTION	DOCTOR	DRUG	HOSPITAL	0,034	0,477	0,020	0,027
TRASH	PAPER	GARBAGE	INK	0,013	0,526	0,456	0,013
DONOR	HEART	BLOOD	LIVER	0,042	0,524	0,067	0,041
DUSK	EVENING	DAWN	MORNING	0,042	0,609	0,454	0,047
BREEZE	SUMMER	WIND	MOSQUITO	0,012	0,606	0,122	0,014
PEDESTRIAN	STREET	WALK	NEIGHBORHOOD	0,032	0,597	0,000	0,034
FRAME	WINDOW	PICTURE	SHINGLE	0,014	0,811	0,316	0,014
VOCABULARY	SCHOOL	WORDS	TEXT	0,013	0,507	0,034	0,013

*Note.* C-T indicates cue-to-target association strength, C-M indicates cue-to-mediator association strength, M-C indicates mediator-to-cue association strength, and R-T indicates related-to-target association strength. Mediator-to-target association strength and related-to-cue association strength was always 0.

# 6

## Summary and Discussion

The testing effect entails that when information is studied and subsequently tested, it is better remembered in the long term than information that is only restudied (Toppino & Cohen, 2009). This is true even when the total study time is kept equal. The testing effect is well established empirically, but its underlying cognitive mechanisms are less clear. Therefore, the studies in this dissertation were set up to investigate the underlying mechanisms that contribute to the emergence of the testing effect. More specifically, the studies in this dissertation assessed the idea that elaborative processing underlies the testing effect.

According to the elaborative processing hypothesis of the testing effect, testing causes more elaborative processing than restudying. This increased elaborative processing in turn causes more semantically related information to be activated and coupled with the to-be-learned information. At a final test, the related information that was activated during learning serves as extra retrieval cues. Therefore, the learned information is easier to retrieve after testing than after restudying. The present dissertation contributes to research on this explanation of the testing effect by testing previously untested predictions of the elaborative processing hypothesis using cue-target pairs and cued recall final tests.

### **Summary of the main findings**

#### **Is elaboration necessary?**

In Chapter 2 we tested whether the testing effect is also present when the to-be-learned materials consist of symbol-word pairs. Participants learned Adinkra symbols paired with nouns. The symbol-word pairs were either studied four times or studied once and tested three times. In the final test after five minutes or seven days, the symbols were shown and participants had to recall the word paired with each symbol. Results showed a clear testing effect: after a short retention interval of five minutes, there was no difference in the mean number of correctly retrieved symbol-word pairs between restudying and testing. After a long retention interval of seven days, there was a difference between conditions: tested symbol-word pairs were more often correctly recalled than restudied symbol-word pairs. This study therefore showed that the testing effect also occurs when the to-be-learned materials are symbol-word pairs.

Although the primary purpose of the study in Chapter 2 was to generalize the testing effect to learning symbol-word pairs, the results also provide information on the elaborative processing hypothesis. The Adinkra symbols used as cues in this study were non-semantic cues and according to a norming study they were difficult to

verbally describe. According to elaborative processing theories of the testing effect, elaboration involves the activation of semantically related information (Carpenter, 2011). In order to activate semantically related information and thus elaborate on the learned materials, a verbal description of the to-be-learned materials must be available. This was not the case for the Adinkra symbols we used in this study; the symbols were difficult to verbalize and therefore it was presumably very difficult for participants to elaborate on the present materials. Still, we found a testing effect. Therefore the findings of the study in Chapter 2 appear to be not in line with theories of the testing effect that rely on semantic elaboration, such as the mediator effectiveness hypothesis. This result led us to investigate the role of elaboration in the emergence of the testing effect in word pairs.

### **Does more elaboration occur during testing than during restudying?**

In the study in Chapter 3 we started to investigate the elaborative processing hypothesis of the testing effect in more detail, by looking at cognitive processes that occur during restudying and testing. We had participants restudy or retrieve word pairs and used a novel approach to investigate cognitive processes during learning; we recorded electroencephalograms (EEG) and analyzed event-related potentials (ERPs). ERPs can provide valuable information about the cognitive processes that occur during learning and therefore elucidate the processes that underlie the testing effect. Although some researchers (e.g., Eriksson, Kalpouzos, & Nybert, 2011) have investigated the testing effect using functional magnetic resonance imaging (fMRI) and one study used oscillatory EEG to investigate the effects of retrieval on subsequent encoding (i.e., Pastötter, Schicker, Niedernhuber, & Bäuml, 2011), to our knowledge ours was the first ERP study into the testing effect.

According to the dual process account of memory (Curran, Tepe, & Piatt, 2006; Yonelinas, 2002), two distinct processes are associated with retrieval: familiarity occurs when a word is recognized based on a 'feeling of knowing' without remembering specific details of the occurrence, while recollection occurs when the specific details of a previous presentation of a word are recalled. Based on the elaborative processing hypothesis, we hypothesized that during retrieval trials familiarity should play a larger role than during restudy trials, because familiarity is focused on the global meaning of a word. By contrast, recollection is focused on physical or contextual details that are presumably less important in order to activate semantically related words. Therefore we predicted that the familiarity-indicating N400 repetition effect should be more pronounced during test trials than during restudy trials. This would suggest that during test trials, processes are more focused on the meaning of the items than during restudy trials. In regard to the recollection-

indicating late positive component (LPC) repetition effect, we predicted that the LPC amplitude would be smaller during testing than during restudying. This would suggest less detail-focused processing during test trials. Taken together, these differences in N400 and LPC repetition effects would indicate a focus on meaning and not on details, which would fit with the elaborative processing hypothesis. Furthermore, we hypothesized that more effort would be allocated to test trials than to restudy trials because of increased elaborative processing, resulting in a larger P300 amplitude.

First of all, a clear testing effect occurred: tested words were more often correctly recalled than restudied words on the final test after two days. The ERP data recorded during learning showed a larger P300 amplitude during testing than during restudying. This suggests that more effort was put into learning the word pairs through testing than through restudying. The differences in the N400 and LPC repetition effects that we predicted were also present. The N400 repetition effect was larger for testing than for restudying and the LPC repetition effect was smaller for testing than for restudying. Combined, these results suggest that during testing more effort is used to learn the word pairs and that this increased effort is at least partly caused by the retrieval of semantic information during learning. Thus, it seems that more semantically related information is activated during testing than during restudying. Therefore, the results of the study in Chapter 3 appear to be consistent with the elaborative processing hypothesis of the testing effect.

### **Does elaboration involve the activation of mediators?**

Taken together, the studies in Chapters 2 and 3 reveal a mixed picture. In the study in Chapter 4 we started to investigate a somewhat more specific version of the elaborative processing hypothesis: the mediator effectiveness hypothesis (Pyc & Rawson, 2010, 2012). According to the mediator effectiveness hypothesis, during testing mediators are activated. A mediator is a word that is strongly associated with the cue and can be easily coupled with the target. To test the activation of mediators we had participants learn pairs of words that were each related to a homonym in the study in Chapter 4. For instance, participants learned the word pair *teller* – *account*. Both *teller* and *account* are related to the homonym *bank* (a financial institution). The word pairs were studied and either restudied twice or tested twice. In a supposedly unrelated second experiment, participants read sentences that included the homonym in the opposite meaning. For instance, after learning the word pair *teller* – *account* (suggesting the financial institution) the sentence was *he lay down in the grass by the bank* (suggesting the alternative meaning of *bank*: a river side). During reading of these sentences, we recorded event-related brain potentials (ERPs). We hypothesized that the homonym *bank* would be more often activated during learning

of *teller – account* through testing than through restudying. Therefore, we predicted that reading the word bank in the opposite meaning after the learning phase would more often result in a conflict between the two meanings of the homonym after testing than after restudying.

Since the N400 is the most commonly used ERP measure of semantic integration, we expected that the amplitude of the N400 would be higher during reading of homonyms of which the related words had been tested than during reading of homonyms of which the related words had been restudied. However, we found no difference in N400 amplitude between the two types of studying. Therefore, we found no evidence of increased elaboration during testing compared to restudying. Thus the results of this experiment suggest that the role of elaboration in the emergence of the testing effect is limited at best.

In the study in Chapter 5, we again set out to validate the mediator effectiveness hypothesis. We did a series of experiments via Mechanical Turk. The original goal of the study was to investigate a peculiarity in the materials used in the experiment of Carpenter (2011, Experiment 2). She investigated the mediator effectiveness hypothesis by having participants learn related word pairs such as *mother – child* through testing or restudying. At the final test, participants received one of three types of cues: (1) The original cue (*mother*), (2) a word weakly associated with the target (*birth*), (3) a word strongly associated with the cue, but with no association with the target (i.e., a mediator, *father*). Her results showed a testing effect when the final test cue was an original cue. Moreover, the testing effect was larger when the final test cue was a mediator than when the final test cue was a word associated with the target. This suggests that during testing, the mediator is activated and coupled with the target, resulting in an association between the mediator and the target.

However, looking at the materials we noticed that some of the mediators had a strong forward association with the original cue. For instance, the mediator *father* has a strong association with the original cue *mother*. In such a case, when given the mediator *father*, the original cue *mother* is immediately activated. Because testing strengthens the cue-target association more than restudying, it will be easier to retrieve the target from the original cue after testing than after restudying. Therefore, when there is a strong association between the mediator and the original cue, this can explain the emergence of the testing effect without assuming mediator activation during studying. To test this alternative explanation of Carpenter's (2011) results, we constructed two lists of stimuli. In one list of word sets, there was no mediator-cue association. In the other list, there was a strong mediator-cue association. We predicted that the first list would not show a larger testing effect for mediators than

for related words. For the second list, we predicted the same effect Carpenter found: a larger testing effect for mediator cues than for related cues.

In contrast to our predictions, we found no interaction between learning condition and mediator-cue association. Apparently, the results of Carpenter were not caused by mediator-cue associations. In addition, and surprisingly, we did not confirm the original finding of Carpenter because there was no difference in the magnitude of the testing effect between mediator and related final test cues. This finding led us to attempt two replications of Carpenter's (2011) study. We analyzed the results of our three studies and those of Carpenter's Experiment 2 together in a small-scale meta-analysis. This analysis was inspired by the 'new statistics' approach (Cumming, 2012). The meta-analysis allowed us to combine all the findings in one clear overview. In addition to providing an overview of the results, the confidence intervals used in the meta-analysis also clarified the precision of the findings from the individual studies and the analysis gave a much more precise (relative to the individual studies) combined estimate of the crucial interaction effect. The analysis confirmed that there was a testing effect in all experiments. However, the magnitude of the testing effect did not differ systematically between related and mediator final test cues. These results are not in line with the mediator effectiveness hypothesis; they suggest that mediators are not activated more often during testing than during restudying.

## Discussion

### Reflection on the conducted studies

In the studies in Chapters 3 and 4, we analyzed ERPs. This technique had not been used before in testing effect research. ERPs can provide valuable information about the cognitive processes that occur during learning. However, there are some drawbacks to using this technique. First, it can be difficult to interpret results, especially when the measured ERP components do not have a single commonly accepted underlying mechanism. In the study in Chapter 3, we interpreted a larger N400 repetition effect as an indication of familiarity processing and a larger LPC repetition effect as evidence for recollection. However, the N400 repetition effect is not only ascribed to familiarity but also to superficial processing of stimuli, and the LPC repetition effect is sometimes attributed to semantic processing (Juottonen, Revonsuo, & Lang, 1996). We interpreted a larger N400 repetition effect as an indication of familiarity processing and a larger LPC repetition effect as evidence for recollection, based on extensive literature on the topic (e.g., Curran, 2000; Curran & Cleary, 2003; Mecklinger, 2000; Rugg, 1995; Rugg, Schloerscheidt, & Mark, 1998).



Therefore, although we could have interpreted our results differently, the choice we made was based on the literature on ERPs associated with familiarity and recollection.

A second drawback to using the ERP technique is that a null effect (i.e., no difference in amplitude between conditions) can be attributed not only to an actual null effect (i.e., no difference between conditions) but also to a variety of other factors. For instance, a null effect can be caused by excessive noise in the EEG data, by a low number of trials used in an average, or even by certain filter settings. Of course, we were aware of the latter factor and thus the null effect in the study in Chapter 4 cannot be attributed to filter settings. However, the other two factors cannot be ruled out. In the study in Chapter 4 we only used fifteen homonyms per condition because of the limited number of homonyms available and because we wanted to make sure the participants remembered most of the word pairs. A low number of trials can cause the ERP average to be unreliable, resulting in an inability to find an effect. However, N400 modulations have been found with fewer than 20 trials per condition (e.g., Arzouan, Goldstein, & Faust, 2007; Franklin, Dien, Neely, Huber, & Waterson, 2007) and we used particularly low-noise apparatus, which makes the average more reliable. Therefore, we do not think the low number of trials caused our null effect.

Another property of the study in Chapter 4 is that there was quite some time (about 15 minutes) between learning and the sentence-reading task. This could be a problem because the effects of semantic priming (which are presumably caused by the same mechanism that underlies elaborative processing) have been shown to be short-lived (e.g., Bentin & Feldman, 1990; Masson, 1995; Meade, Watson, Balota, & Roediger, 2007). For instance, Zeelenberg and Pecher (2002) presented as much as twelve primes related to one target and found no evidence of long-term priming (i.e., after an interval longer than 10 minutes). One could therefore argue that the homonyms in our study were no longer active during the reading task and that is the reason we did not find evidence for a semantic conflict upon encountering the homonym in the sentence reading task. However, this argument appears to be at odds with elaborative processing explanations of the testing effect. This is because testing effects are often found after long retention intervals of hours or even days; retention intervals that are much longer than in the study of Chapter 4. Hence, if the testing effect is caused by the activation of semantically related information this activation should also last several hours or days. According to this line of reasoning, the N400 conflict should have been stronger after testing than after restudying in the study of Chapter 4.

In the study in Chapter 5 we attempted two replications of the study of Carpenter (2011, Experiment 2) and could not replicate the original finding. We

interpreted the results as evidence against the mediator effectiveness hypothesis. A critical reader could note that these two studies were not exact replications of Carpenter's study: there were some differences between the studies. Specifically, we tested participants via internet and consequently had less control over the experimental circumstances. Although we asked participants to perform the task in a place where they could concentrate and had them answer questions about their surroundings, we can never be sure whether the participants were not distracted. In addition, we tested US residents whereas Carpenter tested undergraduate students. Therefore our sample was more diverse than Carpenter's. In particular the age and educational level of participants could play a role in the outcomes. However, we think these differences between Carpenter's and our studies do not pose a threat to our conclusion. First of all, the influence of the differences in samples and test conditions on the outcomes seems limited. In a lab setting it is also possible that participants are distracted during the experiment and the overall final test performance we found was similar to that of Carpenter (2011), suggesting that the samples are comparable. Second, and most importantly, the replications were not intended as direct replications, but as conceptual replications in order to validate the mediator effectiveness hypothesis. The mediator effectiveness hypothesis does not include a restriction to experiments with undergraduate students tested in the lab and therefore the theory gives no reason to assume a different effect with a more diverse sample tested via internet. Therefore we think that our replications provided a valid test of the mediator effectiveness hypothesis and thus that the results provided no evidence for the activation of mediators during testing.

A general property of the studies in the current dissertation that we should note is that we only investigated the testing effect in learning cue-target pairs with a cued recall test as an intervening test. Using these specific materials and learning methods allowed us to build on previous studies in order to formulate new hypotheses. However, it also limits the reach of our conclusions to cue-target learning. For instance, the current dissertation does not provide information on the role of elaboration in learning texts.

### **Theoretical implications**

Overall, the results of the studies in the current dissertation suggest that although there is some evidence that more elaboration takes place during testing than during restudying, this elaboration does not involve the activation of mediators. In Chapter 2, we found a testing effect using materials that did not allow elaboration. This suggests that elaboration is not necessary in order to find a testing effect and therefore that elaborative processing cannot be the only mechanism that underlies

the testing effect in cued recall. Still, because the study in Chapter 2 only used materials that do not allow elaboration, it does not exclude elaboration in learning materials that do allow for elaboration. Indeed, in the study in Chapter 3 using word pairs as to-be-learned materials we found ERP evidence for enhanced semantic processing during testing. This finding is in line with theories that rely on elaboration, such as the mediator effectiveness hypothesis (Pyc & Rawson, 2010). Subsequently, we set out to investigate the mediator effectiveness hypothesis more specifically. In the event-related potential study in Chapter 4 and the Mechanical Turk study in Chapter 5, we found no evidence for the activation of mediators during testing. Since an important assumption of the mediator effectiveness hypothesis is that mediators are activated during testing, these results are not in line with the mediator effectiveness hypothesis. In sum, the studies reported in the current dissertation provide only limited support for the elaborative processing hypothesis of the testing effect.

There is empirical evidence in favor of the elaborative processing hypothesis (Carpenter, 2011; Carpenter & DeLosh, 2006; Pyc & Rawson, 2010, 2012), but the evidence is not conclusive. For instance, Pyc and Rawson (2010, 2012) gave feedback after testing, which makes it difficult to differentiate the effect of testing from the effect of feedback. Carpenter and DeLosh (2006) found a larger testing effect after intervening testing with less informative cues. This finding supports the elaborative processing hypothesis, but only when one assumes that a less informative cue elicits more elaboration. Therefore, this evidence is indirect. Carpenter (2011) provided the only direct evidence for the elaborative processing hypothesis; she found more false memories (i.e., falsely recognized words at a final recognition test) for cue-related words after testing than after restudying. These results suggest that words related to the cue are activated during testing and therefore provide support for the elaborative processing hypothesis. However, the findings can not only be explained by proposing that the related words are activated during learning. According to global matching models of false memory (e.g., Hintzman, 2001), false memories emerge during the final test (as opposed to during learning) because the memory traces of learned words are highly similar to those of the related words that are presented. If the match between the learned word and the word presented at the final test is strong enough, the word is incorrectly recognized. Thus, according to this account, false memories emerge because of a matching process during the final test instead of through activation of related words during learning. The global-matching model can therefore explain Carpenter's results without supposing that testing enhances elaboration.

Thus, the evidence in favor of the elaborative processing account of the testing effect is not conclusive. Recently some evidence against it has been published. For instance, Karpicke and Blunt (2011) demonstrated that testing produced more learning than drawing a concept map. Participants learned science texts through testing or elaborative concept mapping. Final test performance was better after testing than after elaborative concept mapping, even when the final test involved drawing a concept map. As drawing a concept map is a highly elaborative study strategy, this result shows that testing is more effective than elaborative studying and, therefore, that elaboration is not likely to be the key factor in the emergence of the testing effect. Moreover, Karpicke and Smith (2012) compared testing with repeated studying using two elaborative encoding strategies: an imagery-based keyword strategy and a verbal elaboration method. The authors predicted that if the testing effect is indeed caused by elaboration, the two elaborative encoding strategies would have the same effect on learning as testing. However, testing led to better final test performance than elaborative restudying. Karpicke and Smith interpret these results as evidence against the elaborative processing hypothesis.

A problem with the studies of Karpicke and Blunt (2011) and Karpicke and Smith (2012, Experiments 1-3) is that one could argue that the elaboration that occurs during the elaborative study strategies used in the experiments is not the same as the elaboration that might occur during testing. It is possible that although for instance an imagery-based keyword strategy is an elaborative study strategy, the elaboration during studying with this method is not as extensive as the elaboration during testing. However, in their Experiment 4, Karpicke and Blunt (2011) found more conclusive evidence against the elaborative processing hypothesis. When the to-be-learned stimuli included not only unrelated word pairs (e.g., *dog* – *chair*) but also identical word pairs (e.g., *hammer* – *hammer*) that required no elaboration, the identical word pairs still showed a testing effect. Thus, there was a testing effect when the materials did not allow for elaboration. The authors conclude from these results that the testing effect is not likely to be caused by elaboration.

All in all, the studies in the current dissertation and the studies of Karpicke and Blunt (2011) and Karpicke and Smith (2012) discussed above do not provide evidence in favor of the elaborative processing hypothesis. There is no evidence for the activation of mediators during testing (Chapters 4 and 5 of this dissertation), and testing produces more learning than elaborative restudying using various restudying techniques (Karpicke & Blunt, 2011; Karpicke & Smith, 2012). Alternative explanations of the testing effect that do fit with the results of the current dissertation have been proposed. Karpicke and Zaromb (2010) suggested that testing improves memory by

enhancing the diagnostic value of retrieval cues. According to this view, it is not elaboration that strengthens memory during testing, but enhanced discrimination of possible targets. Testing could enhance cue diagnosticity by reducing the match between the cue and competitor targets, which makes it easier to ignore competitors and find the target. Alternatively, testing could reduce the number of possible targets (Karpicke & Smith, 2012), which also makes it easier to recall the correct target.

Cue diagnosticity can also account for the interaction between learning condition and retention interval that is often found, by incorporating the bifurcation model (Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011). According to this model, an item can be retrieved (either on an intervening test or on a final test) when the memory strength of that item is above a certain threshold. All items are learned through an initial learning trial, which is assumed to result in a normal distribution of items over memory strength. Subsequent testing without feedback results in strengthening of a subset of the learned materials, because only a part of the initially learned materials is above the retrieval threshold and thus retrieved on the intervening tests. The distribution of items after testing is therefore bifurcated: the items that were not successfully retrieved on the intervening test are not strengthened, whereas the items that were retrieved on the intervening test are strengthened a lot. By contrast, restudying after the initial study trial strengthens the entire set of learned materials, resulting in a small (smaller than for successfully retrieved items) upward shift of the entire distribution of restudied items. When the final test is administered directly after learning, more restudied than tested items are above the retrieval threshold. This is because although items that are successfully retrieved during learning are strengthened more than restudied items, not all tested items are successfully retrieved. However, after a retention interval in which the memory strength of all items declines, more tested than restudied items are above the retrieval threshold. In this way, the bifurcation model can account for the interaction of the testing effect with retention interval.

### **Suggestions for future research**

The cue diagnosticity account of the testing effect is not dependent on the activation of mediators or other elaborative processes during testing and therefore does not conflict with the findings in the current dissertation. However, the current dissertation does not provide direct evidence in favor of the importance of cue diagnosticity. Moreover, to our knowledge there is no direct evidence in favor of the cue diagnosticity hypothesis of the testing effect. To investigate this hypothesis, future work could investigate the activation of related words during the final test. Testing could enhance cue diagnosticity by reducing the number of competing

targets. A probable competing target is an item that is highly related or similar to the target (for instance with nonsense syllables, for the syllable pair *pud* - *lor*, a competing target could be *jor*). If cue diagnosticity causes the testing effect by reducing the number of competing targets, this effect should be measurable on a final test. In the case of the nonsense syllables, in the final test the cue *pud* could be given with four target alternatives, including the target *lor*, two non-similar distractors such as *pib* and *dar*, and the syllable *jor* that is highly similar to the target. If cue diagnosticity is indeed enhanced after testing through reducing the number of competing targets, then errors on the final test would more often involve choosing the highly similar syllables (as opposed to the non-similar syllables) after restudying than after testing.

In addition, to investigate the bifurcation model future research could look at the effect of the level of difficulty of the final test on the testing effect. The bifurcation model predicts that if the final test is difficult enough and thus the retrieval threshold is high enough, a short-term testing effect will emerge. If this prediction is confirmed, this would be evidence in favor of the bifurcation model. Indeed, this finding has been reported by Halamish and Bjork (2011, Experiments 1 and 2). They found a short-term testing effect when the final test involved free recall, which is a difficult test. When the final test involved cued recall, which is easier than free recall, there was an advantage of restudying. In addition, Verkoeijen, Bouwmeester, and Camp (2012) had participants learn word lists in Dutch and take a short-term final test either in Dutch or in English. This language manipulation increases the difficulty of the final test; a final test in the same language as the learning phase is easier than a final test in a different language. The authors found a short-term testing effect when the final recognition test was in a different language than the learning phase, but a restudying advantage when the final test was in the same language as the learning phase. These results can also be interpreted as evidence for the bifurcation model. Other manipulations of final test difficulty could include manipulating the retrieval threshold itself, for instance by giving some sort of punishment for incorrect responses, or using highly similar distractors at a recognition test.

## Conclusion

Although we did find evidence for enhanced elaborative processing during testing compared to restudying (Chapter 3), we also found a testing effect when the learned materials did not allow elaboration (Chapter 2). In addition, we did not find any evidence of the activation of mediators during testing (Chapters 4 and 5). Therefore, the results of the current dissertation do not support the mediator

effectiveness hypothesis. The testing effect does not seem to be caused by the activation of mediators during testing. Alternative explanations of the testing effect have been offered and should be investigated in future research.





# Samenvatting

Dutch summary

Het testeffect houdt in dat informatie beter wordt onthouden op de lange termijn wanneer de informatie is bestudeerd en daarna getest, dan wanneer het alleen is bestudeerd (Toppino & Cohen, 2009). Dit effect treedt ook op wanneer de totale studietijd gelijk wordt gehouden. Het testeffect is een robuuste empirische bevinding, maar de onderliggende cognitieve mechanismen zijn minder duidelijk. Daarom zijn de studies in deze dissertatie opgezet om de onderliggende mechanismen die bijdragen aan het ontstaan van het testeffect te onderzoeken. De studies in deze dissertatie onderzochten het idee dat elaboratieve processen ten grondslag liggen aan het testeffect.

Volgens de elaboratieve verwerking-hypothese van het testeffect veroorzaakt testen meer elaboratieve verwerking dan herstuderen. Deze verhoogde elaboratieve verwerking zorgt ervoor dat er meer semantisch gerelateerde informatie wordt geactiveerd en gekoppeld met de informatie die geleerd moet worden. Tijdens de eindtest dient de gerelateerde informatie die is geactiveerd tijdens het leren als extra ophaalcues. Daardoor is de geleerde informatie makkelijker op te halen na testen dan na herstuderen. Deze dissertatie draagt bij aan onderzoek naar deze verklaring van het testeffect door tot nu toe niet-geteste voorspellingen van de elaboratieve verwerking-hypothese te toetsen. Hierbij wordt gebruik gemaakt van *cue-target* paren en *cued recall* eindtesten.

## Samenvatting van de belangrijkste bevindingen

### Is elaboratie nodig?

In Hoofdstuk 2 testten we of het testeffect ook optreedt als het leermateriaal bestaat uit symbool-woordparen. Deelnemers leerden paren van Adinkrasymbolen en zelfstandig naamwoorden. De symbool-woordparen werden vier keer bestudeerd of een keer bestudeerd en drie keer getest. In de eindtest na vijf minuten of zeven dagen werden de symbolen getoond en de deelnemers moesten het woord ophalen dat bij het symbool hoorde. Er trad een duidelijk testeffect op: na het korte retentie-interval van vijf minuten was er geen verschil tussen herstuderen en testen in het gemiddelde aantal correct opgehaalde woorden. Na het lange retentie-interval van zeven dagen was er echter wel een verschil: geteste woorden werden vaker correct opgehaald dan herbestudeerde woorden. Deze studie toonde daarom aan dat het testeffect ook optreedt wanneer de leermaterialen bestaan uit symbool-woordparen.

Hoewel het hoofddoel van de studie in Hoofdstuk 2 was om te onderzoeken of het testeffect ook optreedt bij symbool-woordparen, geven de resultaten ook informatie over de elaboratieve verwerking-hypothese. De Adinkrasymbolen die we

gebruikten in deze studie waren niet-semantiche cues en een normatieve studie toonde aan dat ze moeilijk verbaal te beschrijven waren. Volgens de elaboratieve verwerking-hypothese is elaboratie de activatie van semantisch gerelateerde informatie (Carpenter, 2011). Om semantisch gerelateerde informatie te activeren en dus te elaboreren moet een verbale omschrijving van de leermaterialen beschikbaar zijn. Dit was niet het geval bij de Adinkrasymbolen die we gebruikten in de studie in Hoofdstuk 2; de symbolen waren moeilijk te verbaliseren en dus was het waarschijnlijk moeilijk om te elaboreren tijdens het leren van de symbool-woordparen. Toch vonden we een testeffect. Daarom lijken de resultaten van de studie in Hoofdstuk 2 niet in lijn met theorieën van het testeffect die uitgaan van elaboratie, zoals de *mediator effectiveness*-hypothese. Deze uitkomst leidde ons ertoe de rol van elaboratie in het ontstaan van het testeffect in woordparen te onderzoeken.

### **Vindt meer elaboratie plaats tijdens testen dan tijdens herstuderen?**

In de studie in Hoofdstuk 3 onderzochten we de elaboratieve verwerking-hypothese van het testeffect in meer detail, door te kijken naar cognitieve processen die plaatsvinden tijdens herstuderen en testen. We lieten deelnemers woordparen herbestuderen of ophalen en gebruikten een nieuwe benadering om cognitieve processen tijdens het leren te onderzoeken: we maten elektro-encefalogram (EEG) en analyseerden gebeurtenis-gerelateerde potentialen (*event-related potentials*, ERP's). ERP's kunnen waardevolle informatie verschaffen over de cognitieve processen die plaatsvinden tijdens het leren en daardoor de processen die ten grondslag liggen aan het testeffect duidelijk maken. Hoewel sommige onderzoekers (bijv. Eriksson, Kalpouzos & Nybert, 2011) het testeffect hebben onderzocht door middel van functionele kernspintomografie (*functional magnetic resonance imaging*, fMRI) en een studie oscillatoire EEG heeft gebruikt om de effecten van ophalen op latere encoding te onderzoeken (Pastötter, Schicker, Niedernhuber & Bäuml, 2011), was onze studie bij ons weten de eerste ERP-studie naar het testeffect.

Volgens de *dual process*-verklaring van het geheugen (Curran, Tepe & Piatt, 2006; Yonelinas, 2002) zijn twee gescheiden processen betrokken bij het ophalen van informatie uit het geheugen: *familiarity* treedt op wanneer een woord wordt herkend op basis van een 'gevoel van weten' zonder specifieke details van een eerdere presentatie, terwijl *recollection* optreedt wanneer de specifieke details van een eerdere presentatie van het woord worden herinnerd. Gebaseerd op de elaboratieve verwerking-hypothese veronderstelden wij dat *familiarity* een grotere rol speelt tijdens testtrials dan tijdens herstudietrials, omdat *familiarity* gericht is op de globale betekenis van een woord. *Recollection* daarentegen is gericht op fysieke of contextuele details die minder belangrijk zijn voor het activeren van semantisch

gerelateerde woorden. Daarom voorspelden we dat het N400 herhalingseffect, dat *familiarity* aangeeft, groter is tijdens testtrials dan tijdens herstudietrials. Dit zou suggereren dat de processen tijdens testtrials meer gericht zijn op de betekenis van de items dan tijdens herstudietrials. Voor het late positieve component (LPC)-herhalingseffect, dat *recollection* aangeeft, voorspelden we dat de LPC-amplitude kleiner zou zijn tijdens testen dan tijdens herstuderen. Dit zou wijzen op minder detailgerichte verwerking tijdens testtrials. Tezamen zouden deze verschillen in N400- en LPC-herhalingseffecten wijzen op een nadruk op betekenis en niet op details, wat zou passen binnen de elaboratieve verwerking-hypothese. Ook veronderstelden we dat meer moeite zou worden gedaan voor testtrials dan voor herstudietrials door verhoogde elaboratieve verwerking, en voorspelden we dat dit zou resulteren in een hogere P300-amplitude.

Er trad een duidelijk testeffect op: geteste woorden werden vaker correct opgehaald dan herbestudeerde woorden op de eindtest na twee dagen. De ERP-data die waren opgenomen tijdens het leren lieten een grotere P300-amplitude zien tijdens testen dan tijdens herstuderen. Dit wijst erop dat de proefpersonen meer moeite deden tijdens testen dan tijdens herstuderen. De verschillen in de N400- en LPC-herhalingseffecten die we voorspelden waren ook aanwezig. Het N400-herhalingseffect was groter voor testen dan voor herstuderen en het LPC-herhalingseffect was kleiner voor testen dan voor herstuderen. Gecombineerd geven deze resultaten aan dat tijdens testen meer moeite wordt gedaan voor het leren dan tijdens herstuderen en dat deze moeite tenminste gedeeltelijk wordt veroorzaakt door het ophalen van semantisch gerelateerde informatie. Het lijkt er dus op dat tijdens testen meer semantisch gerelateerde informatie wordt geactiveerd dan tijdens herstuderen. De resultaten van de studie in Hoofdstuk 3 lijken daarom consistent te zijn met de elaboratieve verwerking-hypothese van het testeffect.

### **Worden tijdens elaboratie *mediators* geactiveerd?**

Samengenomen laten de studies in Hoofdstuk 2 en 3 een gemengd beeld zien. In de studie in Hoofdstuk 4 begonnen we een wat specifiekere versie van de elaboratieve verwerking-hypothese te onderzoeken: de *mediator effectiveness*-hypothese (Pyc & Rawson, 2010, 2012). Volgens de *mediator effectiveness*-hypothese worden tijdens testen *mediators* geactiveerd. Een *mediator* is een woord dat sterk geassocieerd is met de cue en gemakkelijk kan worden gekoppeld aan de target. Om de activatie van *mediators* te testen lieten we in de studie in Hoofdstuk 4 deelnemers paren leren van woorden die beide gerelateerd waren aan een homoniem. Deelnemers leerden bijvoorbeeld het woordpaar *teller* – *account* (*bankbediende* – *rekening*). De woorden *bankbediende* en *rekening* zijn beide gerelateerd aan het homoniem *bank* (een

financiële instelling). De woordparen werden bestudeerd en dan twee keer herbestudeerd of twee keer getest. In een zogenaamd niet-gerelateerd 'tweede experiment' lazen de deelnemers zinnen waarin het homoniem in de andere betekenis voorkwam. Na het leren van *teller – account* (woorden gerelateerd aan de financiële instelling *bank*) lazen ze bijvoorbeeld de zin *he lay down in the grass by the bank* (een Nederlands voorbeeld zou kunnen zijn: *tijdens het wandelen ging hij zitten op een bank*). In de zin was de andere betekenis van het homoniem duidelijk. Tijdens het lezen van de zinnen werden ERP's gemeten. We veronderstelden dat het homoniem *bank* vaker zou worden geactiveerd tijdens het leren van *teller – account* door testen dan door herstuderen. Daarom voorspelden we dat het lezen van het woord *bank* in de tegenovergestelde betekenis na de leerfase vaker zou resulteren in een conflict tussen de twee betekenissen van het homoniem na testen dan na herstuderen.

Omdat de N400 de meest gebruikte ERP-maat voor semantische integratie is, verwachtten we dat de amplitude van de N400 hoger zou zijn tijdens het lezen van homoniemen waarvan de gerelateerde woorden waren getest dan tijdens het lezen van homoniemen waarvan de gerelateerde woorden waren herbestudeerd. We vonden echter geen verschil in N400-amplitude tussen de twee manieren van leren. We vonden dus geen bewijs voor meer elaboratie tijdens testen dan tijdens herstuderen. De resultaten van dit experiment suggereren daarom dat de rol van elaboratieve verwerking in het ontstaan van het testeffect op zijn best beperkt is.

Ook de studie in Hoofdstuk 5 werd opgezet om de *mediator effectiveness*-hypothese te valideren. We voerden een reeks experimenten uit via Mechanical Turk. Het oorspronkelijke doel van de studie was het onderzoeken van een opvallende eigenschap van de materialen die werden gebruikt in het experiment van Carpenter (2011, Experiment 2). Zij onderzocht de *mediator effectiveness*-hypothese door deelnemers gerelateerde woordparen zoals *moeder – kind* te laten leren door middel van testen of herstuderen. Op de eindtest kregen deelnemers een van de volgende drie typen cues: (1) de originele cue (*moeder*), (2) een woord dat zwak geassocieerd is met de target (*geboorte*), (3) een woord dat sterk geassocieerd is met de cue, maar niet met de target (i.e., een *mediator*: *vader*). De resultaten lieten een testeffect zien wanneer de eindtestcue de originele cue was. Bovendien was het testeffect groter wanneer de eindtestcue een *mediator* was dan wanneer dit een target-gerelateerd woord was. Dit suggereert dat tijdens leren de *mediator* wordt geactiveerd en gekoppeld aan de target, wat resulteert in een associatie tussen de *mediator* en de target die voorheen niet bestond.

Echter, na een nadere blik op de materialen merkten we op dat sommige van de *mediators* een sterke voorwaartse associatie hadden met de originele cue.

Bijvoorbeeld de *mediator vader* heeft een sterke associatie met de originele cue *moeder*. In zo'n geval wordt de originele cue *moeder* direct geactiveerd bij presentatie van de *mediator vader*. Omdat testen de cue-target associatie meer versterkt dan herstuderen is het na testen makkelijker om de target op te halen op basis van de originele cue dan na herstuderen. Daarom kan een sterke associatie tussen de *mediator* en de originele cue het ontstaan van het testeffect verklaren zonder aan te nemen dat de *mediator* wordt geactiveerd tijdens testen. Om deze alternatieve verklaring van Carpenter's (2011) resultaten te verklaren, maakten we twee lijsten van woordensets. In de ene lijst was er geen *mediator*-cue associatie. In de andere lijst was er een sterke *mediator*-cue associatie. We voorspelden dat de eerste lijst geen sterker testeffect voor *mediator* eindtestcues dan voor target-gerelateerde eindtestcues zou opleveren. Voor de tweede lijst voorspelden we hetzelfde effect dat Carpenter vond: een groter testeffect voor *mediator* eindtestcues dan voor target-gerelateerde eindtestcues.

In tegenstelling tot onze voorspellingen vonden we geen interactie tussen leerconditie en mediator-cue-associatie. De resultaten van Carpenter (2011) werden blijkbaar niet veroorzaakt door sterke associaties tussen de mediator en de originele cue. Bovendien, en tot onze verrassing, konden we de originele bevinding van Carpenter niet bevestigen omdat er geen verschil was in de grootte van het testeffect tussen *mediator* en target-gerelateerde eindtestcues. Deze bevinding leidde ons ertoe twee pogingen tot replicatie van Carpenter's studie uit te voeren. We analyseerden de resultaten van onze drie studies en die van Carpenter's Experiment 2 in een kleinschalige meta-analyse. Deze analyse was geïnspireerd door de 'new statistics'-benadering (Cumming, 2012). De betrouwbaarheidsintervallen uit de analyse illustreerden de beperkte precisie van de bevindingen. De meta-analyse stelde ons in staat alle bevindingen te combineren in een duidelijk overzicht. Bovendien gaf de meta-analyse vergeleken met de afzonderlijke studies een veel nauwkeuriger schatting van het cruciale interactie-effect. De analyse bevestigde dat er een testeffect was in alle experimenten. De grootte van het testeffect verschilde echter niet systematisch tussen target-gerelateerde en *mediator* eindtestcues. Deze resultaten zijn niet in lijn met de *mediator effectiveness*-hypothese; ze suggereren dat *mediators* niet vaker worden geactiveerd tijdens testen dan tijdens herstuderen.

## Conclusie

Hoewel we bewijs vonden voor verhoogde elaboratieve verwerking tijdens testen vergeleken met herstuderen (Hoofdstuk 3), vonden we ook een testeffect

wanneer de geleerde materialen geen elaboratie toelieten (Hoofdstuk 2). Bovendien vonden we geen bewijs voor de activatie van *mediators* tijdens testen (Hoofdstukken 4 en 5). Daarom ondersteunen de resultaten van deze dissertatie de *mediator effectiveness*-hypothese niet. Het testeffect lijkt niet te worden veroorzaakt door de activatie van *mediators* tijdens testen.





# References

- Arzouan, Y., Goldstein, A., & Faust, M. (2007). Brainwaves are stethoscopes: ERP correlates of novel metaphor comprehension. *Brain Research*, 1160, 69–81. doi:10.1016/j.brainres.2007.05.034
- Baayen, R. H., Piepenbrock, R., & Van Rijn, H. (1993). *The CELEX lexical database [CD-ROM]*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.
- Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, 85, 89–99. doi:10.1080/00220671.1991.10702818
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.
- Bouwmeester, S., & Verkoeijen, P. P. J. L. (2011). Why do some children benefit more from testing than others? Gist trace processing to explain the testing effect. *Journal of Memory and Language*, 65, 32–41. doi:10.1016/j.jml.2011.02.005
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6, 3–5. doi:10.1177/1745691610393980
- Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, 19, 514–527. doi:10.1080/09541440701326097
- Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 1563–1569. doi:10.1037/a0017021
- Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 1547–1552. doi:10.1037/a0024140
- Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, 19, 619–636. doi:10.1002/acp.1101
- Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, 34, 268–276. doi:10.3758/BF03193405
- Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, 14, 474–478. doi:10.3758/BF03194092

- Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, 36, 438–448. doi:10.3758/MC.36.2.438
- Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, 20, 633–642. doi:10.3758/BF03202713
- Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridge of optimal retention. *Psychological Science*, 19, 1095–1102. doi:10.1111/j.1467-9280.2008.02209.x
- Chan, J. C. K., McDermott, K. B., & Roediger, H. L., III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, 135, 553–571. doi:10.1037/0096-3445.135.4.553
- Chwilla, D. J., Kolk, H. H., & Vissers, C. T. W. M. (2007). Immediate integration of novel meanings: N400 support for an embodied view of language comprehension. *Brain Research*, 1183, 109–123. doi:10.1016/j.brainres.2007.09.014
- Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82, 407–428. doi:10.1037/0033-295X.82.6.407
- Coppens, L. C., Gootjes, L., & Zwaan, R. A. (2012). Incidental picture exposure affects later reading: evidence from the N400. *Brain and Language*, 122, 64–69. doi:10.1016/j.bandl.2012.04.006
- Coppens, L. C., Verkoeijen, P. P. J. L., & Rikers, R. M. J. P. (2011). Learning Adinkra symbols: The effect of testing. *Journal of Cognitive Psychology*, 23, 351–357. doi:10.1080/20445911.2011.507188
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.
- Curran, T. (1999). The electrophysiology of incidental and intentional retrieval: ERP old/new effects in lexical decision and recognition memory. *Neuropsychologia*, 37, 771–785. doi:10.1016/S0028-3932(98)00133-X
- Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory & Cognition*, 28, 923–938. doi:10.3758/BF03209340
- Curran, T., & Cleary, A. M. (2003). Using ERPs to dissociate recollection from familiarity in picture recognition. *Brain Research. Cognitive Brain Research*, 15, 191–205. doi:10.1016/S0926-6410(02)00192-1
- Curran, T., Tepe, K., & Piatt, C. (2006). ERP explorations of dual processes in recognition memory. In H. D. Zimmer (Ed.), *Handbook of Binding and*

- Memory: Perspectives from Cognitive Neuroscience*. Oxford, UK: Oxford University Press.
- Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, 7, 555–561. doi:10.1177/1745691612459059
- Franklin, M. S., Dien, J., Neely, J. H., Huber, E., & Waterson, L. D. (2007). Semantic priming modulates the N400, N300, and N400RP. *Clinical Neurophysiology*, 118, 1053–1068. doi:10.1016/j.clinph.2007.01.012
- Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, 40(2), 1–104.
- Glass, A., Brill, G., & Ingate, M. (2008). Combined online and in-class pretesting improves exam performance in general psychology. *Educational Psychology*, 28, 483–503. doi:10.1080/01443410701777280
- Glover, J. A. (1989). The “testing” phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, 81, 392–399. doi:10.1037/0022-0663.81.3.392
- Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, 55, 468–484. doi:10.1016/0013-4694(83)90135-9
- Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 371–377. doi:10.1037/0278-7393.15.3.371
- Gunter, T. C., Wagner, S., & Friederici, A. D. (2003). Working memory and lexical ambiguity resolution as revealed by ERPs: A difficult case for activation theories. *Journal of Cognitive Neuroscience*, 15, 643–657. doi:10.1162/jocn.2003.15.5.643
- Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, 304, 438–441. doi:10.1126/science.1095455
- Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 801–812. doi:10.1037/a0023219
- Hintzman, D. L. (2001). Similarity, global matching, and judgments of frequency. *Memory and Cognition*, 29, 547–556. doi:10.3758/BF03200456

- Hintzman, D. L., Summers, J. J., & Block, R. A. (1975). Spacing judgments as an index of study-phase retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 1, 31–40. doi:10.1037/0278-7393.1.1.31
- Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, 30, 513–541. doi:10.1016/0749-596X(91)90025-F
- Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, 101, 621–629. doi:10.1037/a0015183
- Juottonen, K., Revonsuo, A., & Lang, H. (1996). Dissimilar age influences on two ERP waveforms (LPC and N400) reflecting semantic context effect. *Cognitive Brain Research*, 4(2), 99–107. doi:10.1016/0926-6410(96)00022-5
- Kang, S. H. K. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory and Cognition*, 38, 1009–1017. doi:10.3758/MC.38.8.1009
- Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, 21, 157–163. doi:10.1177/0963721412443552
- Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331, 772–775. doi:10.1126/science.1199327
- Karpicke, J. D., & Roediger, H. L., III. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 704–719. doi:0278-7393.33.4.704
- Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language*. doi:10.1016/j.jml.2012.02.004
- Karpicke, J. D., & Zaroomb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language*, 62, 227–239. doi:10.1016/j.jml.2009.11.010
- Koole, S. L., & Lakens, D. (2012). Rewarding Replications A Sure and Simple Way to Improve Psychological Science. *Perspectives on Psychological Science*, 7, 608–614. doi:10.1177/1745691612462586
- Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, 65, 85–97. doi:10.1016/j.jml.2011.04.002

- Kutas, M., & Federmeier, K. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, 4, 463–470. doi:10.1016/S1364-6613(00)01560-6
- Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207, 203–205. doi:10.1126/science.7350657
- Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In *Practical aspects of memory* (pp. 625–632). London: Academic Press.
- Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.
- MacDonald, J. (2009, July 29). West African Wisdom: Adinkra Symbols & Meanings. *Adinkra Symbols of West Africa*. Retrieved November 17, 2009, from [www.adinkra.org](http://www.adinkra.org)
- Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, 7, 537–542. doi:10.1177/1745691612460688
- McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, 19, 494–513. doi:10.1080/09541440701326154
- Mecklinger, A. (2000). Interfacing mind and brain: A neurocognitive model of recognition memory. *Psychophysiology*, 37, 565–582. doi:10.1111/1469-8986.3750565
- MettingVanRijn, A. C., Peper, A., & Grimbergen, C. A. (1990). High quality recording of bioelectric events. I: Interference reduction, theory and practice. *Medical & Biological Engineering & Computing*, 28, 389–397. doi:10.1007/BF02441666
- MettingVanRijn, A. C., Peper, A., & Grimbergen, C. A. (1991). High-quality recording of bioelectric events: Part 2 Low-noise, low-power multichannel amplifier design. *Medical & Biological Engineering & Computing*, 29, 433–440. doi:10.1007/BF02441666
- Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. Retrieved from <http://www.usf.edu/FreeAssociation/>
- Otten, M., & Van Berkum, J. (2008). Discourse-based word anticipation during language processing: Prediction or priming? *Discourse processes*, 45, 464–496. doi:10.1080/01638530802356463

- Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, 14, 187–193. doi:10.3758/BF03194050
- Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7, 528–530. doi:10.1177/1745691612465253
- Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, 118, 2128–2148. doi:10.1016/j.clinph.2007.04.019
- Polich, J., & Criado, J. R. (2006). Neuropsychology and neuropharmacology of P3a and P3b. *International Journal of Psychophysiology*, 60, 172–185. doi:10.1016/j.neulet.2007.11.044
- Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, 60, 437–447. doi:10.1016/j.jml.2009.01.004
- Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, 330, 335–335. doi:10.1126/science.1191465
- Pyc, M. A., & Rawson, K. A. (2012). Why is test-restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 737–746. doi:10.1037/a0026166
- Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15, 20–27. doi:10.1016/j.tics.2010.09.003
- Roediger, H. L., III, & Karpicke, J. D. (2006a). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science*, 17, 249–255. doi:10.1111/j.1467-9280.2006.01693.x
- Roediger, H. L., III, & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181–210. doi:10.1111/j.1745-6916.2006.00012.x
- Roediger, H. L., III, Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin and Review*, 8, 385–407. doi:10.3758/BF03196177
- Rugg, M. D. (1995). ERP studies of memory. In M. D. Rugg & M. Coles (Eds.), *Electrophysiology of Mind: Event-Related Brain Potentials and Cognition* (pp. 133–170). Oxford, UK: Oxford University Press.

- Rugg, M. D., Schloerscheidt, A. M., & Mark, R. E. (1998). An Electrophysiological Comparison of Two Indices of Recollection. *Journal of Memory and Language*, 39, 47–69. doi:10.1006/jmla.1997.2555
- Schnyer, D. M. (1997). Event-related brain potential examination of implicit memory processes: Masked and unmasked repetition priming. *Neuropsychology*, 11, 243–260. doi:10.1037/0894-4105.11.2.243
- 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, 716–727. doi:10.1037/0278-7393.14.4.716
- Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, 56, 252–257. doi:10.1027/1618-3169.56.4.252
- Van Strien, J. W., Verkoeijen, P. P. J. L., Van der Meer, N., & Franken, I. H. A. (2007). Electrophysiological correlates of word repetition spacing: ERP and induced band power old/new effects with massed and spaced repetitions. *International Journal of Psychophysiology*, 66, 205–214. doi:10.1016/j.ijpsycho.2007.07.003
- Vanderplas, J. M., & Garvin, E. A. (1959). The association value of random shapes. *Journal of Experimental Psychology*, 57.
- Verkoeijen, P. P. J. L., Bouwmeester, S., & Camp, G. (2012). A short-term testing effect in cross-language recognition. *Psychological Science*, 23, 567–571. doi:10.1177/0956797611435132
- Voss, J. L., & Paller, K. A. (2009). Remembering and knowing: Electrophysiological distinctions at encoding but not retrieval. *NeuroImage*, 46, 280–289. doi:10.1016/j.neuroimage.2009.01.048
- Wagner, S., & Gunter, T. C. (2004). Determining inhibition - Individual differences in the “lexicon context” trade-off during lexical ambiguity resolution in working memory. *Experimental Psychology*, 51, 290–299. doi:10.1027/1618-3169.51.4.290
- Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory*, 11, 571–580. doi:10.1080/09658210244000414
- Yonelinas, A. (2002). The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language*, 46, 441–517. doi:10.1006/jmla.2002.2864



# Dankwoord

## Acknowledgements

Graag wil ik alle mensen bedanken die op wat voor manier dan ook hebben bijgedragen aan dit proefschrift.

Allereerst, veel dank aan alle proefpersonen (ruwe schatting: 2000). Vooral psychologiestudenten, maar ook collega's en vrienden (voor pilotstudies), Mechanical Turkers, en anderen die toevallig in de buurt waren. Zonder proefpersonen geen psychologisch onderzoek!

Prof. dr. Rikers, mijn promotor. Remy, dank voor je snelle en enthousiaste reacties op stukken en voor je commentaar vanuit een nieuwe invalshoek. Ondanks je drukke agenda maakte je altijd tijd voor me. Je goede adviezen en motiverende woorden hielpen me steeds vooruit.

Dr. Verkoeijen, mijn copromotor en dagelijks begeleider. Peter, dank voor je motiverende, humorvolle en positieve begeleiding. Jij zag het altijd nog zitten en je deur stond altijd open.

Leden van de leescommissie: dr. Gino Camp, dr. Katinka Dijkstra en professor Tamara van Gog. Bedankt dat jullie de tijd en moeite hebben genomen om mijn proefschrift te lezen en beoordelen en dat jullie met mij willen discussiëren over mijn proefschrift. Leden van de grote commissie, professor Guido Band, dr. Huib Tabbers en professor Jan van Strien, bedankt dat jullie met mij van gedachten willen wisselen tijdens de verdediging van mijn proefschrift.

Collega's van de C&L-groep: Mario, Gerdien, Nicole, Huib, Samantha, Peter, Gabriela, Jan. Bedankt voor de discussies, nuttige suggesties, kritische vragen en gezelligheid.

Collega's van O&O, en in het bijzonder de deelnemers aan de pub group 'with double meaning', dank voor jullie major, minor, major minor en minor major comments op mijn stukken. En ook van het reviewen van jullie stukken heb ik veel geleerd.

De mannen van het EBL, dank voor alle hulp en belangstelling.

Karen, mijn ex-kamergenootje en nu mijn paranimf! Op magische wijze zaten we steeds in dezelfde fase. Stuk afgewezen, onderwijs geven, bergen data analyseren, schrijven, zelfs een kind krijgen deden we bijna tegelijk. Bedankt voor de mooie reizen, de herkenning, dat ik af en toe je hersenen mocht lenen en tegen je aan mocht (mag) klagen. Ik hoop dat onze dochters nog lang samen zullen spelen.

Anja en Marije, ARP'ers van het eerste uur. Nu hebben we het alle drie gehaald! Dank voor de etentjes, koffie, kletsavondjes. Heerlijk om het soms niet te hoeven uitleggen en om eens niet te horen 'het is maar werk'. Marije, dankjewel dat je mijn paranimf wilt zijn!

G12'ers: Jojanneke, Joris, Michèle, Naomi, Robert. Heerlijk om soms eens te horen dat er veel meer belangrijke zaken zijn in het leven, dat jullie met allerlei andere dingen bezig zijn, kortom dat promoveren ook maar werk is. Joris, bedankt dat je me binnen hebt gelooft bij je eigen werkgever!

Mijn familie: Pap en mam, ik kan me geen betere ouders wensen. Vol vertrouwen laten jullie me vrij en als ik het nodig heb, geven jullie precies het juiste advies. Wat is het fantastisch om jullie nu als grootouders te zien. Coen, hoewel je niet altijd begreep waar ik mee bezig was, was het voor mij wel motiverend dat je het stoer vond. Carmen, nu ben ik dan eindelijk afgestudeerd! ;-)) En Pepijn, kleine lachebek, met jou lach ik vanzelf mee.

Ardjoena, bedankt dat je het al zo lang met me uithoudt, met mijn gekke plannen en gestress. Heel veel dank voor je steun en vertrouwen!

Last but not most, Rosalinde. Meis, bedankt dat je er bent en me blij maakt. Je laat me de wereld anders bekijken en je bent je er waarschijnlijk niet van bewust, maar jouw relativerend vermogen kent geen grenzen.

*Ook aan iedereen die ik niet persoonlijk genoemd heb, maar die wel heeft bijgedragen aan dit proefschrift: hartelijk bedankt!*



# Curriculum Vitae



Leonora Coppens was born in Capelle aan den IJssel, the Netherlands, on June 11<sup>th</sup> 1984. She completed her secondary education in 2002 at the Comenius College in Capelle aan den IJssel. She received her Master's degree in Biological and Cognitive Psychology in 2008 at the Erasmus University Rotterdam. Her published master's thesis was on the role of visual information in language comprehension. For the next year, she combined working as a scientific teacher with the Advanced Research Program at the Erasmus University Rotterdam. In 2009, she started working as a PhD student at the Erasmus University Rotterdam to investigate the cognitive mechanisms that underlie the testing effect, of which the present dissertation is the result. In addition to performing research, she was involved in a number of educational tasks. She supervised bachelor and master students with writing their thesis and taught courses in cognitive, biological and educational psychology and statistics & methodology.

## **Publications**

### **International peer-reviewed papers**

- Coppens, L. C., Gootjes, L. & Zwaan, R. A. (2012). Incidental picture exposure affects later reading: Evidence from the N400. *Brain & Language*, 122, 64-69.
- Coppens, L. C., Verkoeijen, P. P. J. L., & Rikers, R. M. J. P. (2011). Learning Adinkra symbols: The effect of testing. *Journal of Cognitive Psychology*, 23, 351-357. (Chapter 2)
- Gootjes, L., Coppens, L. C., Zwaan, R. A., Franken, I. H. A., & Van Strien, J. W. (2011). Effects of recent word exposure on emotion-word Stroop interference: An ERP study. *International Journal of Psychophysiology*, 79, 356-363.

### **Submitted for publication**

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





