GESA VAN DEN BROEK

BENEFITS OF MEMORY RETRIEVAL FOR VOCABULARY LEARNING

A NEUROCOGNITIVE PERSPECTIVE

BENEFITS OF MEMORY RETRIEVAL FOR VOCABULARY LEARNING

A Neurocognitive Perspective

Gesa van den Broek

The research presented in this thesis was supported by a grant from The Netherlands Organisation for Scientific Research, National Initiative Brain and Cognition (grant number: 056-33-014; https://www.hersenenencognitie.nl) and a Language Learning dissertation grant to Gesa van den Broek.

ISBN
978-94-92380-61-6
Cover
Bureau Brouns, Utrecht. www.burobrouns.nl
Lay-Out
Joska Sesink, www.persoonlijkproefschrift.nl
Print
GVO drukkers & vormgevers B.V.

BENEFITS OF MEMORY RETRIEVAL FOR VOCABULARY LEARNING

A Neurocognitive Perspective

Proefschrift
ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op maandag 2 oktober 2017
om 14.30 uur precies

door

Gesa Sonja Elsa van den Broek
geboren op 9 oktober 1985
te Kleve (Duitsland)

Promotoren
Prof. dr. L.T.W. Verhoeven
Prof. dr. G.S.E. Fernández
Prof. dr. P.C.J. Segers

Copromotor
Dr. A. Takashima

Manuscriptcommissie
Prof. dr. J.M. McQueen
Prof. dr. A.P.J. van den Bosch
Prof. dr. L. Kester (UU)

BENEFITS OF MEMORY RETRIEVAL FOR VOCABULARY LEARNING

A Neurocognitive Perspective

Doctoral Thesis

to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.H.J.M. van Krieken,
according to the decision of the Council of Deans
to be defended in public on Monday, October 2, 2017
at 14.30 hours

by

Gesa Sonja Elsa van den Broek
Born on October 9, 1985
in Kleve (Germany)

Supervisors
Prof. dr. L.T.W. Verhoeven
Prof. dr. G.S.E. Fernández
Prof. dr. P.C.J. Segers

Co-Supervisor
Dr. A. Takashima

Doctoral Thesis Committee
Prof. dr. J.M. McQueen
Prof. dr. A.P.J. van den Bosch
Prof. dr. L. Kester (UU)

CONTENTS

GENERAL INTRODUCTION

Abstract. The work described in this dissertation is an attempt to gain insight into practice conditions that lead to effective vocabulary learning in a foreign language, and in particular, into the benefits of memory retrieval for word retention. Many psychological studies have shown that practicing the retrieval of words from memory is a powerful way to boost the retention of words over time. In order to capitalize on this so-called testing effect, a good understanding is needed of the cognitive mechanisms involved, and the prerequisites and boundary conditions that apply during vocabulary exercises. The first aim of this thesis is therefore to provide insight into neural and cognitive mechanisms involved in retrieval practice. The second aim of this thesis is to investigate the effect of retrieval opportunities during vocabulary exercises. This first chapter of the dissertation provides an introduction to the key theoretical accounts and terms discussed in the thesis, and describes the overarching aims and research questions.

Vocabulary acquisition is an essential part of mastering a language. In their first language, learners acquire much of their vocabulary incidentally, as a by-product of being exposed to words in context (Krashen, 1989; Swanborn & de Glopper, 1999). In contrast, learners of a foreign language often do not experience the conditions that are needed to incidentally learn vocabulary (e.g., Nation, 2001c). For example, foreign language learners may not encounter words frequently enough to remember them over time. Therefore, intentional vocabulary practice is important for foreign language acquisition. Many studies have shown that such intentional practice enhances vocabulary learning compared to purely incidental exposure and that some practice conditions lead to better retention of vocabulary than other conditions (overviews in Hulstijn & Laufer, 2001; Schmitt, 2008). The work described in this thesis is an attempt to gain insight into practice conditions that lead to successful vocabulary learning in a foreign language, and in particular, into benefits of memory retrieval for vocabulary learning.

1.1 EFFECTIVE VOCABULARY LEARNING WITH MEMORY RETRIEVAL

The term *vocabulary* refers to the body of words known to an individual person ("vocabulary," 2017). Word knowledge is a complex concept, given that words are both a unit of speech and thought (Koenig & Woodward, 2007). As Miller (1999, p. 2) put it: "knowing a word is generally considered to be a matter of knowing the word's meaning, and meaning is one of those concepts of great importance for understanding the nature and limits of psychology." In applied research, vocabulary knowledge in a foreign language is therefore often defined by the tasks that learners master when they know a word (Schmitt & Meara, 1997). These tasks include the recall of the form – for example, sound or spelling – and meaning of a word, but also the correct grammatical use and the understanding of a word's register and collocations (Nation, 2001a). The encoding of a previously unknown word form (e.g., "keha") and the association of that word form with meaning (e.g., "keha" = house) are considered an important step in this word learning process (e.g., Deconinck, Boers, & Eyckmans, 2015).

Everyone who has looked up a word in the dictionary and moments later found him- or herself at a loss of that same word again, recognizes that remembering a new form-meaning association is a gradual process that often requires multiple repetitions (Pigada & Schmitt, 2006; Webb, 2007). Moreover, it is important to keep in mind that word learning depends on the way in which learners process a word. According to Schmitt (2008, p. 329), "the overriding principle for maximizing vocabulary learning is to increase the amount of engagement learners have with lexical items". For more

than 40 years, the notion of *processing depth* (Craik & Lockhart, 1972) has been used to describe such learner engagement. Conditions that cause *deeper* or more effortful processing are said to lead to better retention than conditions that lead to superficial or shallow processing. For word learning, deep processing is often equated to semantic as opposed to structural elaboration, because early experiments showed that learners remembered words better after categorizing them based on their meaning than after categorizing based on font or sound (Craik & Tulving, 1975). However, while the concept of processing depth is widely used, it is also abstract, which makes it difficult to determine which of several activities triggers deeper processing (e.g., Baddeley, 1978; Eysenck, 1978).

Different frameworks have been presented to clarify which activities stimulate deep processing of vocabulary words. According to the *involvement load* framework (Laufer & Hulstijn, 2001), for example, a word's retention is predicted by the learner's need to use or understand the word, the learner's engagement in the evaluation of the word (e.g., of the word's fit into a particular context), and the degree to which the learner searches for information about the word by consulting others or a dictionary (Laufer & Hulstijn, 2001). Similarly, Nation (2001b) proposed that three general psychological processes contribute to the retention of words: noticing, generative use, and retrieval. *Noticing* involves the learner paying attention to a word as a unit of language, which has a meaning outside of its immediate context. *Generative use* refers to the use or recognition of words in varying contexts to enrich word knowledge. Regarding retrieval, Nation distinguished *receptive retrieval*, which "involves perceiving the form and having to retrieve its meaning when the word is met in listening or reading" and *productive retrieval*, which "involves wishing to communicate the meaning of the word and having to retrieve its spoken or written form." (2001b, p. 67). Unlike the external search component of the involvement load framework (Laufer & Hulstijn, 2001), Nation thus put a *mental* search for words forward as a crucial mechanism to enhance the retention of words over time. He claimed that "each retrieval of a word strengthens the path linking form and meaning and makes subsequent retrieval easier." (p. 67). Such benefits of retrieval have not been documented extensively in vocabulary learning research (cf. Barcroft, 2015; Folse, 2006; Nakata, 2016) but are well known in the psychological literature (Roediger & Karpicke, 2006b).

Many psychological studies have shown that retrieving information from memory is a powerful way to boost the retention of that information over time (reviews in Nunes & Karpicke, 2015; Roediger & Butler, 2011; Roediger & Karpicke, 2006b; Rowland, 2014; van Gog & Sweller, 2015). To get an idea of what retrieval practice constitutes in the context of vocabulary learning, imagine a student practicing a list of foreign words. After reading the words and translations a few times, she covers the translations with

her hand to see if she can translate them from memory[1]. By doing so, she practices the retrieval of words from memory, which is an effective means to enhance later recall (meta-analyses in Adesope, Trevisan, & Sundararajan, 2017; Rowland, 2014). Karpicke and Roediger (2008), for example, demonstrated the benefits of memory retrieval for vocabulary learning in a study with university students who learned the English translation of Swahili words. Once students could produce the translation of a Swahili word, the word was either dropped from practice, was further practiced by restudying the word with its translation, was further practiced by retrieving the translation from memory, or was practiced by both restudying and retrieval. The conditions that included repeated retrieval led to significantly better later recall than the other conditions, including the restudy conditions. Such a positive effect of retrieval practice compared to restudying is known as *testing effect*[2]. The testing effect is a robust phenomenon in cognitive psychology that has been replicated with different age groups, materials, and settings (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). However, in spite of a large amount of literature on the benefits of retrieval practice, the theoretical understanding of the effect lags behind (Carpenter & Yeung, 2017; Roediger & Butler, 2011).

1.2 COGNITIVE MECHANISMS AND NEURAL CORRELATES OF RETRIEVAL PRACTICE

In order to capitalize on the benefits of retrieval practice for vocabulary learning, a good understanding is needed of the cognitive mechanisms involved. Several explanations, such as the *retrieval effort hypothesis* (Pyc & Rawson, 2009) and *transfer appropriate processing* (e.g., Roediger & Karpicke, 2006a, 2006b) explain testing effects in abstract terms. They propose that 'effort' underlies testing effects because difficult retrieval has greater benefits than easier retrieval (e.g., Carpenter, 2009; Carpenter & Delosh, 2006; Karpicke & Bauernschmidt, 2011) or that recall on the final test is facilitated because it resembles cognitive processes during retrieval practice (e.g., Roediger & Karpicke, 2006a, 2006b). Another influential theory is the *New theory of disuse* (Bjork & Bjork, 1992), which holds that the momentary accessibility of a memory (called *retrieval strength*) is different from its *storage strength*, which determines the memory's long-term retention. According to this theory, retrieval

1 Or in a modern scenario, she could enter the words into a computer program, which would then repeatedly prompt her to translate the words from memory.

2 The expressions "*testing effect"* and *"benefits of retrieval practice"* are used interchangeably in the present thesis, to refer to enhanced retention after retrieval practice compared to restudy practice.

practice increases storage strength more if the momentary retrieval strength is low, for example, due to demanding practice conditions. Though frequently cited, these explanations do not specify what constitutes mental effort or low retrieval strength, and why retrieval becomes facilitated with practice (Karpicke, Lehman, & Aue, 2014). More specific descriptions of the cognitive and neural mechanisms that make (effortful) retrieval beneficial for retention are needed in order to understand its prerequisites and boundary conditions (Carpenter & Yeung, 2017; Whiffen & Karpicke, 2017).

1.2.1 COGNITIVE MECHANISMS

Mechanistic accounts of retrieval practice often distinguish between retrieval cues, target information, and retrieval routes. Retrieval cues are extracted from input that a learner perceives, and lead to the activation of corresponding mental representations (e.g., when reading the foreign word "keha"). Target information is relevant information stored in memory that becomes activated in response to available cues (e.g., the translation "house"). The mental associations that allow the activation of target information based on cues, are called cue-target associations or retrieval routes (e.g., Roediger & Butler, 2011). A consensus among different accounts of testing effects is that retrieval practice strengthens retrieval routes (e.g., Carpenter, 2011; Roediger & Butler, 2011; Roediger & Karpicke, 2006b). However, an open question is how the retrieval routes change. Two possible mechanisms are prominent in the current literature: on the one hand, the elaboration account (Carpenter, 2009, 2011; Carpenter & Yeung, 2017) and on the other hand, controlled inhibition (Thomas & McDaniel, 2013; Wimber, Alink, Charest, Kriegeskorte, & Anderson, 2015) or context reinstatement accounts (Jacoby, Shimizu, Daniels, & Rhodes, 2005; Karpicke et al., 2014; Whiffen & Karpicke, 2017), which I summarize here as selection accounts.

Both elaboration and selection accounts of retrieval practice imply that mental representations of words in memory arise from activation in semantic networks. This is based on classical spreading activation theories of semantic memory (e.g., Collins & Loftus, 1975), which hold that words are represented in networks defined by their associations with other words or word features. The recognition of specific words is considered to be the result of the spread of activation between interconnected nodes in this network, a process which has also been described in psycholinguistic models of the so-called mental *lexicon*[3] (for an overview, see Gaskell, 2007). A comprehensive discussion of these models is beyond the scope of this chapter but what is important

3 Note that the mental lexicon is thought to include word features like phonology (sounds), semantics (meaning), and orthography (spelling) (e.g., Dijkstra, 2007), whereas accounts of testing effects explicitly mention only semantic information and associations between words rather than between word features.

for the present purpose is the basic idea that nodes in the network can be more or less activated and pass on activation via excitatory and inhibitory connections with other nodes, thereby activating nodes with which they are consistent and inhibiting nodes with which they are inconsistent (Gaskell, 2007). In bilingual speakers, activation may also spread between languages (e.g., Dijkstra, 2007). In such a connectionist framework, learning is considered to influence the future spread of activation within the network through the addition of new connections or nodes to the network, or the change of existing connections (Gaskell, 2007).

The elaboration account (Carpenter, 2009, 2011) and selection accounts (Karpicke et al., 2014; Thomas & McDaniel, 2013) make different predictions about the nature of retrieval-induced changes in semantic networks. According to elaboration accounts, the mental search during retrieval activates various semantic nodes in the network which subsequently become associated with cue and target information (Carpenter & Yeung, 2017). This elaboration of the semantic network creates additional connections which later function as alternative retrieval routes to access the target information. In contrast, selection accounts hold that the number of activated nodes decreases with repeated retrieval such that target information can be more easily selected amongst a smaller search-set of candidate responses (Karpicke & Blunt, 2011; Karpicke & Smith, 2012; Thomas & McDaniel, 2013). This search-set reduction is thought to occur through a refinement of representations of the encoding context (Karpicke et al., 2014; Lehman, Smith, & Karpicke, 2014; Whiffen & Karpicke, 2017) and/or through monitoring processes that select cue-target associations and suppress competing irrelevant connections (Thomas & McDaniel, 2013; see also Wimber et al., 2015). Elaboration and selection accounts thus make fundamentally different assumptions about the cognitive processes that make retrieval beneficial, and in turn about the cognitive processes that should be stimulated during retrieval practice to enhance learning. It is therefore relevant to investigate which account is more accurate. In the present thesis, this is done by combining behavioral and neuro-imaging measures to understand the timing and neural correlates of testing effects.

1.2.1.1 THE TIMING OF TESTING EFFECTS. Cognitive accounts of testing effects are often evaluated by their ability to predict and explain behavioural results (Karpicke et al., 2014), such as the finding that more effortful retrieval leads to stronger benefits than easier retrieval (Carpenter & Delosh, 2006; Pyc & Rawson, 2009). An empirical finding which has complicated mechanistic explanations of retrieval practice in this respect is the finding that benefits of retrieval compared to restudying often only become visible after a delay but not immediately after learning (e.g., Kornell, Bjork, & Garcia, 2011; Toppino & Cohen, 2009). These delayed testing effects cannot readily be explained by elaboration or selection accounts, because both accounts assume that connections in the semantic network change directly during retrieval practice

and therefore should have immediate benefits for recall. Alternative explanations have been proposed which hold that retrieval somehow reduces forgetting over time whereas restudying increases the initial encoding, but these do not specify the cognitive mechanisms that reduce forgetting (Wheeler, Ewers, & Buonanno, 2003).

The discrepancy between mechanistic accounts and empirical findings of delayed testing effects can also be solved by making two assumptions that are combined in the so-called *bifurcation model* (Halamish & Bjork, 2011; Kornell et al., 2011). First, the model assumes that items have on average higher memory strength after successful retrieval practice than after restudying, but this difference is not visible when the lower memory strength of the restudied items suffices for recall on an easy immediate test. After some time has passed, however, only the strongest memories may still be recalled, and testing effects (i.e., higher memory strength of retrieved items than of restudied items) become visible. Second, retrieval is only considered beneficial compared to restudying if the learner can access the target information (i.e., when the retrieval is successful or failed retrieval is followed by feedback) (Jang, Wixted, Pecher, Zeelenberg, & Huber, 2012; Rowland & DeLosh, 2015). Therefore, if retrieval success is limited and there is no feedback, more restudied items may be remembered than (successfully) retrieved items. Together, the two assumptions of the bifurcation model can explain how even if successful retrieval immediately strengthens memory representations, testing effects can still emerge only over time (a more detailed description is given in Chapter 2).

The bifurcation model is important for the theoretical understanding of testing effects because it implies that mechanistic accounts which predict that retrieval induces immediate beneficial changes in semantic associations, like the elaboration (Carpenter, 2009, 2011) or selection accounts (Lehman et al., 2014; Thomas & McDaniel, 2013), do not need to be changed to include mechanisms that cause a delayed effect or influence forgetting. However, so far, the empirical evidence for the model is limited. Two studies have supported the bifurcation model by manipulating the difficulty of the test to show that harder tests produce earlier testing effects than easy tests, supposedly because hard tests are more sensitive to differences in memory strength (Halamish & Bjork, 2011; Kornell et al., 2011). In the present thesis, a different approach was taken to test the two major assumptions of the model directly. First, we measured response times to tap into differences in the memory strength of correctly recalled items. More accessible, stronger memories lead to shorter (cf. Anderson, 1981; MacLeod & Nelson, 1984; Wixted & Rohrer, 1993). This allowed us to test the prediction that memory strength is higher after successful retrieval than after restudying. Second, we compared later recall of words that were or were not successfully retrieved during practice to test if retrieval success indeed moderates testing effects. This approach is further introduced in Chapter 2.

1.2.2 NEURAL CORRELATES

The classical view on learning is that experiences are transformed into a physical memory trace in the brain through changes in synaptic connections (for reviews, see Kandel, Dudai, & Mayford, 2014; Sekeres, Moscovitch, & Winocur, 2017). This *synaptic consolidation* results in assemblies of interlinked neurons, which form an *engram* (Semon, 1921, in Schacter, Eich, & Tulving, 1978) or physical substrate of memories, to enable the storage and later retrieval of memories through the reinstatement of cortical activity present during encoding (Danker & Anderson, 2010; Josselyn, Köhler, & Frankland, 2015; Li et al., 2016). Engrams are not static; over time, their physical and chemical organization changes (Dudai, 2012). Studies with humans and animals show that the hippocampus, a structure in the medial temporal lobe, is initially critical for storage and retrieval of memories, possibly because it functions as a unifier or pointer to representations across different neocortical areas (Squire, Genzel, Wixted, & Morris, 2015). Over time, the hippocampus becomes gradually less important and permanent representations develop in distributed regions of the neocortex (for reviews, see Dudai, 2012; Squire et al., 2015). This *system consolidation* process can continue for weeks or months, and is partly a function of the reactivation or retrieval of memories, which temporarily destabilizes the engram and initiates new synaptic consolidation cycles (Dudai, 2012; Josselyn et al., 2015; Sekeres et al., 2017).

While learning is thought to involve the strengthening of connections between neurons distributed in the brain, learning and memory processes can be probed at different scales and levels of analysis (Josselyn et al., 2015). One approach is to study the brain regions involved during encoding and later retrieval. Many studies have employed the so-called „subsequent memory paradigm“ to identify regions of the brain in which activation changes during the successful encoding of memories (e.g., Otten, Henson, & Rugg, 2001; Wagner et al., 1998). Strikingly, these studies show that brain activation during the initial processing of information (partly) predicts whether that information can later be recalled. In the subsequent memory paradigm, brain activity is measured while participants study different items. Recall of these items on a later test (i.e., subsequent memory) is then used to make a comparison between brain activity during the study of items that were subsequently remembered and brain activity during the study of items that were subsequently forgotten. Meta-analyses of this contrast show a set of areas including the medial temporal lobe, and subregions of the prefrontal cortex and the posterior parietal cortex, which predict subsequent memory (Kim, 2011; Spaniol et al., 2009).

Whereas the subsequent memory effect has been used to identify the neural correlates of successful initial encoding, other paradigms have been used to study later retrieval success. These studies have compared, for example, brain activity during the correct recognition of items and the correct rejection of unknown items (Spaniol

et al., 2009). A meta-analysis by Spaniol et al. (2009) linked activations in the medial temporal lobe, left prefrontal cortex, parietal cortex, and posterior midline regions to this contrast. The brain areas mediating encoding and retrieval partly overlap, which has been explained with reinstatement of encoding-related activity during retrieval (Danker & Anderson, 2010; Rissman & Wagner, 2012). However, some areas, amongst which the inferior parietal lobe, seem to have different functions during encoding and retrieval, as their activation is related to both unsuccessful encoding and successful memory retrieval (Daselaar et al., 2009; Uncapher & Wagner, 2009).

1.2.2.1 THE NEURAL CORRELATES OF TESTING EFFECTS. It is clear that retrieval is not the endpoint of learning. Memories are not fixed; their reactivation during retrieval places them in a state in which they can be changed or strengthened (Dudai, 2012; Sekeres et al., 2017). Nevertheless, in neuroimaging experiments, retrieval is often the endpoint of data collection (Wing, Marsh, & Cabeza, 2013), leaving open how brain activations during retrieval influence *subsequent* memory. Identifying areas that, when active during retrieval, predict subsequent memory, could provide insight into crucial neural processes during retrieval practice. Such analyses could also inform the debate about cognitive accounts of testing effects. Patterns of activation in different areas of the brain have been related to broad cognitive functions which feature in the existing accounts of testing effects. For example, prefrontal areas have been related to controlled, non-automatic, capacity-limited processing (Race, Kuhl, Badre, & Wagner, 2009). They might interact with the medial temporal lobe during memory retrieval to allow top-down selection of relevant information (Badre & Wagner, 2007; Blumenfeld & Ranganath, 2007). Explanations of testing effects that focus on retrieval effort (e.g., Pyc & Rawson, 2009) and on selection processes (Thomas & McDaniel, 2013) therefore predict that prefrontal areas are more active during retrieval compared to restudying. Moreover, posterior cortical regions including temporal and inferior parietal areas have been related to the representation of retrieved information (Vilberg & Rugg, 2008, 2009) and binding of semantic information (Binder, Desai, Graves, & Conant, 2009; Lau, Phillips, & Poeppel, 2008), which are important processes in elaboration and selection accounts. The differential involvement of these areas during retrieval and restudy, and their relation to subsequent memory, can therefore be used to evaluate the existing cognitive theories to improve insight into the beneficial mechanisms of retrieval. For example, it can be tested whether activations are more in line with increasingly elaborate or selective semantic associations to distinguish between semantic elaboration and selection accounts, which is difficult to achieve with behavioral measures (Lehman et al., 2014). How the neural correlates of memory retrieval can be related to accounts of testing effects is further explained in Chapters 3 and 4 of the thesis. In these chapters, the neural mechanisms underlying testing effects are described using functional

magnetic resonance imaging (fMRI) in order to determine whether patterns of brain activation during retrieval are in line with or contradict accounts that attribute testing effects to retrieval effort, semantic elaboration, or selection.

1.3 MEMORY RETRIEVAL IN VOCABULARY EXERCISES

Many psychological studies have shown benefits of retrieval practice for long-term retention (most recent meta-analysis in Adesope et al., 2017), suggesting that retrieval practice could be beneficial for certain aspects of word learning in a foreign language. However, most of what is known about retrieval practice comes from psychological laboratory experiments with explicit recall tasks (e.g., "keha = ?"). Retrieval might also be triggered with a variety of other tasks that prompt learners to retrieve word knowledge from memory (Nunes & Karpicke, 2015). The prerequisites and boundary conditions of using retrieval in this way for vocabulary learning have not been established clearly. Two aspects that need to be taken into account for such practical applications of retrieval practice are the need for feedback and the effect of contextual information.

1.3.1 RETRIEVAL PRACTICE AND FEEDBACK

Retrieval practice has a strong potential to enhance learning, especially if it is difficult (Bjork & Bjork, 1992). However, there is trade-off between the potency of retrieval, which increases when the retrieval is difficult, and the chance that retrieval is successful, which decreases when the retrieval is difficult (e.g., Finley, Benjamin, Hays, Bjork, & Kornell, 2011). Effortful retrieval is more beneficial than easier retrieval (Carpenter & DeLosh, 2005; Carpenter & Delosh, 2006; Pyc & Rawson, 2009) but if the retrieval fails and a learner does not gain access to the target information, the retrieval attempt has no benefits (Jang et al., 2012; Kornell et al., 2011). One approach to deal with this trade-off is to aim for the highest retrieval difficulty that still avoids retrieval failure, for example, by gradually reducing the amount of hints available (Finley et al., 2011) or by expanding spacing between repetitions over the course of practice (e.g., Karpicke & Roediger, 2007; Storm, Bjork, & Storm, 2010). Computational models of the forgetting rate of individual learners can be used to predict the best moment to repeat a word, which is supposedly just before the word cannot be retrieved from memory anymore (Pavlik & Anderson, 2008; Sense, Behrens, Meijer, & van Rijn, 2016). Still, retrieval failure cannot always be prevented. Therefore, a second crucial approach is to reduce the impact of retrieval failures with corrective feedback. Feedback allows learners to access the correct answer to learn from their errors, which significantly increases benefits of retrieval practice (e.g., Butler & Roediger, 2008; Thomas & McDaniel, 2013).

Different forms of feedback exist (Shute, 2008). Elaborate feedback that triggers deep processing of the correct answer might lead to better retention than standard feedback which only displays the correct answer (van der Kleij, Feskens, & Eggen, 2015). Such elaborate feedback can, for example, consist of additional explanations or hints to help learners retrieve the correct answer (Finn & Metcalfe, 2010). In the study reported in Chapter 5, we tested whether creating an additional retrieval opportunity during the feedback phase enhanced learning. Including retrieval seemed like a promising approach to enhance retention of the answer, given that retrieval practice normally leads to better retention than restudying (e.g., Roediger & Karpicke, 2006b). On the other hand, previous studies have shown that elaborate feedback comes at a cost because it increases feedback processing times, and this can reduce the time available for further repetitions (Hall, Adams, & Tardibuono, 1968; Hays, Kornell, & Bjork, 2010). Moreover, it is not clear if retrieval supported with hints, also leads to better recall when the hints are no longer available (see the transfer-appropriate processing idea, e.g. Lockhart, 2002). The effect of feedback with hints that create a new retrieval opportunity was therefore investigated in Chapter 5, based on a series of experiments with high school students who practiced foreign language vocabulary with an adaptive computer program in a classroom environment.

1.3.2 RETRIEVAL PRACTICE AND CONTEXT

Many people's intuition is that words should be practiced in a rich context rather than in isolation. This is because exposure to a rich and informative context can help learners understand new concepts and expand their vocabulary (e.g., Beck, McKeown, & McCaslin, 1983; Seibert, 1945). However, understanding a word in context does not necessarily lead to the acquisition of that word (Pressley, Levin, & McDaniel, 1987; Verspoor & Lowie, 2003). Usually, words need to be presented multiple times before they are remembered (Horst, Cobb, & Meara, 1998; Hulstijn, 1992; Hulstijn, Hollander, & Greidanus, 1996; Webb, 2007). Moreover, just as in other vocabulary exercises, the way in which words are processed in context, predicts how well they are remembered (Laufer & Hulstijn, 2001). If a learner, for example, focuses only on a word's meaning, retention of the word form is limited (Barcroft, 2002, 2003).

During repeated exposures to a word in context, learners can understand words through the retrieval of word knowledge gained during previous encounters from memory, or through inferences of word meaning from context (Nation, 2015). Contextual information influences this process because contextual clues can facilitate the correct inference of word meaning (Mondria & Wit-de Boer, 1991; Webb, 2008). In contrast, when words occur in an uninformative context, their meaning cannot be inferred. In this case, learners have to search for the word meaning in memory or in a dictionary. For example, to understand the sentence "Where is the funguo?",

a reader needs to retrieve the meaning of "funguo" from memory. In contrast, the sentence "I want to unlock the door. Where is the *funguo*?" provides stronger clues to the meaning of "funguo" (key) and allows the inference of word meaning from context. Given the benefits of retrieval practice, the question arises if it can be beneficial to trigger retrieval during practice of words in context by reducing the amount of contextual information. On the other hand, the inference of word meaning from a semantic context could also trigger beneficial, deep processing that enhances retention (e.g., Carpenter, Sachs, Martin, Schmidt, & Looft, 2012; but see Mondria, 2003). The effect of inferences and retrieval during the practice of vocabulary words in context are addressed in Chapter 6.

1.4 THE PRESENT THESIS

1.4.1 AIMS AND RESEARCH QUESTIONS

The first aim of this thesis is to provide insight into the neurocognitive base of the benefits of memory retrieval for vocabulary learning. For this purpose, reaction time data and neuroimaging data were used to test predictions derived from cognitive accounts of the testing effect. The overarching question is split into two subquestions:
Question 1. Which cognitive processes underlie the benefits of retrieval practice for word retention?

- Question 1a. Do behavioral measures including reaction times confirm the predictions of the bifurcation model that memory strength is higher after successful retrieval than after restudying, and that retrieval success during practice moderates testing effects?
- Question 1b. What do the neural correlates of retrieval and restudying measured with fMRI reveal about cognitive accounts that retrieval effort, semantic elaboration, and selection processes are crucial for testing effects?

The second aim of this thesis is to investigate the effect of retrieval opportunities during vocabulary learning. Specifically, manipulations of feedback and context that trigger memory retrieval were studied, in order to answer the following question:
Question 2. What is the effect of creating retrieval opportunities during vocabulary exercises on the long-term retention of words?

- Question 2a. What is the effect of feedback that includes a retrieval opportunity compared to standard feedback that presents the answer for restudy?
- Question 2b. What is the effect of a sentence context that creates a need for memory retrieval compared to a context that allows the inferences of word meaning?

Finally, a third aim of this thesis is to come to practical recommendations regarding the use of retrieval practice for vocabulary learning, based on the integration of findings regarding the cognitive mechanisms of retrieval practice, and findings regarding the effect of retrieval manipulations during selected vocabulary exercises with feedback and context.

1.4.2 OUTLINE

The first three chapters of the thesis address the neurocognitive mechanisms of retrieval practice (Question 1). Reaction time measures and functional magnetic resonance imaging were used to complement previous behavioral studies and test the cognitive accounts introduced in this chapter. In **Chapter 2**, reaction time data were used to test the basic assumptions of the bifurcation model that explain empirical findings that benefits of retrieval practice become visible over time. In Chapter 3 and 4, fMRI measurements during learners' practice of words through retrieval and restudying are discussed. **Chapter 3** reports an fMRI study in which brain activations during retrieval and restudying were compared and related to subsequent memory. The data were analyzed against existing cognitive accounts, in particular, elaboration and selection accounts. **Chapter 4** builds on these data by presenting an overview and integration of fMRI studies on testing effects, many of which became available after the data in Chapter 3 had been published. This chapter provides a comprehensive summary of the most recent neuroimaging results on testing effects.

The later chapters of this thesis investigate the effect of memory retrieval during specific vocabulary exercises (Question 2). The experiments reported in **Chapter 5** are focused at the effect of retrieval opportunities during the processing of feedback after errors. Feedback was manipulated in a series of experiments with high school students who practiced foreign language vocabulary with an adaptive computer program in a classroom environment. **Chapter 6** addresses the effect of retrieval when words are presented in context. It tests in particular whether an informative sentence context that allows the inference of word meaning leads to different learning outcomes than an uninformative sentence context that requires the retrieval of word meaning from memory.

The thesis concludes with a summary and discussion of the reported findings, in **Chapter 7**. This last chapter provides an overview of implications for the theoretical understanding of retrieval practice effects and their practical application in vocabulary learning. I also present six practical recommendations for the design of vocabulary exercises with retrieval practice. Open questions for future research are discussed.

1.5 REFERENCES

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Anderson, J. R. (1981). Interference: The relationship between response latency and response accuracy. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 326–343. https://doi.org/10.1037/0278-7393.7.5.326

Baddeley, A. D. (1978). The trouble with levels: A reexamination of Craik and Lockhart's framework for memory research. [Editorial]. *Psychological Review*, *85*(3), 139–152. https://doi.org/10.1037/0033-295X.85.3.139

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

Barcroft, J. (2002). Semantic and structural elaboration in l2 lexical acquisition. *Language Learning*, *52*(2), 323–363. https://doi.org/10.1111/0023-8333.00186

Barcroft, J. (2003). Effects of questions about word meaning during L2 Spanish lexical learning. *The Modern Language Journal*, *87*(4), 546–561. https://doi.org/10.1111/1540-4781.00207

Barcroft, J. (2015). Can retrieval opportunities increase vocabulary learning during reading? *Foreign Language Annals*, *48*(2), 236–249. https://doi.org/10.1111/flan.12139

Beck, I. L., McKeown, M. G., & McCaslin, E. S. (1983). Vocabulary development: All contexts are not created equal. *The Elementary School Journal*, *83*(3), 177–181. https://doi.org/10.1086/461307

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796. https://doi.org/10.1093/cercor/bhp055

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

Blumenfeld, R. S., & Ranganath, C. (2007). Prefrontal cortex and long-term memory encoding: An integrative review of findings from neuropsychology and neuroimaging. *The Neuroscientist*, *13*(3), 280–291. https://doi.org/10.1177/1073858407299290

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604–616. https://doi.org/10.3758/MC.36.3.604

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*(6), 1563–1569. https://doi.org/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*(6), 1547–1552. https://doi.org/10.1037/a0024140

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, *19*(5), 619–636. https://doi.org/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*(2), 268–276. https://doi.org/10.3758/BF03193405

Carpenter, S. K., Sachs, R. E., Martin, B., Schmidt, K., & Looft, R. (2012). Learning new vocabulary in German: The effects of inferring word meanings, type of feedback, and time of test. *Psychonomic Bulletin & Review*, *19*(1), 81–86. https://doi.org/10.3758/s13423-011-0185-7

Carpenter, S. K., & Yeung, K. L. (2017). The role of mediator strength in learning from retrieval. *Journal of Memory and Language*, *92*, 128–141. https://doi.org/10.1016/j.jml.2016.06.008

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

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

Danker, J. F., & Anderson, J. R. (2010). The ghosts of brain states past: Remembering reactivates the brain regions engaged during encoding. *Psychological Bulletin*, *136*(1), 87–102. https://doi.org/10.1037/a0017937

Daselaar, S. M., Prince, S. E., Dennis, N. A., Hayes, S. M., Kim, H., & Cabeza, R. (2009). Posterior midline and ventral parietal activity is associated with retrieval success and encoding failure. *Frontiers in Human Neuroscience*, *3*(13). https://doi.org/10.3389/neuro.09.013.2009

Deconinck, J., Boers, F., & Eyckmans, J. (2015). "Does the form of this word fit its meaning?" The effect of learner-generated mapping elaborations on L2 word recall. *Language Teaching Research*, *21*(1), 31–53. https://doi.org/10.1177/1362168815614048

Dijkstra, T. (2007). The multilingual lexicon. In M. G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics*. [E-reader version]: https://doi.org/10.1093/oxfordhb/9780198568971.001.0001.

Dudai, Y. (2012). The restless engram: Consolidations never end. *Annual Review of Neuroscience*, *35*(1), 227–247. https://doi.org/10.1146/annurev-neuro-062111-150500

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Eysenck, M. W. (1978). Levels of processing: A critique. *British Journal of Psychology*, *69*(2), 157–169. https://doi.org/10.1111/j.2044-8295.1978.tb01643.x

Finley, J. R., Benjamin, A. S., Hays, M. J., Bjork, R. A., & Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language*, *64*(4), 289–298. https://doi.org/10.1016/j.jml.2011.01.006

Finn, B., & Metcalfe, J. (2010). Scaffolding feedback to maximize long-term error correction. *Memory & Cognition*, *38*(7), 951–961. https://doi.org/10.3758/MC.38.7.951

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*(2), 273–293. https://doi.org/10.2307/40264523

Gaskell, M. G. (2007). Statistical and connectionist models of speech perception and word recognition. In M. G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics*. [E-reader version]: https://doi.org/10.1093/oxfordhb/9780198568971.001.0001.

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*(4), 801–812. https://doi.org/10.1037/a0023219

Hall, K. A., Adams, M., & Tardibuono, J. (1968). Gradient- and full-response feedback in computer assisted instruction. *Journal of Educational Research*, *61*(5). Retrieved from http://search.proquest.com/docview/1290534868/citation/795E05A01EF24809PQ/1

Hays, M. J., Kornell, N., & Bjork, R. A. (2010). The costs and benefits of providing feedback during learning. *Psychonomic Bulletin & Review*, *17*(6), 797–801. https://doi.org/10.3758/PBR.17.6.797

Horst, M., Cobb, T., & Meara, P. (1998). Beyond A Clockwork Orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, *11*(2), 207–223.

Hulstijn, J. H. (1992). Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 113–125). London: Maxmillan.

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal*, *80*(3), 327–339. https://doi.org/10.1111/j.1540-4781.1996.tb01614.x

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*(3), 539–558. https://doi.org/10.1111/0023-8333.00164

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review*, *12*(5), 852–857. https://doi.org/10.3758/BF03196776

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: the role of retrievability. *The Quaterly Journal of Experimental Psychology*, *65*(5), 962–975. https://doi.org/10.1080/17470218.2011.638079

Josselyn, S. A., Köhler, S., & Frankland, P. W. (2015). Finding the engram. *Nature Reviews Neuroscience*, *16*(9), 521–534. https://doi.org/10.1038/nrn4000

Kandel, E. R., Dudai, Y., & Mayford, M. R. (2014). The molecular and systems biology of memory. *Cell*, *157*(1), 163–186. https://doi.org/10.1016/j.cell.2014.03.001

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1250–1257. https://doi.org/10.1037/a0023436

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775. https://doi.org/10.1126/science.1199327

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning. *Psychology of Learning and Motivation*, *61*, 237–284. https://doi.org/10.1016/B978-0-12-800283-4.00007-1

Karpicke, J. D., & Roediger, H. L. (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*(4), 704–719. https://doi.org/10.1037/0278-7393.33.4.704

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

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

Kim, H. (2011). Neural activity that predicts subsequent memory and forgetting: A meta-analysis of 74 fMRI studies. *NeuroImage*, *54*(3), 2446–2461. https://doi.org/10.1016/j.neuroimage.2010.09.045

Koenig, M. A., & Woodward, A. (2007). Word learning. In M. Gareth Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics*. [E-reader version]: https://doi.org/10.1093/oxfordhb/9780198568971.001.0001.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. https://doi.org/10.1016/j.jml.2011.04.002

Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Journal*, *73*(4), 440–464. https://doi.org/10.1111/j.1540-4781.1989.tb05325.x

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics:(de) constructing the N400. *Nature Reviews Neuroscience*, *9*, 920–933. https://doi.org/10.1038/nrn2532

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*(1), 1–26. https://doi.org/10.1093/applin/22.1.1

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1787–1794. https://doi.org/10.1037/xlm0000012

Li, L., Sanchez, C. P., Slaughter, B. D., Zhao, Y., Khan, M. R., Unruh, J. R., .. Si, K. (2016). A putative biochemical engram of long-term memory. *Current Biology*, *26*(23), 3143–3156. https://doi.org/10.1016/j.cub.2016.09.054

Lockhart, R. S. (2002). Levels of processing, transfer-appropriate processing, and the concept of robust encoding. *Memory*, *10*(5–6), 397–403. https://doi.org/10.1080/09658210244000225

MacLeod, C. M., & Nelson, T. O. (1984). Response latency and response accuracy as measures of memory. *Acta Psychologica*, *57*(3), 215–235. https://doi.org/doi: 10.1016/0001-6918(84)90032-5

Miller, G. A. (1999). On knowing a word. *Annual Review of Psychology*, *50*(1), 1–19. https://doi.org/10.1146/annurev.psych.50.1.1

Mondria, J.-A. (2003). The effects of inferring, verifying, and memorizing on the retention of L2 word meanings. *Studies in Second Language Acquisition*, *25*(04), 473–499. https://doi.org/https://doi.org/10.1017/S0272263103000202

Mondria, J.-A., & Wit-de Boer, M. (1991). The effects of contextual richness on the guessability and the retention of words in a foreign language. *Applied Linguistics*, *12*(3), 249–267. https://doi.org/10.1093/applin/12.3.249

Nakata, T. (2016). Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. *Studies in Second Language Acquisition*, Advance online publication. https://doi.org/10.1017/S0272263116000280

Nation, I. S. P. (2001a). Knowing a word. In *Learning Vocabulary in Another Language* (pp. 23–59). Cambridge: Cambridge University Press. Retrieved from http://ebooks.cambridge.org/ref/id/CBO9781139524759

Nation, I. S. P. (2001b). Teaching and explaining vocabulary. In *Learning Vocabulary in Another Language* (pp. 60–113). Cambridge: Cambridge University Press. Retrieved from http://ebooks.cambridge.org/ref/id/CBO9781139524759

Nation, I. S. P. (2001c). Vocabulary learning strategies and guessing from context. In *Learning Vocabulary in Another Language* (pp. 217–262). Cambridge: Cambridge University Press. Retrieved from http://ebooks.cambridge.org/ref/id/CBO9781139524759

Nation, I. S. P. (2015). Principles guiding vocabulary learning through extensive reading. *Reading in a Foreign Language*, *27*(1), 136–145.

Nunes, L. D., & Karpicke, J. D. (2015). Retrieval-based learning: research at the interface between cognitive science and education. In R. A. Scott & S. M. Kosslyn (Eds.), *Emerging Trends in the Social and Behavioral Sciences [E-reader version]*. (pp. 1–16). https://doi.org/10.1002/9781118900772.etrds0289

Otten, L. J., Henson, R. N. A., & Rugg, M. D. (2001). Depth of processing effects on neural correlates of memory encoding: Relationship between findings from across- and within-task comparisons. *Brain*, *124*(2), 399–412. https://doi.org/10.1093/brain/124.2.399

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*(2), 101–117. https://doi.org/10.1037/1076-898X.14.2.101

Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language*, *18*(1), 1–28.

Pressley, M., Levin, J. R., & McDaniel, M. A. (1987). Remembering versus inferring what a word means: Mnemonic and contextual approaches. In M. G. McKeown & M. E. Curtis (Eds.), *The Nature of Vocabulary Acquisition* (pp. 107–127). Hillsdale, NJ: Lawrence Erlbaum.

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*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Rissman, J., & Wagner, A. D. (2012). Distributed representations in memory: Insights from functional brain imaging. *Annual Review of Psychology*, *63*(1), 101–128. https://doi.org/10.1146/annurev-psych-120710-100344

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

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

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

Rowland, C. A., & DeLosh, E. L. (2015). Mnemonic benefits of retrieval practice at short retention intervals. *Memory*, *23*(3), 403–419. https://doi.org/10.1080/09658211.2014.889710

Schacter, D. L., Eich, J. E., & Tulving, E. (1978). Richard Semon's theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *17*, 721–744.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363. https://doi.org/10.1177/1362168808089921

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, *19*(1), 17–36.

Seibert, L. C. (1945). A study on the practice of guessing word meanings from a context. *The Modern Language Journal*, *29*(4), 296–322. https://doi.org/10.2307/318219

Sekeres, M. J., Moscovitch, M., & Winocur, G. (2017). Mechanisms of memory consolidation and transformation. In N. Axmacher & B. Rasch (Eds.), *Cognitive Neuroscience of Memory Consolidation* (pp. 17–44). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-45066-7

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An individual's rate of forgetting is stable over time but differs across materials. *Topics in Cognitive Science*, *8*(1), 305–321. https://doi.org/10.1111/tops.12183

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Spaniol, J., Davidson, P. S. R., Kim, A. S. N., Han, H., Moscovitch, M., & Grady, C. L. (2009). Event-related fMRI studies of episodic encoding and retrieval: Meta-analyses using activation likelihood estimation. *Neuropsychologia*, *47*(8–9), 1765–1779. https://doi.org/10.1016/j.neuropsychologia.2009.02.028

Squire, L. R., Genzel, L., Wixted, J. T., & Morris, R. G. (2015). Memory Consolidation. *Cold Spring Harbor Perspectives in Biology*, *7*(8), a021766. https://doi.org/10.1101/cshperspect.a021766

Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. *Memory & Cognition*, *38*(2), 244–253. https://doi.org/10.3758/MC.38.2.244

Swanborn, M. S. L., & de Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, *69*(3), 261–285. https://doi.org/10.2307/1170540

Thomas, R. C., & McDaniel, M. A. (2013). Testing and feedback effects on front-end control over later retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 437–450. https://doi.org/10.1037/a0028886

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, *56*(4), 252–257. https://doi.org/10.1027/1618-3169.56.4.252

Uncapher, M. R., & Wagner, A. D. (2009). Posterior parietal cortex and episodic encoding: Insights from fMRI subsequent memory effects and dual-attention theory. *Neurobiology of Learning and Memory*, *91*(2), 139–154. https://doi.org/10.1016/j.nlm.2008.10.011

van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, *85*(4), 475–511. https://doi.org/10.3102/0034654314564881

van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, *27*(2), 247–264. https://doi.org/10.1007/s10648-015-9310-x

Verspoor, M., & Lowie, W. (2003). Making sense of polysemous words. *Language Learning*, *53*(3), 547–586. https://doi.org/10.1111/1467-9922.00234

Vilberg, K. L., & Rugg, M. D. (2008). Memory retrieval and the parietal cortex: A review of evidence from a dual-process perspective. *Neuropsychologia*, *46*(7), 1787–1799. https://doi.org/10.1016/j.neuropsychologia.2008.01.004

Vilberg, K. L., & Rugg, M. D. (2009). Left parietal cortex is modulated by amount of recollected verbal information. *Neuroreport*, *20*(14), 1295–1299. https://doi.org/10.1097/WNR.0b013e3283306798

vocabulary. (2017). In *Oxford English Dictionaries*. Retrieved from https://en.oxforddictionaries.com/definition/vocabulary

Wagner, A. D., Schacter, D. L., Rotte, M., Koutstaal, W., Maril, A., Dale, A. M., .. Buckner, R. L. (1998). Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science*, *281*(5380), 1188–1191. https://doi.org/10.1126/science.281.5380.1188

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65. https://doi.org/10.1093/applin/aml048

Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language*, *20*(2), 232–245.

Wheeler, M., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*(6), 571–580. https://doi.org/10.1080/09658210244000414

Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*, *43*(7), 1036–1046. https://doi.org/10.1037/xlm0000379

Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., & Anderson, M. C. (2015). Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nature Neuroscience*, *18*(4), 582–589. https://doi.org/10.1038/nn.3973

Wing, E. A., Marsh, E. J., & Cabeza, R. (2013). Neural correlates of retrieval-based memory enhancement: An fMRI study of the testing effect. *Neuropsychologia*, *51*(12), 2360–2370. https://doi.org/10.1016/j.neuropsychologia.2013.04.004

Wixted, J. T., & Rohrer, D. (1993). Proactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(5), 1024–1039. https://doi.org/10.1037/0278-7393.19.5.1024

DO TESTING EFFECTS CHANGE OVER TIME? INSIGHTS FROM IMMEDIATE AND DELAYED RETRIEVAL SPEED

This chapter is based on: van den Broek, G. S. E., Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? Insights from immediate and delayed retrieval speed. *Memory, 22*(7), 803–812. https://doi.org/10.1080/09658211.2013.831455. The authors thank Liesbeth Linssen for statistical support.

Abstract. Retrieving information from memory improves recall accuracy more than continued studying, but this testing effect often only becomes visible over time. In contrast, the present study documents testing effects on recall speed both immediately after practice and after a delay. Forty participants learned the translation of 100 Swahili words and then further restudied the words with translations or retrieved the translations from memory during testing. As in previous experiments, recall accuracy was higher for restudied words than for tested words immediately after practice, but higher for tested words after seven days. Response times for correct answers, however, showed a different result: Learners were faster to recall tested words than restudied words both immediately after practice and after seven days. These results are interpreted in light of recent suggestions that testing selectively strengthens cue-response associations. An additional outcome was that testing effects on recall accuracy were related to perceived retrieval success during practice. When several practice retrievals were successful, testing effects on recall accuracy were significant already immediately after practice. Together with the reaction time data, this supports recent models that attribute changes in testing effects over time to limited item retrievability during practice.

2.1 INTRODUCTION

Numerous studies have documented *testing effects*, i.e., the phenomenon that retrieving information from memory improves the long-term retention of that information more than continued studying (review in Roediger & Butler, 2011). For example, learners benefit less from *restudying* a foreign vocabulary item and its translation than from retrieving the translation from memory like on a *test* (e.g., Carrier & Pashler, 1992; Metcalfe & Kornell, 2007). However, these benefits of testing are often only visible after a delay and not immediately after practice, when outcomes may even be better for restudied materials than for tested materials (for reviews, see Kornell, Bjork, & Garcia, 2011; Roediger & Karpicke, 2006; Toppino & Cohen, 2009). In the present study, we investigated why this is the case by analyzing response times after restudy and testing practice[1], and by relating later recall to judgments of retrieval success during practice.

Although there is a growing literature on the cognitive mechanisms that might underlie testing effects, it is not yet clear why testing effects change over time. For example, a prominent account is that testing improves the efficiency of later recall processes (Karpicke & Smith, 2012), such that relevant information comes to mind earlier and fewer irrelevant associations are activated (Thomas & McDaniel, 2013). The exact mechanisms of this process have not been established yet, but they could involve increased suppression of competing irrelevant information after repeated selection of target-information during testing (M. C. Anderson, Bjork, & Bjork, 1994, 2000). Also, the search set of items treated as candidates in response to retrieval cues could be reduced (Karpicke & Smith, 2012), for example, due to refined mnemonic associations (Pyc & Rawson, 2010) or improved recapitulation of the encoding context (Jacoby, Shimizu, Daniels, & Rhodes, 2005).

Such mechanistic accounts, however, do not readily explain why benefits of testing practice are typically only visible after a delay and not immediately after learning. In the literature, this timing of testing effects is usually discussed in terms of reduced forgetting after testing in comparison to restudying (Carpenter, Pashler, Wixted, & Vul, 2008; Wheeler, Ewers, & Buonanno, 2003) and it has been suggested that the accessibility of items in memory decreases faster for weak (restudied) than for strong (tested) memories (Bjork & Bjork, 1992). The present study investigates an alternative explanation which was recently presented by Halamish and Bjork (2011) and Kornell, Bjork, and Garcia (2011), who suggested that the timing of testing effects can be explained without assuming differences in forgetting rates, referring only to limited retrieval success during testing practice.

1 The terms testing and retrieval practice are used interchangeably in this chapter.

Limited retrieval success during testing practice could explain the timing of testing effects because it leads to a "bifurcation" of items (Kornell et al., 2011, p. 85) into some tested items with high memory strength (those that were successfully retrieved during practice) and some with low memory strength (those that were not retrieved during practice). In contrast, restudying should lead to a comparably large number of items with moderate memory strength, assuming that restudying is not as effective as successful testing but more effective than unsuccessful testing (Kornell et al., 2011). Assuming further that the high memory strength of successfully tested items and the moderate memory strength of restudied items but not the low memory strength of unsuccessfully tested items is sufficient for recall on a later test, the situation can arise that more restudied items than tested items are recalled although the average memory strength of the restudied items is lower than the average memory strength of the successfully tested items (Kornell et al., 2011). However, memory strength decays over time and due to their initially higher memory strength, (successfully) tested items are more likely than restudied items to remain accessible enough for recall over time, leading to higher recall on delayed tests. Note that this explanation of changes in testing effects over time does not require that the memory decay over time is different for tested and restudied items.

The bifurcation model was based on previous studies of testing effects on recall accuracy: The strongest support for the model comes from experiments showing that increasing the difficulty of performance measures can make testing effects visible already immediately after learning, arguing that only (successfully tested) items with high memory strength but not restudied items with moderate memory strength can be recalled on such relatively difficult tests (Halamish & Bjork, 2011). However, in order to directly test the bifurcation model, measures of recall accuracy do not suffice because they only provide information on the outcome of the recall (recalled or not recalled), and not on the difficulty of the recall. In the present study, we therefore measured response times to collect additional information on the difficulty of the retrieval act and on the accessibility of the target information among competing representations in memory, assuming that longer reaction times reflect more difficulty in retrieving information (J. R. Anderson, 1981; MacLeod & Nelson, 1984; Wixted & Rohrer, 1993).

The first purpose of this study was to investigate whether testing practice (in comparison to restudying) influences later retrieval speed at all. The facilitation of later retrieval processes as described by mechanistic accounts of testing effects has so far almost always been measured in terms of the amount of information which the learners recalled but it is likely that recalls also become *faster* if more efficient retrieval routes become available. Although there has been some interest in changes of response times over the course of repeated retrieval practice (e.g., Karpicke & Roediger, 2007), very few studies measured response times *after* restudy and testing

practice. The first study that we found dates back to the 1980s, when MacLeod and Nelson (1984) reported shorter response times but lower recall success immediately after four testing cycles in comparison to three study cycles and one testing cycle. Testing effects on response times did not reach statistical significance in their study, but the authors concluded that accuracy and response times reflect different dimensions of memory, with accuracy depending on whether an item is sufficiently encoded to be retrieved at all, and response times depending on processing steps necessary during retrieval (MacLeod & Nelson, 1984). More support for the relevance of testing effects on response times comes from recent neuroimaging studies of testing effects focusing at its neural correlates, in which significant response time effects were reported as a side result (Keresztes, Kaiser, Kovács, & Racsmány, 2014; van den Broek, Takashima, Segers, Fernández, & Verhoeven, 2013). Therefore, the present study was set up to more systematically investigate whether testing not only improves recall accuracy but also recall speed indicating that testing practice reduces the amount of processing needed for later memory retrieval.

The second purpose of this study was to test the bifurcation explanation of changes in testing effects over time. This was done in two ways. First, we investigated if and how testing effects on response times change over time. The bifurcation model predicts that testing effects on response times should, unlike testing effects on recall accuracy, be visible already immediately after learning and remain visible over time. The reason for this is that the memory strength of those tested items that are successfully retrieved during practice and thus remembered over time should be higher than that of restudied items, both immediately after learning (even when at that moment overall less tested items than restudied items are recalled) and on delayed tests. From this, we derived the hypothesis that both immediately after learning as well as on a delayed test, response times for correctly remembered items would be shorter for tested than for restudied items. A different possible outcome would be that testing effects on response times change over time in a similar way as testing effects on recall accuracy, such that testing only leads to shorter response times after a delay but not immediately after learning. In that case, the data would directly contradict the bifurcation model but be in line with the idea that testing effects only appear after a delay because memory representations of tested items are more resistant to forgetting over time than representations of restudied items (Carpenter et al., 2008; Wheeler et al., 2003).

As a second test of the bifurcation model, we collected judgments of retrieval success during practice to investigate the prediction that testing effects are restricted to items that are successfully retrieved during practice. So far, there has been limited direct research on this topic. In one recent study, Jang and colleagues used an initial test to establish retrievability of items, and then exposed participants to

further restudy and testing practice (Jang, Wixted, Pecher, Zeelenberg, & Huber, 2012). By dividing the data into retrievable and nonretrievable items, they showed that immediate benefits of restudy over testing were almost completely explained by effects on initially nonretrievable items, whereas delayed benefits of testing over restudy were explained fully by testing effects on initially retrievable items. In the present study, we further explored the relation between item retrievability and the timing of testing effects.

2.2 METHOD

2.2.1 PARTICIPANTS

Forty female university students (M_{age}= 19.5 years, SD_{age} = 2.1) from a Psychology Participant Pool took part in the experiment for course credits or a monetary compensation (10 Euro per hour). To increase their motivation, there was an additional bonus of 10 Euro for the 10% best performing participants. Participants reported investing a high amount of mental effort during practice, with an average score of 15.9 (*SD* = 2.6) on a 20-point rating scale (0 = *very low effort*, 20 = *very high effort*). All participants spoke Dutch fluently (88% native speakers), and none of them had prior knowledge of Swahili.

2.2.2 STIMULI

The stimuli were 100 Swahili nouns with Dutch translations, which were pronounceable for Dutch speakers, such as *bustani* (garden), *kaza* (work), *anga* (sky), *samaki* (fish), *jiwe* (stone), *tofaa* (apple).

2.2.3 OVERVIEW OF THE EXPERIMENT

There were two sessions: The first session comprised an initial encoding phase, a practice phase with testing and restudy trials, and an immediate test. The second session seven days later comprised a second test. Session 1 took about 1 hour and 40 minutes; Session 2 took about 20 minutes.

2.2.3.1 ENCODING PHASE. The purpose of the initial encoding phase was to ensure that participants learned the meaning of the majority of the words and to control for item-selection differences between testing and restudy condition, based on a paradigm by Karpicke and Smith (2012). For this purpose, we used an adaptive study program that presented the word-pairs one at a time, in a randomized order, and let participants indicate after each presentation whether they thought they knew the word-pair or not. The presentation of each pair continued until the participants had indicated in two consecutive encoding rounds that they knew the pair. In addition,

all word-pairs were presented one more time at the end of the encoding phase to control for recency effects. The presentation durations for the word-pairs were reduced in steps of 500 ms for every encoding round from 4000 ms in the first round to a minimum duration of 2000 ms. To minimize opportunities for retrieval during the encoding phase, Swahili words were always presented simultaneously with their translation. At the end of encoding, words were randomly assigned to the testing, restudy, or control condition for every participant in such a way that the mean number of presentations during the encoding phase was equal in all conditions (M_T = 4.6, SD_T = 3.1; M_{RS} = 4.6, SD_{RS} = 3.1).

2.2.3.2 PRACTICE PHASE. The critical experimental manipulation took place in the practice phase, when the participants practiced 40 of the 100 previously encoded words in a *testing* condition and 40 of the words in a *restudy* condition. The remaining 20 words served as controls that were not presented during practice. The difference between the conditions was that the complete word-pair was visible on the screen in the *restudy* condition (e.g., *roho - soul*), whereas only the Swahili word was visible in the *testing* condition (e.g., *roho – xxx*). The words were presented for 800 ms before they were replaced by a prompt to make a retrieval success judgment. There were three practice blocks, in which trials were presented in a randomized order. Each block lasted about 9 minutes.

Analysis of perceived retrieval success during practice. To obtain a measure of perceived retrieval success during practice, the participants answered the question "Did you already know the translation?" with "Yes" or "No" after each practice trial. Responses for the three practice rounds were then summarized in five categories: No/No/No (*NNN*), No/No/Yes (*NNY*), No/Yes/Yes (*NYY*), Yes/Yes/Yes (*YYY*), and any other combination in a rest category. For example, NYY indicates words to which participants responded "No" in the first practice block, and "Yes" in the second and third practice block. The words in the "Rest" category (4.8 % of all words) were not included in the analysis reported here, as they form a less interpretable category. However, including them did not change the overall picture of results.

2.2.3.3 IMMEDIATE AND DELAYED TEST. Every participant was tested on a random selection of 10 words from each condition immediately after practice and on the remaining words on a delayed test after seven days. During both tests, the participants saw the Swahili words (one at a time) on a computer screen and entered the Dutch translation with the keyboard. Responses were later categorized as either correct or incorrect. The test program (Inquisit 3.0.4.0 (2009). Seattle, WA: Millisecond Software LLC) recorded how long it took the students to fill in the translation and to click on a button to proceed to the next word, after the Swahili word had appeared on screen. The students received no instruction to respond fast. Only response times for correct responses were included in the following data analyses,

in order to avoid confounding by performance differences between the conditions (correct and incorrect responses often differ in terms of response times (e.g., J. R. Anderson, 1981)). Individual response times that deviated more than three standard deviations from the participant's average response time (these were 1.3 % of all correct responses) were excluded before response times were summarized per participant for further statistical analysis.

2.2.4 DATA ANALYSIS

Data on participant level were subjected to two 3 x 2 repeated measures analyses of variance (ANOVA) with Practice Condition (Test, Restudy, Control) and Testing Moment (immediate, delayed) as within subject factors and Later Recall (i.e., the mean proportion of correctly translated words) or Response Times for correct responses as dependent variables. In a second step, the word-specific data were subjected to a repeated measures logistic regression analysis with SPSS Generalized Estimating Equations function to account for the hierarchical structure of the data (words in participants) (Hanley, Negassa, & Forrester, 2003). We entered Practice Condition (Test, Restudy), Testing Moment (immediate, delayed), and Retrieval Success during Practice (NNN, NNY, NYY, YYY) as predictors and Later Recall Success (correct = 1, not correct = 0) as dependent variable. Note that we could not analyze the relation between retrieval success during practice and response times for correct answers in the same way because there were not enough correct answers for some categories of retrieval success (in particular, very few words were later correctly recalled on the test if they could not be retrieved during practice before). All analyses were performed using SPSS version 15.01.

2.3 RESULTS

2.3.1 RECALL SUCCESS

Table 2.1 contains summary statistics for later recall success (the proportion of correctly translated words) and response times for correct answers. There were significant main effects of Time, $F(1, 39) = 87.94$, $p < .001$, $\eta_p^2 = .69$ and Practice Condition, $F(2, 78) = 15.67$, $p < .001$, $\eta_p^2 = .29$ on Recall Success; as well as an interaction between the two factors, $F(2, 78) = 17.21$, $p < .001$, $\eta_p^2 = .31$. Further investigation of this interaction with t-tests for paired samples revealed a classic testing effect: On the immediate test, performance was significantly better for restudied words than for tested words, $t(39) = -3.58$, $p = .001$, $d = 0.57$, and control words, $t(39) = 3.85$, $p < .001$, $d = 0.61$, whereas the tested and control words did not differ from each other, $t(39) = 0.64$, $p = .53$, $d = 0.10$. On the delayed test after seven days, the effect was reversed: Performance was

significantly better for tested words than for restudied words, $t(39) = 5.57$, $p < .001$, $d = 0.88$, and control words, $t(39) = 7.38$, $p < .001$, $d = 1.17$. Performance was marginally better for restudied words than for control words, $t(39) = 2.015$, $p = .051$, $d = 0.32$.

Table 2.1 Average proportion of Swahili words translated correctly (short: Recall Success) and average response times for correct responses (in ms), per practice condition, as measured immediately and seven days after practice.

		Retrieval		**Restudy**		**Control**	
Testing moment	**Dependent variable**	*M*	*SD*	*M*	*SD*	*M*	*SD*
Immediate	Recall Success	0.69	0.24	0.77	0.22	0.67	0.29
	Response time	4105	897	4654	1313	4826	1332
After 7 days	Recall Success	0.56	0.25	0.45	0.25	0.40	0.24
	Response time	5275	1104	5478	1567	5520	2066

Note. Response times are based on correct responses only.

2.3.2 RESPONSE TIMES

There were significant main effects of Time, $F(1, 34)^2 = 16.33$, $p < .001$, $\eta_p^2 = .32$, and Practice Condition, $F(2, 68) = 6.70$, $p = .002$, $\eta_p^2 = .17$ on Response Times for correct responses, but no interaction between the two factors, $F(2, 68) = 1.08$, $p = .35$, $\eta_p^2 = .031$. The main effect of Time was caused by shorter response times immediately after practice than on the test after seven days. The main effect of Practice Condition was caused by shorter overall response times for tested words (estimated marginal mean: 4690 ms) than for restudied words (estimated marginal mean: 5066 ms), $F(1, 34) = 10.95$, $p = .002$, $\eta_p^2 = .24$, and shorter response times for tested words than for control words (estimated marginal mean: 5173 ms), $F(1, 34) = 11.39$, $p = .002$, $\eta_p^2 = .25$. The difference in response times between control and restudied words was not significant, $F(1, 34) = .47$, $p = .50$, $\eta_p^2 = .01$.

2 Three participants were not included in this analysis because they did not correctly recall any words from the restudy condition on the second test, and therefore had a missing value for response times for correct recalls. Two more participants were excluded because their score on at least one variable was a univariate outlier (z-score > 3.29). Excluding these outlier cases did not change the direction or significance of results.

2.3.3 PERCEIVED RETRIEVAL SUCCESS DURING PRACTICE

To check the reliability of participants' judgments of retrieval success we compared the retrieval success judgments of test items during practice with the recall accuracy of the same items on the immediate test. We found that in case participants indicated during the last practice round that they knew the translation of a word, they correctly recalled the translation on the immediate test a few minutes later in 85.6% (n = 268, retrieval condition) or 81.4% (n = 301, restudy condition) of the cases. This indicates that the retrieval judgments were quite reliable.

To test the research questions related to the bifurcation model, we further investigated the relation between perceived retrieval success during practice and later recall on the two testing moments (see Figure 2.1). The number of words (percentage of all words in parentheses) per retrieval success category were as follows: tested words 221 NNN (15%), 43 NNY (2.9%), 77 NYY (5.2%), and 1133 YYY (76.9%); restudied words 42 NNN words (2.7%), 38 NNY (2.5%), 57 NYY (3.7%), and 1396 YYY words (91.1%). We further investigated these data with repeated measures logistic regression analyses with words within participants as units of analysis. First, we tested a simple model with main effects of Practice Condition, Retrieval Success during Practice, and Time. All main effects were significant, due to, respectively, higher later recall success for tested words than for restudied words, $\chi^2(1) = 64.02$, $p < .001$, higher later recall success when words were retrieved more often during practice, $\chi^2(3) = 125.15$, $p < .001$, and higher recall success on the immediate test than on the test after seven days, $\chi^2(1) = 93.31$, $p < .001$ (a complete overview of B values, SE_B and confidence intervals of the odds ratios can be found in Table 2.2 in the Appendix). In a second step, we added the interaction between Practice Condition and Retrieval Success to the model, which was significant, $\chi^2(3) = 12.03$, $p = .007$ due to the fact that testing effects on Later Recall were significant for the YYY, $\chi^2(1) = 52.70$, $p < .001$, and the NYY words $\chi^2(1) = 22.44$, $p < .001$, but not significant[3] for the NNY, $\chi^2(1) = 0.001$, $p = .97$, or the NNN words, $\chi^2(1) = 0.76$, $p = .38$. The graphs in Figure 2.1 suggest that this effect was more pronounced on the delayed test than on the immediate test, but in subsequent analyses, the 3-way interaction between Practice Condition, Retrieval Success during Practice, and Time was not significant, $\chi^2(1) = 0.49$, $p = .92$.

3 Statistical power to investigate testing effects on the 81 NNY and 263 NNN words was limited due to the relatively small number of words in relation to the very small observed differences in recall success (0.03 and 0.02 respectively). However, performance in these conditions was actually slightly higher for the restudied words than for the tested words. Therefore, it is unlikely that the absence of significant benefits of testing is simply due to a lack of power.

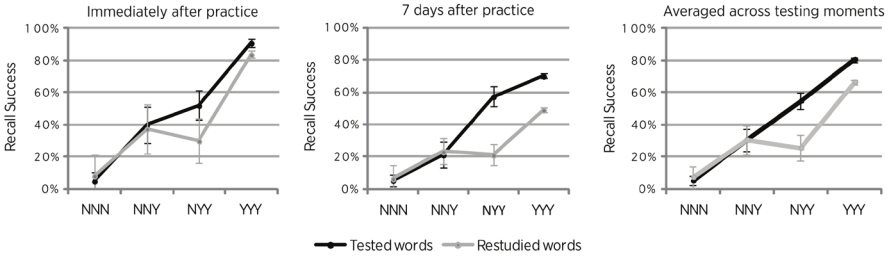


Figure 2.1 Percentage of words recalled correctly on the final performance tests as a function of Practice Condition (test or restudy) and Retrieval Success during the three practice rounds (*NNN* = No/No/No, *NNY* = No/No/Yes, *NYY* = No/Yes/Yes, and *YYY* = Yes/Yes/Yes); shown separately for the immediate test (left graph), the delayed test after 7 days (middle graph), and averaged across the two testing moments (right graph). Error bars denote standard error of the mean.

2.4 DISCUSSION

In the present study, we investigated immediate and delayed effects of successful retrieval during testing practice on later recall accuracy and response times. There were three main results. First, testing improved not only later recall accuracy but also response times in comparison to restudying. Second, the timing of these effects differed: As in previous studies, testing effects on recall accuracy only became visible over time (overviews in Kornell et al., 2011; Roediger & Karpicke, 2006). In contrast, testing effects on response times were visible already immediately after practice as well as after seven days. Third, testing effects on later recall accuracy were related to retrieval success during practice: For those words for which participants indicated that they successfully retrieved the translation in at least two practice rounds, there were testing effects already immediately after practice as well as after seven days. Together, these results indicate that testing improves memory both in terms of later recall success and recall speed but affects only those items that are retrieved successfully during practice. Such limited item retrievability could explain why overall testing effects on recall success only became visible on the delayed test, whereas testing effects on response times for correct answers were already visible immediately after practice.

First, the fact that learners not only recalled *more* tested words than restudied words on the delayed test, but also recalled the tested words *faster*, suggests that successful retrieval practice increases both the chance that information can later be

recalled and the accessibility of that information in memory in terms of processing steps (i.e., time) needed for recall. This interpretation of reaction time results fits well with recent accounts that testing effects could partly be due to increased efficiency of practiced recall processes (Karpicke & Smith, 2012). In terms of the present study, testing may have facilitated the activation of the correct translation in response to the Swahili cue or increased the suppression of incorrect translations. Importantly, this facilitation of later retrieval processes has so far only been measured in terms of the *amount* of recalled information but if testing works by "narrowing the scope of the memory search [during later retrieval] to hone in on targeted information" (Thomas & McDaniel, 2013), a straight forward prediction is that retrieval should also become *faster*. Therefore, the shorter response times that we found after testing than after restudying support mechanistic accounts that explain testing with the (selective) strengthening of cue-response associations.

The present results converge with the few previous studies that reported reaction time outcomes after testing and restudy practice (Keresztes et al., 2014; MacLeod & Nelson, 1984; van den Broek et al., 2013). Note that Keresztes et al. (2014) reported significant testing effects on reaction times both immediately after learning and after a delay of one week, similar to the results reported here, but used onset latencies to measure reaction times whereas submission latencies were used in the present study. The fact that the pattern of results was the same in both studies, suggests that testing effects on response times generalize across different measurements (i.e., onset and submission latencies).

Second, the reported results support the bifurcation model in two ways. First, the pattern of changes over time that we found for response times and recall accuracy support the bifurcation idea that items that are remembered after testing practice may have a higher memory strength than restudied items, even at a moment when the number of recalled tested items is smaller than the number of recalled restudied items (Halamish & Bjork, 2011; Jang et al., 2012; Kornell et al., 2011). Reaction times were shorter for tested words than for restudied words already immediately after learning, although at that moment overall recall was higher for restudied than for tested words. There was, however, no difference in the rate with which response speed decreased over time for restudied and tested materials. Therefore, the present results do not support theories that changes in testing effects over time are due to differences in forgetting rates (Carpenter et al., 2008; Wheeler et al., 2003). Further research is needed with more measurement moments to determine how exactly reaction times change after repeated testing and restudy practice. However, as far as the present study goes, the timing of effects on reaction times can be explained just by referring to limited item retrievability during practice.

The present results also support this bifurcation idea in a second way because when participants indicated that two or three practice retrievals were successful, recall success was better for tested than for restudied items, and this was the case already immediately after learning as well as after seven days. These results are in line with the bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011). However, a replication of the reported results with a more objective measure of retrieval success is desirable because in the present study, the accuracy of judgments could differ for testing and restudy trials (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008), which could partly explain differences in recall success. To control for this potential confound, we repeated our analyses with a measure of word difficulty as covariate (the average recall for each specific word when used as control word) to correct for differences between word difficulty of tested and restudied words within the categories of retrieval success judgments. This analysis again showed that testing led to higher later recall success than restudying at both measurement moments (only) when two or three retrievals were successful during practice. Hence, the conclusion seems warranted that testing without feedback can indeed improve recall success already immediately after learning if several practice retrievals are successful. This is in line with previous studies showing strong benefits of repeated retrieval over a single retrieval opportunity (e.g., Karpicke & Roediger, 2008). However, more research is needed to establish how many successful retrievals are necessary to produce such immediate testing effects.

To conclude, the present study showed that successful retrieval during testing increases not only the amount of information that is remembered over time but also the speed with which that information is accessed. We documented these testing effects on response times at a moment when testing effects on recall success were not yet visible, which supports the idea that limited item retrievability could explain why overall testing effects on recall success only became visible over time. These results open up interesting new possibilities to investigate changes in the accessibility of memories after repeated testing practice even when recall accuracy is at a ceiling level. The reported results further improve insight into the powerful memory-enhancing effects of testing as a tool for learning by measuring not only response accuracy but also response times.

2.5 REFERENCES

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*(7), 861-876. http://dx.doi.org/10.1002/acp.1391

Anderson, J. R. (1981). Interference: The relationship between response latency and response accuracy. *Journal of Experimental Psychology: Human Learning and Memory, 7*(5), 326-343. http://dx.doi.org/10.1037/0278-7393.7.5.326

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1063-1087. http://dx.doi.org/10.1037/0278-7393.20.5.1063

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin & Review, 7*(3), 522-530. http://dx.doi.org/10.3758/BF03214366

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

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

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*(2), 438-448. http://dx.doi.org/10.3758/mc.36.2.438

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*(6), 633-642. http://dx.doi.org/10.3758/BF03202713

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*(4), 801-812. http://dx.doi.org/10.1037/a0023219

Hanley, J. A., Negassa, A., & Forrester, J. E. (2003). Statistical analysis of correlated data using generalized estimating equations: An orientation. *American journal of epidemiology, 157*(4), 364-375. http://dx.doi.org/10.1093/aje/kwf215

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review, 12*(5), 852-857. http://dx.doi.org/10.3758/BF03196776

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability. *The Quaterly Journal of Experimental Psychology., 65*(5), 962-975. http://dx.doi.org/10.1080/17470218.2011.638079

Karpicke, J. D., & Roediger, H. L. (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*(4), 704-719. http://dx.doi.org/10.1037/0278-7393.33.4.704

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966-968. http://dx.doi.org/10.1126/science.1152408

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

Keresztes, A., Kaiser, D., Kovács, G., & Racsmány, M. (2014). Testing promotes long-term learning via stabilizing activation patterns in a large network of brain areas. *Cerebral Cortex, 24*(11), 3025-3035. http://dx.doi.org/10.1093/cercor/bht158

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65*(2), 85-97. http://dx.doi.org/10.1016/j.jml.2011.04.002

MacLeod, C. M., & Nelson, T. O. (1984). Response latency and response accuracy as measures of memory. *Acta Psychologica, 57*(3), 215-235. http://dx.doi.org/10.1016/0001-6918(84)90032-5

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review, 14*(2), 225-229. http://dx.doi.org/10.3758/BF03194056

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*(6002), 335. http://dx.doi.org/10.1126/science.1191465

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20-27. http://dx.doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term memory. *Psychological Science, 17*(3), 249-255. http://dx.doi.org/10.1111/j.1467-9280.2006.01693.x

Thomas, R. C., & McDaniel, M. A. (2013). Testing and feedback effects on front-end control over later retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(2), 437-450. http://dx.doi.org/10.1037/a0028886

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology, 56*(4), 252-257. http://dx.doi.org/10.1027/1618-3169.56.4.252

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

Wheeler, M., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*(6), 571-580. http://dx.doi.org/10.1080/09658210244000414

Wixted, J. T., & Rohrer, D. (1993). Proactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(5), 1024-1039. http://dx.doi.org/10.1037/0278-7393.19.5.1024

2.6 APPENDIX

Table 2.2 Test Statistics for the Logistic Regression Analysis of Word-Level Data of Later Recall (1 = correct, 0 = incorrect) against Practice Condition, Perceived Retrieval Success during Practice, and Testing Moment.

	Model with main effects		**Main effects and Interaction PC x PRSP**	
	B (SE_B)	**OR [95% CI]**	**B (SE_B)**	**OR [95% CI]**
Intercept				
Intercept	-2.51(0.36)	0.08 [0.04, 0.16]	-1.34 (0.48)	0.26 [0.10, 0.67]
Practice Condition (PC)				
Restudy[a]	0	1	0	1
Testing	0.76*** (0.09)	2.13 [1.77, 2.56]	-0.49 (0.50)	0.61 [0.23, 1.62]
Perceived Retrieval Success during Practice (PRSP)	Wald $\chi^2(3) = 125.15$, $p < .001$		Wald $\chi^2(3) = 102.99$, $p < .001$	
No-No-No[a]	0	1	0	1
No-No-Yes	1.86*** (0.31)	6.43 [3.52, 11.75]	1.15** (0.41)	3.15 [1.40, 7.07]
No-Yes-Yes	2.67*** (0.34)	14.50 [7.43, 28.30]	1.04$^{p=.068}$ (0.57)	2.827 [0.93, 8.62]
Yes-Yes-Yes	3.60*** (0.33)	36.56 [19.21, 69.57]	2.47*** (0.48)	11.79 [4.60, 30.26]
Testing Moment (TM)				
Immediate[a]	0	1	0	1
Delayed	-1.29*** (0.13)	0.276 [0.21, 0.36]	-1.29*** (0.13)	0.27 [0.21, 0.36]
Interaction PC x PRSP	NI		Wald $\chi^2(3) = 12.029$, $p = .007$	
Restudy x NNN	NI	NI	0[a]	1
Restudy x NNY	NI	NI	0[a]	1
Restudy x NYY	NI	NI	0[a]	1
Restudy x YYY	NI	NI	0[a]	1
Retrieval x NNN	NI	NI	0[a]	1
Retrieval x NNY	NI	NI	0.51 (0.57)	1.67 [0.54, 5.11]
Retrieval x NYY	NI	NI	2.04** (0.61)	7.71 [2.31, 25.72]
Retrieval x YYY	NI	NI	1.25* (0.50)	3.48 [1.31, 9.24]

	Model with main effects		Main effects and Interaction PC x PRSP	
	B (SE_B)	**OR [95% CI]**	**B (SE_B)**	**OR [95% CI]**
Interaction PC x TM	NI		NI	
Interaction TM x PRSP	NI		NI	
Immediate x NNN	NI	NI	NI	NI
Immediate x NNY	NI	NI	NI	NI
Immediate x NYY	NI	NI	NI	NI
Immediate x YYY	NI	NI	NI	NI
Delayed x NNN	NI	NI	NI	NI
Delayed x NNY	NI	NI	NI	NI
Delayed x NYY	NI	NI	NI	NI
Delayed x YYY	NI	NI	NI	NI
3-way interaction PC x TM x PRSP	NI	NI	NI	NI

Note: OR = odds ratio; CI = confidence interval; NI = not included in model; PC = Practice Condition; PRSP = Perceived retrieval success during the three practice rounds (NNN = No/No/No, NNY = No/No/Yes, NYY = No/Yes/Yes, and YYY = Yes/Yes/Yes); TM = Testing Moment.

[a] set to zero because parameter is redundant.

*** $p < .001$, ** $p < .01$, * $p < .05$.

NEURAL CORRELATES OF TESTING EFFECTS IN VOCABULARY LEARNING

This chapter is based on: van den Broek, G. S. E., Takashima, A., Segers, E., Fernández, G., & Verhoeven, L. (2013). Neural correlates of testing effects in vocabulary learning. NeuroImage, 78, 94–102. https://doi.org/10.1016/j.neuroimage.2013.03.071

Abstract. Tests that require memory retrieval strongly improve long-term retention in comparison to continued studying. For example, once learners know the translation of a word, restudy practice, during which they see the word and its translation again, is less effective than testing practice, during which they see only the word and retrieve the translation from memory. In the present functional magnetic resonance imaging (fMRI) study, we investigated the neuro-cognitive mechanisms underlying this striking testing effect. Twenty-six young adults without prior knowledge of Swahili learned the translation of 100 Swahili words and then further practiced the words in an fMRI scanner by restudying or by testing. Recall of the translations on a final memory test after one week was significantly better and faster for tested words than for restudied words. Brain regions that were more active during testing than during restudying included the left inferior frontal gyrus, ventral striatum, and midbrain areas. Increased activity in left inferior parietal and left middle temporal areas during testing but not during restudying predicted better recall on the final memory test. Together, results suggest that testing may be more beneficial than restudying due to processes related to targeted semantic elaboration and selective strengthening of associations between retrieval cues and relevant responses, and may involve increased effortful cognitive control and modulations of memory through striatal motivation and reward circuits.

3.1 INTRODUCTION

Tests that require memory retrieval improve long-term retention more than continued studying (Roediger & Karpicke, 2006). For example, once learners know the translation of a word, *restudy* practice, during which they see the word and translation again, is less effective than *testing* practice, during which they see only the word and retrieve the translation from memory (Karpicke & Roediger, 2008). This *testing effect* has received much attention from behavioral studies, but its neural correlates are still largely unknown (Roediger & Butler, 2011).

To the best of our knowledge, only two fMRI studies have, so far, explicitly investigated testing effects. Eriksson, Kalpouzos, and Nyberg (2011) scanned participants during a final recall test following prior testing practice, and interpreted correlations between anterior cingulate activation and the amount of prior testing in terms of enhanced memory consolidation. Hashimoto, Usui, Taira, and Kojima (2011) investigated brain activity related to repeated testing and showed both repetition enhancement and attenuation at the final recall. Both of these studies documented facilitated retrieval processes *after* prior testing. In the present study, we took a different approach and investigated the testing practice phase itself. We directly compared the brain activity related to testing and restudying in order to gain insight into the neurocognitive mechanisms by which testing improves memory more than restudying.

Most explanations of testing effects assume that testing improves memory more than restudying because it involves more effortful semantic processing (Roediger & Karpicke, 2006). More specifically, testing is thought to enhance cognitive effort (e.g., Pyc & Rawson, 2009), which is defined somewhat vaguely as an index of the amount of goal-directed, non-automatic processing (Roediger & Butler, 2011). In this context, testing has also been said to constitute a *desirable difficulty* during learning because it increases beneficial deep semantic processing (Bjork & Bjork, 1992). This could lead to a strengthening of the association between retrieval cues and target information and an improved efficiency of search processes during later recall (e.g., Karpicke & Smith, 2012; Karpicke & Zaromb, 2010), such that irrelevant associations are suppressed and target information comes to mind earlier in response to retrieval cues (Thomas & McDaniel, 2012). Alternatively, testing could improve memory because searching for the correct answer during memory retrievals extends semantic networks around the target information with additional associations, thereby increasing the number of available retrieval cues that can lead to later recall (Carpenter, 2009).

Although these explanations of testing effects are rather abstract, some predictions about possible neural substrates can be derived. First, the inferior frontal gyrus (IFG) has consistently been related to controlled, effortful processing during memory retrieval (Race, Kuhl, Badre, & Wagner, 2009). More specifically, IFG is

thought to maintain retrieval plans to favor the activation of relevant information, and to be involved in the selection among competing representations (Badre & Wagner, 2007). Furthermore, IFG activity has been related to semantic processing (Gabrieli et al., 1996; Wagner et al., 1998), during which frontal control processes are thought to act on semantic representations stored in more posterior regions of the brain (Whitney, Kirk, O'Sullivan, Lambon Ralph, & Jefferies, 2011). Although semantic representations are probably distributed across multiple brain areas, a recent meta-analysis of 120 studies suggested that the middle temporal gyrus (MTG) and the inferior parietal lobe (IPL) could function as association areas that integrate different aspects of semantic concepts (Binder, Desai, Graves, & Conant, 2009). More specifically, MTG and IPL seem to mediate the storage and retrieval of word meaning and the integration of information into larger units for semantic processing (Lau, Phillips, & Poeppel, 2008). Therefore, it is likely that the coordinated activity of IFG, MTG, and IPL is involved in testing if effortful, elaborate semantic processing enhances the memory trace.

To test our predictions about neural correlates of testing effects, we collected fMRI data while Dutch participants practiced previously encoded Swahili-Dutch word-pairs by looking at the whole pair (restudying), and while retrieving the translation from memory upon seeing only the Swahili word (testing) (Figure 3.1C). Based on earlier studies (e.g., Roediger & Karpicke, 2006), we expected that testing would lead to better recall than restudying on a later memory test. With respect to brain activity, we derived two hypotheses from the idea that testing increases semantic elaborations and effortful cognitive control: First, we expected higher activity in IFG, IPL, and MTG during testing than during restudying. Second, we expected that activity in these areas during testing and perhaps also during restudying would predict later recall.

3.2 MATERIALS AND METHODS

3.2.1 PARTICIPANTS

Twenty-six female first-year university students (Mage= 19.5 years, SDage = 1.9) participated in the experiment for course credits. The native language of all participants was Dutch and they had no prior knowledge of Swahili. All participants reported that they were right-handed, had normal or corrected-to-normal vision, no neurological or psychiatric history and no language-impairments. The data of 22 participants were included in the analyses; the other four participants were excluded because they had too few trials in specific conditions of interest (i.e., less than ten remembered or less than ten forgotten words). To increase motivation, there was a small financial reward (10 Euro) for the 10% of participants who performed best.

3.2.2 STIMULI

The stimuli were 100 Swahili nouns with their Dutch translation, and 20 control words (also Swahili nouns) of which no translation was given, but which were randomly paired with the Dutch word for "left" or "right". All Swahili words were pronounceable for Dutch native speakers, e.g. "kiti" (chair), "panya" (mouse).

3.2.3 PROCEDURE

The experiment consisted of two sessions, which were both conducted at the same laboratory. Session 1 began with an extensive initial encoding phase, followed by testing and restudy practice in the MR scanner. There was a delay of about fifteen minutes between the initial encoding phase and the practice phase in the scanner, due to preparation of the participants for scanning. Session 2 was conducted one week later, and contained the final memory test (see Figure 3.1A).

3.2.3.1 INITIAL ENCODING. The purpose of the initial encoding phase was to let the participants learn the translations of the 100 Swahili words. For this purpose, they studied the Swahili-Dutch word-pairs at the computer with four different tasks (Figure 3.1 B). Throughout these tasks, the Swahili words were presented simultaneously with their translation to minimize opportunities for retrieval during initial encoding. First, the participants saw all word-pairs once for 8 seconds each and were instructed to think of an association to remember the words. Second, they typed in a short description of each association when cued with the complete word-pairs. Third, the participants practiced with an adaptive computer program that presented the complete word-pairs, one at a time. After each presentation, the participants were asked to make a judgment of learning by pressing a button for either "Yes, I already know the translation" or "No, I don't know the translation yet". Presentations of each word-pair continued until the participants had responded with "Yes" in two consecutive encoding rounds. The number of rounds necessary to learn each word was then used to assign the words to the experimental conditions in such a way that the mean number of rounds during initial encoding was equal for the 50 restudied words and the 50 tested words for each participant. The control word-pairs (Swahili words paired with the word "left" or "right") were presented during the first two encoding rounds and the participants responded by pressing the indicated (left or right) button. The participants were told that they did not have to remember the control words. Fourth, at the end of the encoding phase, all word-pairs were presented one more time and participants again pressed a button to make learning judgments. In total, the initial encoding phase took about 1 hour and 15 minutes, with variations depending on the number of rounds that the participants required to learn each word.

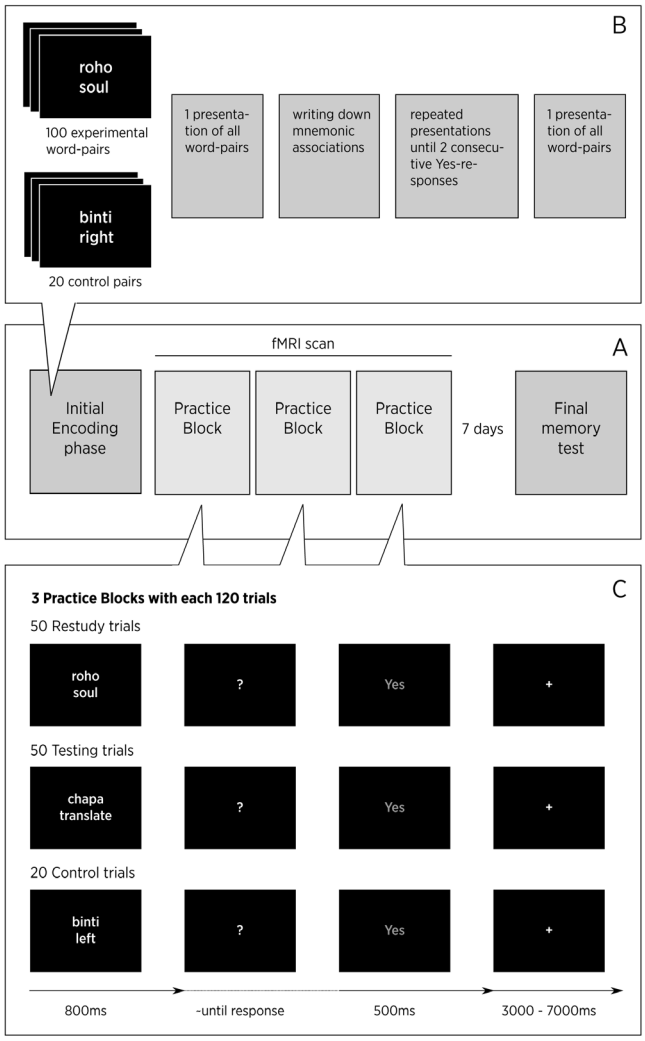


Figure 3.1 Experimental procedure

A) Overview of the complete experiment that consisted of an extensive initial encoding phase before scanning, testing and restudy practice in the MR scanner, and a memory test one week later. B) Overview of the four initial encoding tasks with which the participants studied 100 experimental words and 20 control words. C) Overview of the practice trials in the fMRI scanner. This phase contained the critical experimental manipulation: 50 word-pairs were presented in a testing condition with retrieval opportunity, and 50 word-pairs were presented in a restudy condition. In the response phase of both testing and restudy trials, participants pressed a button to indicate whether they thought that they knew the translation of the Swahili word. The response (Yes or No) was displayed for 500ms. Note that all non-Swahili words were presented in the participants' native language Dutch during the experiment.

3.2.3.2 TESTING AND RESTUDY PRACTICE IN THE FMRI SCANNER. The critical experimental manipulation took place in the fMRI scanner, where the participants practiced 50 words in a testing condition and the other 50 words in a restudy condition. The difference between the conditions was that the complete word-pair was visible on the screen in the *restudy* condition, whereas only the Swahili word was visible in the *testing* condition, together with the word "translate". In both conditions, the participants responded by pressing a button with their left hand to indicate whether or not they knew the translation (See Figure 3.1C for details on the timing of the trials). There was no other overt response. Participants were instructed to do their best to further improve their memory for the presented words during scanning and to devote enough attention to each word to make a good judgment of whether they knew the translation of the word or not. The 20 *control* words were randomly paired with the word "left" or "right" in every practice block, and the participants responded with the left or right button. The participants completed three practice blocks in the fMRI scanner, in each of which they saw all 120 word-pairs once in the assigned condition, in a randomized order. Each practice block took approximately 17 minutes.

3.2.3.3 FINAL MEMORY TEST. Seven days after scanning, the participants took a computerized test, during which they saw the trained Swahili words in a randomized order (one word at a time) and were instructed to type in the Dutch translation. There was no time pressure during responding.

Behavioral data analysis. Responses on the final test were categorized as either correct or incorrect. In addition, response times were obtained by covertly recording how long it took the participants to fill in the translation and click on a button to proceed to the next word, after the Swahili word had appeared on the screen. Only response times for correct responses were analyzed.

3.2.4 MRI DATA ACQUISITION

A 3T MR scanner (Magnetom TIM TRIO, Siemens Medical Systems, Erlangen, Germany) was used to acquire T2*-weighted images of the whole brain with an echo-planar imaging (EPI) sequence (35 slices, slice thickness: 3.0 mm, slice gap: 0.3 mm, ascending slice acquisition, repetition time (TR) = 2.22 s, echo time (TE) = 30 ms, flip angle = 80°, matrix size = 64x64, field of view: 212 mm). In addition, a structural T1-weighted image was obtained using a magnetization-prepared, rapid-acquisition gradient echo sequence (192 slices, slice thickness: 1.0 mm, TR = 2300 ms, TE = 3.03 ms, flip angle = 8°, matrix = 256 x 256, field of view: 256 mm).

3.2.5 MRI DATA ANALYSIS

Image preprocessing and statistical analyses were performed with SPM8 (*Statistical Parametric Mapping;* Wellcome Department of Cognitive Neurology, London, UK; www.fil.ion.ucl.ac.uk) implemented in Matlab 7.11 (MathWorks, Natick, MA).

3.2.5.1 PREPROCESSING. The first five volumes of each participant's functional EPI data were discarded to allow for T1 equilibration. The EPI images were realigned to the participant mean EPI image, which was co-registered to the corresponding structural image. Both functional and structural scans were spatially normalized to a common Montreal Neurological Institute (MNI) reference brain as defined by the SPM8 T1.nii template (resampled at voxel size 2 x 2 x 2 mm), as well as spatially filtered by convolving the functional images with an isotropic three-dimensional (3D) Gaussian kernel (8 mm full width at half maximum). Slow signal drifts were removed with a high-pass filter with a cutoff period of 128 s.

3.2.5.2 STATISTICAL ANALYSES. As a first step, the data were analyzed separately for each participant for each of the three practice blocks. Trials were categorized based on the practice condition (testing, restudy, control) and the result at the final test (LR, LF): Later remembered testing trials (LR_T), later forgotten testing trials (LF_T), later remembered restudy trials (LR_{RS}), later forgotten restudy trials (LF_{RS}), and control trials (C). Only practice trials in which the participants responded with "Yes, I know the translation" were used; trials with the answer "No, I don't remember" were modeled as trials of no interest in a separate sixth category. Neural activations corresponding to the six categories were modeled by separate stick functions, which were time-locked to the presentation of the word-pairs and convolved with a canonical hemodynamic response function and its temporal derivative provided by SPM8, to yield twelve regressors in a general linear model of the BOLD response. The design matrix also included six head motion regressors (three translations and three rotations determined from the realignment step). Parameter estimates were calculated and summarized in contrast images against the control trials: LR_T – C; LF_T – C; LR_{RS} – C; and LF_{RS} – C. In the second step, these single-subject contrast images were included in a group-level ANOVA with the factors Block (1, 2, 3), Practice Condition (Testing, Restudy) and Memory (LR, LF), in which the participants were treated as random factors. For the statistical analyses, we used an uncorrected threshold of $p < .001$ at voxel-level, and applied a threshold of $p < .05$ (family wise error corrected) at the cluster-level (cf., for example, Hayasaka & Nichols, 2003).

3.3 RESULTS

3.3.1 BEHAVIORAL RESULTS

3.3.1.1 INITIAL ENCODING. Prior to scanning, the participants studied all 100 experimental words and translations to the same criterion (see Method section for details). The words were then, for each participant, assigned to the two practice conditions in such a way that the average number of presentations during encoding was identical for the 50 tested and the 50 restudied words (across participants $M_{testing}$= 3.3 (*SD* = 2.11), $M_{restudy}$= 3.3 (*SD* = 2.14)).

3.3.1.2 PRACTICE PHASE IN THE SCANNER. During testing- and restudy practice in the scanner, participants responded with "Yes, I know the translation" to on average 91.1 of the 100 experimental words (the rest of the words were modeled as trials of no interest in the analysis of fMRI data, as described in the Method section).

3.3.1.3 TRANSLATION PERFORMANCE AFTER ONE WEEK (FIGURE 3.2). At the final memory test seven days after practice, participants recalled more translations of the tested words than of the restudied words, $t(21) = 7.436$, $p < .001$, $d = 1.62$. The average performance difference between the two conditions was 8.2%. At the same time, participants were on average 596 ms faster to (correctly) fill in translations of tested words than of restudied words, $t(21) = 3.257$, $p = .004$, $d = 0.71$. When only those words were taken into account to which participants responded "Yes, I know the

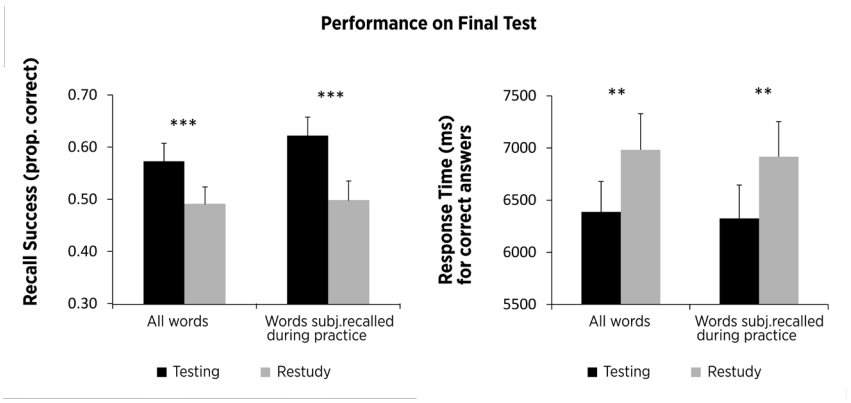


Figure 3.2 Performance on the memory test seven days after testing and restudy practice.
Proportion of words translated correctly and reaction times (for correct responses only) per practice condition, as measured on the final recall test after seven days. Results are displayed separately for all words (the two left bars of each figure) and for those words to which the participants responded "Yes, I know the translation" during practice (the two right bars of each figure). Error bars indicate standard errors of the mean. In all cases, performance was significantly better for the tested than for the restudied words. *** $p < .001$, ** $p < .01$

translation" during practice, the performance difference on the final test increased to 12.3%, $t(21) = 8.682$, $p < .001$, $d = 1.89$, whereas response time differences remained approximately the same ($M_{T\text{-}RS}$ = 593 ms). In sum, behavioral testing effects were large and were found both in terms of the amount of information that was remembered and in terms of response times (Figure 3.2).

3.3.2 NEUROIMAGING RESULTS

3.3.2.1 TESTING VERSUS RESTUDY. To determine which regions were differentially activated during testing and restudying, the two conditions were compared in a factorial design (see Method section for details). As shown in Table 3.1, when trials were combined across the three practice blocks and across levels of subsequent memory, testing engaged a large set of brain areas in comparison to restudying (see also Figure 3.3A). This included bilateral anterior and mid-IFG in pars orbitalis (~BA 47; local maximum (hereafter abbreviated) [-30; 24; 0]) and pars triangularis (~BA 45; [-40; 24; 24]), and the left posterior IFG in pars opercularis (~BA 44; [-42; 4; 34]). Other regions that were more engaged during testing than during restudying included the bilateral ventral striatum [12;10;-2] and midbrain areas [8; -20;-12], left supplementary motor areas [-6;18;50], left middle occipital gyrus [-32;-56;46] and bilateral lingual gyrus [-8;-82;10].

The results for the reversed comparison of restudy over testing trials are reported in Table 3.2. The right MTG [50;-70; 28] and bilateral IPL [-54;-66;42]; the right middle cingulate gyrus [8;-50;40], right middle frontal gyrus [28; 34; 48] and left middle orbital gyrus [-8; 58; 4] were more active during restudying than during testing.

Table 3.1 Brain regions showing more activity during testing than during restudy.

			Local maxima				
Cluster	**Cluster size**	***p***	**Anatomical area**	**x**	**y**	**z**	***t***
1	3382	< .0001	Left Anterior IFG, p. orbitalis (-BA 47)	-30	24	0	7.90
			Left Mid-IFG, p. triangularis (BA 45)	-40	24	24	5.48
			Left Posterior IFG, p. opercularis (BA 44)	-42	4	34	6.80
2	728	.0001	Right Anterior IFG, p. orbitalis (-BA 47)	34	24	-4	7.38
			Right Mid-IFG, p. triangularis (BA 45)	42	20	10	3.47
3	1718	< .0001	Left supplementary motor area	-6	18	50	6.65
			Right middle cingulate	10	20	44	4.75
4	1764	< .0001	Right caudate nucleus	12	10	-2	5.89
			Left putamen	-12	6	-4	5.26
			Left thalamus	-6	-10	6	4.39
			Left midbrain	-8	-20	-12	4.07
			Right midbrain	8	-20	-12	4.06
5	970	< .0001	Left inferior parietal lobe	-32	-56	46	4.96
			Left middle occipital gyrus	-26	-72	42	4.66
6	1073	< .0001	Left lingual gyrus	-8	-82	10	4.48
			Right lingual gyrus	16	-68	8	3.86
			Right calcarine gyrus	12	-76	12	3.83

Note: The table contains all clusters that were significantly more activated during testing than restudying, when combining later remembered and later forgotten trials over the three practice blocks. Statistical tests were performed with an uncorrected threshold of p <. 001 at voxel-level, and a FWE-corrected threshold of p <.05 at cluster-level. The table lists the cluster-size in number of voxels, cluster-level p-value and information about local maxima (anatomical labels, MNI coordinates and t-values). BA = Brodmann area, IFG = inferior frontal gyrus.

Table 3.2 Brain regions showing more activity during restudy than during testing.

			Local maxima				
Cluster	**Cluster size**	**p**	**Anatomical area**	**x**	**y**	**z**	**t**
1	1612	< .001	Right IPL	50	-70	28	4.97
			Right IPL	56	-56	46	3.93
			Right supramarginal gyrus (IPL; ~ BA 40)	54	-46	40	3.76
2	863	< .001	Right middle cingulate gyrus	8	-50	40	4.61
3	711	< .001	Right middle frontal gyrus	28	34	48	4.43
				26	28	42	4.26
4	1103	< .001	Left middle orbital gyrus	-8	58	4	4.39
			Right superior medial gyrus	12	60	10	4.36
			Right superior frontal gyrus	24	62	6	3.68
5	290	.02	Right middle temporal gyrus	64	-16	-14	4.26
6	643	< .001	Left angular gyrus (IPL; ~ BA 39)	-54	-66	42	3.95
				-48	-60	36	3.88
				-58	-64	30	3.70

Note: Structured like Table 1. IPL = Inferior parietal lobe, BA = Brodmann area.

3.3.2.2 PRACTICE EFFECTS. The results on the final memory test seven days after practice were used to categorize the practice trials into practice of later-remembered (LR) and later-forgotten (LF) words, which were then compared to each other to find areas in which activity predicted later memory. Note that we refer to this contrast as "practice effect" to distinguish it from classic subsequent memory effects (e.g., A. S. N. Kim, 2011), which are based on data obtained during a single encoding opportunity and not during additional practice, as in the present study.

For the restudy items, the LR-LF comparison revealed activity in the bilateral rectal gyrus extending to left superior orbital gyrus [-2;40;-2]. For the testing items, a large set of brain areas was predictive of later memory, including the superior medial and superior frontal gyrus [-12;56;8], the left middle cingulate cortex and left precuneus [-6;-56;22], the left and right middle temporal gyrus ([-46;-56;28] and [56;-14;-16]), and the left and right inferior parietal lobe ([-54;-58;30] and [52;-50;40]). The reversed contrast, LF-LR, showed no significant clusters for the restudy items and showed activity in the occipital lobe [-10;78;10] and in the supplementary motor area [-6;8;56] for the testing items.

Differences in Practice effects between Testing and Restudy trials. Practice effects were visible in different areas for the testing and the restudy trials. To test in which brain areas this difference was significant, we calculated interaction effects between practice condition and later memory. The interaction effect showed areas in the supramarginal and angular gyrus in the left IPL ([-56;-52;44] and [-54;-60;44]) and the left MTG [-64;-46;-6] (statistics in Table 3.3, activation map in Figure 3.3B) that were predictive of later memory in the test condition but not in the restudy condition.

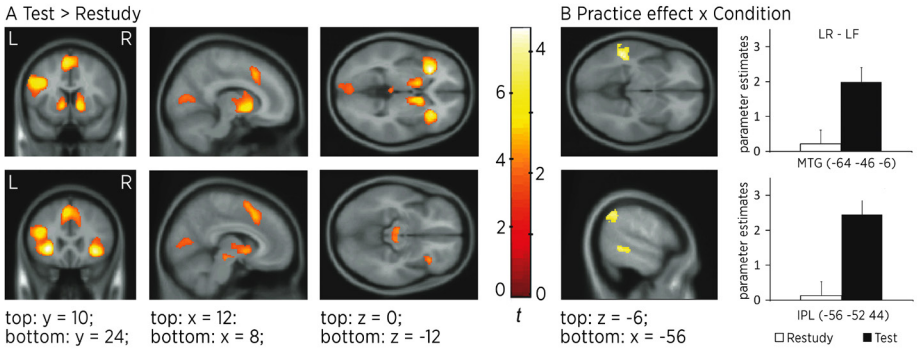


Figure 3.3 Brain activity related to beneficial effects of testing.
A) Clusters that were significantly more activated during testing than during restudying. Color coding as indicated on the left scale of the color map. B) Clusters in left inferior parietal lobe (IPL) and middle temporal gyrus (MTG) that showed an interaction effect between practice condition and later memory. Activity in these regions during testing, but not during restudying was predictive of later memory. Color coding as indicated on the right scale of the color map. Statistical tests were performed with an uncorrected threshold of $p < .001$ at voxel-level, and corrected for multiple comparisons with a FWE-corrected threshold of $p < .05$ at cluster-level. Contrast estimates for the comparison of later remembered (LR) and later forgotten (LF) items are shown for two local maxima, error bars indicate 90% confidence intervals.

Table 3.3 Brain regions showing (A) Practice effect during restudy, i.e., more activity during restudying of words that were later remembered than during restudying of words that were later forgotten; (B) Practice effect during testing; (C) Different practice effects during restudy and during testing.

			Local maxima				
Cluster	**Cluster size**	***p***	**Anatomical area**	**x**	**y**	**z**	***t***
(A) $LR_{RS} > LF_{RS}$							
1	242	.044	Left rectal gyrus	-2	40	-20	4.41
			Right rectal gyrus	6	52	-14	3.70
			Left superior orbital gyrus	-10	52	-14	3.36
(B) $LR_T > LF_T$							
1	5505	< .001	Left superior medial gyrus	-12	56	8	6.51
			Left superior frontal gyrus	-12	56	28	6.14
			Left superior frontal gyrus	-10	52	38	5.67
2	2353	< .001	Left middle temporal gyrus	-46	-56	28	6.06
			Left angular gyrus (IPL)	-54	-58	30	5.80
			Left Inferior Parietal Lobe	-56	-52	44	5.71
3	2926	< .001	Left middle cingulate cortex	-6	-52	40	6.02
			Left precuneus	-6	-56	22	4.85
4	1822	< .001	Right supramarginal gyrus	52	-50	40	5.66
			Right angular gyrus	44	-60	34	5.07
			Right superior temporal gyrus	56	-54	28	4.59
5	1405	< .001	Left middle temporal gyrus	-64	-46	-6	5.37
			Left middle temporal gyrus	-56	-42	-4	5.03
6	625	< .001	Right middle temporal gyrus	56	-14	-16	4.91
			Right inferior temporal gyrus	52	-26	-18	4.42
			Right inferior temporal gyrus	48	-10	-24	3.64
(C) $(LR_T - LF_T) > (LR_{RS} - LF_{RS})$							
1	253	.04	Left middle temporal gyrus	-64	-46	-6	3.49
2	371	.01	Left Supramarginal gyrus (IPL)	-56	-52	44	3.93
			Left Angular gyrus (IPL)	-54	-60	44	3.78
			Left Supramarginal gyrus (IPL)	-56	-50	40	3.70
			Left Angular gyrus (IPL)	-48	-68	44	3.67

Note: Structured like Table 3.1. LR_T = Activity during testing of later remembered items; LF_T = Activity during testing of later forgotten items; LR_{RS} = Activity during restudying of later remembered items; LF_{RS} = Activity during restudying of later forgotten items. For all regions reported in the third section, the difference between later remembered and later forgotten items was larger in the testing than in the restudy condition. The reverse interaction (larger practice effect in restudy than in testing condition) showed no significant clusters. BA = Brodmann area. The supramarginal gyrus and the angular gyrus together form a part of the inferior parietal lobe (IPL).

3.4 DISCUSSION

In this study, we investigated neural correlates of testing effects by comparing testing and restudy practice in an fMRI experiment. Replicating previous behavioral results, delayed recall was better and faster for tested words than for restudied words (e.g., Karpicke & Roediger, 2008; Roediger & Karpicke, 2006). Several areas in the brain were more active during testing than during restudy, including bilateral IFG and striatal areas. Areas that were more active during restudying than testing included the right MTG and bilateral IPL. Further analyses revealed that later memory was predicted by more activity in left MTG and IPL during testing- but not restudying. Together, results show that testing improves memory retention more than restudying and that (1) this practice effect is related to greater activity in IPL and MTG during testing but not during restudy, (2) IFG activity is enhanced during testing in comparison to restudy, and (3) increased activity in striatal and midbrain areas during testing may contribute to memory strengthening.

First, based on the notion that testing effects involve increased semantic elaboration of the connection between words and translations (Carpenter, 2009), we hypothesized that IPL and MTG would be more active during testing than restudying, and that activity in these areas would predict later memory. Results did not support the first prediction. On the contrary, activity in parts of IPL and MTG was higher during restudy than testing. However, the second prediction was partly confirmed: Activity in left IPL and MTG predicted later memory, yet only during testing and not during restudy. These results suggest that activity in IPL and MTG reflects a cognitive function that is important for the beneficial effects of testing but not restudying.

IPL is an association cortex that is engaged in different higher cognitive functions, presumably supporting the integration of complex information and knowledge retrieval (Binder et al., 2009). During semantic elaboration, IPL is thought to integrate semantic information into context and to combine separate concepts into a larger coherent meaning (Lau et al., 2008). Memory studies have related IPL activity to both unsuccessful encoding and successful retrieval (e.g., Daselaar et al., 2009; Uncapher & Wagner, 2009). The relation with unsuccessful encoding has been attributed to increased elaboration of irrelevant information, such as during mind-wandering (Daselaar et al., 2009; A. S. N. Kim, Daselaar, & Cabeza, 2010; Vannini et al., 2011). As a case in point, both the IPL and the MTG have been associated with the so-called default mode network (DMN), a set of brain areas that tends to be activated when thoughts are not focused on a specific task, for example, during rest and self-referential thoughts (e.g., Buckner, Andrews-Hanna, & Schacter, 2008; Mason et al., 2007). Other areas which were more activated during restudying than testing, such as the middle cingulate and medial orbitofrontal cortex, also show overlap with the

DMN, suggesting that some of the activation during restudying could reflect increased task-unrelated semantic processing.

On the other hand, there is an overlap between the DMN and cortical regions that are consistently engaged during successful episodic retrieval together with medial temporal lobe structures (Rugg & Vilberg, 2012). Areas of the DMN, including the angular gyrus in the IPL, tend to show greater activity during the recollection of stronger episodic memories than during familiarity responses to weaker memories (review in H. Kim, 2010). Moreover, the functional connectivity between DMN areas and the left hippocampus appears to increase during successful deep as compared to more shallow encoding, possibly reflecting the encoding of novel episodes into the larger scale self-referential DMN (Schott et al., 2013). Therefore, one interpretation of practice effects in IPL during testing could be the involvement of general recollection networks, an idea that is further supported by studies that link IPL activity to retrieval success: IPL activity increases when more information is retrieved (Vilberg & Rugg, 2008, 2009), feelings of remembering are strong (Wagner, Shannon, Kahn, & Buckner, 2005), or more attention is drawn towards retrieved information (Cabeza, Ciaramelli, Olson, & Moscovitch, 2008; Ciaramelli, Grady, Levine, Ween, & Moscovitch, 2010). So while overall higher activity in IPL during restudying than testing might reflect processes involved in self-referential thought or mind-wandering, the higher IPL activity during testing of later remembered than later forgotten words suggests that differences between retrieved representations predict later memory. Note that we only analyzed trials in which participants indicated that they successfully retrieved a translation – activity is therefore likely to be driven by the amount or quality of the retrieved information and not by mere retrieval success.

The left MTG and neighboring regions are commonly associated with the long-term storage of lexical representations (Hagoort, 2005). Some argue that access to the meaning of words occurs in MTG (Jamal, Piche, Napoliello, Perfetti, & Eden, 2012; Pugh, Sandak, Frost, Moore, & Mencl, 2005), with this area acting as a store of conceptual features of semantic representations or as a hub that connects lexical representations to distributed semantic networks (Lau et al., 2008; Zhuang, Randall, Stamatakis, Marslen-Wilson, & Tyler, 2011). Because activity in MTG predicted later memory only during testing and not during restudying, it seems that only processing of actively retrieved representations predicted later memory whereas processing of representations evoked by passive restudying did not. One possible explanation for this is that testing, more than restudying, activated relevant memory representations that facilitate later access to the translation, for example, mediators that link characteristics of the Swahili word to its translation (Carpenter, 2011; Pyc & Rawson, 2010).

In sum, activity in left IPL and MTG during testing but not during restudying was predictive of later memory. This does not support the idea that semantic processing

is in general enhanced during testing in comparison to restudying, as was put forward in earlier testing effect papers (Carpenter & Delosh, 2006). Instead, results suggest that semantic processing during testing is more beneficial for memory than semantic processing during restudying, possibly because it is more focused on relevant associations. This explanation is in line with recent suggestions that testing improves later recall because it influences the specification of search sets that are activated in response to available retrieval cues, such that relevant target information is activated more effectively (Karpicke & Blunt, 2011; Karpicke & Smith, 2012; Karpicke & Zaromb, 2010). In terms of the present study, testing may have increased the suppression of incorrect translations that would otherwise be activated in response to the Swahili words and/or may have facilitated the activation of the correct translations. This idea that testing facilitated later recall by strengthening the association between the presented Swahili cues and the recalled translations is further supported by the behavioral outcome that tested words were translated significantly faster than restudied words on the final test.

A second major result that supports the conclusion that testing might selectively improve target associations was the enhanced activity in IFG during testing than restudying, which we had predicted based on accounts that mental effort is important for testing effects (e.g., Pyc & Rawson, 2009). IFG has repeatedly been related to intentional, non-automatic processing in memory studies (e.g., Race et al., 2009). During retrieval, IFG is thought to be involved in the controlled access to relevant information in memory and in the selection among competing representations (Badre & Wagner, 2007; Blumenfeld & Ranganath, 2007). Higher activation during testing than during restudying therefore supports the idea that testing involves more intentional, effortful processing than restudying. Possibly, the memory search during testing recruits control-processes in IFG for the activation of and selection among possible translations. Increased effort could also underlie the observed activations in lingual gyrus, which responds to visual processing demands (Mechelli, Humphreys, Mayall, Olson, & Price, 2000) and in supplementary motor areas, which have been linked to effortful word selection processes in language production (Alario, Chainay, Lehericy, & Cohen, 2006).

These results are particularly interesting in light of behavioral findings that testing effects increase with test difficulty: IFG activity during memory retrieval increases when cues are weak (e.g., Crescentini, Shallice, & Macaluso, 2010; Danker, Gunn, & Anderson, 2008), and likewise, behavioral testing effects increase when cues are weak (Carpenter, 2009; Carpenter & Delosh, 2006). Vice versa, IFG activity decreases during repeated retrieval acts (Petersson, Elfgren, & Ingvar, 1999), and likewise, the amount of memory improvement per retrieval act decreases with repetition, especially when the delay between retrievals is short (Pyc & Rawson, 2009). These neural and

behavioral results have both been explained with changing demands on controlled, effortful processing (e.g., Danker et al., 2008; Kelly & Garavan, 2005; Pyc & Rawson, 2009). Interpreting IFG activity in the present study in terms of enhanced cognitive control is thus in line with previous imaging and behavioral studies about repeated testing practice as well as with theoretical claims that testing constitutes a desirable difficulty during learning that improves memory (Bjork & Bjork, 1992).

However, whereas activity in IFG during encoding has consistently been related to effective memory formation, in particular for verbal information (meta-analysis by A. S. N. Kim, 2011), we found only indirect proof of such a relation in this study: IFG activity was higher and later memory was better for tested than for restudied items but there was no direct relation between IFG activity during practice and later memory (i.e., no practice effect). This could be due to the fact that - unlike previous studies - we measured brain activity during additional practice of stimuli that had already been studied extensively before. It is plausible that learners invested more effort to practice words that they found difficult to remember than to practice words that they found easy, which could conceal positive effects of effort on memory if the difficult words were more likely to be forgotten.

In sum, testing increased activity in IFG in comparison to restudying, possibly reflecting higher demands on effortful control processes necessary for the selective activation of the correct translations, but the amount of this processing as such was not predictive of better memory retention.

Additional regions that were involved in testing more than restudying included parts of the midbrain and the ventral striatum. This is interesting because these are key structures of the brain's motivation and reward-system (Shohamy & Adcock, 2010). Dopaminergic neurons that project from tegmental areas in the midbrain to the ventral striatum highlight motivationally significant information (Camara, Rodriguez-Fornells, Ye, & Münte, 2009), and direct attention toward relevant or 'adaptive' information during memory encoding (Wittmann, Schiltz, Boehler, & Düzel, 2008; for a review, see Shohamy & Adcock, 2010; Wittmann et al., 2005). Increased activity in these areas could reflect an additional mechanism by which testing strengthens the memory trace by highlighting information as relevant and enhancing attention. This is in line with speculations that during testing, interactions between the hippocampus and dopaminergic neurons in ventral tegmental midbrain areas could enhance long-term potentiation in the hippocampus and thereby learning (Roediger & Butler, 2011). In addition, genetic determinants of dopamine projections to the prefrontal cortex have been related to retrieval-induced suppression of irrelevant information, which presumably reduces future interference (Wimber et al., 2011). As dopaminergic activations are higher during more effortful tasks, it has been speculated that dopaminergic regions might be involved in a gating mechanism that adjusts the

amount of cognitive resources for the processing of incoming information (Boehler et al., 2011). Involvement of such a gating mechanism would offer a plausible explanation for testing effects from an evolutionary point of view: Information that is readily available in the environment (as during restudying), is likely to remain available in the future. In contrast, information that must be retrieved from memory with effort (as during testing) is likely to cost cognitive capacities again during future retrievals. Therefore, investing resources to better remember tested information is more useful on average than to remember restudied information, because remembering tested information is more likely to reduce future processing costs.

3.4.1 CONCLUSION

We report three major findings on mechanisms potentially underlying testing effects: First, semantic association areas in left IPL and MTG were more active during testing of later remembered than later forgotten words, but showed no such relation to later memory for the restudied items. Activity in these areas might reflect the selective enrichment of semantic associations that improve later access to the target-information during testing. Second, testing increased activity in IFG in comparison to restudying. This supports claims that testing requires more effortful cognitive control than restudying due to the suppression of irrelevant responses and the selective activation of target information. Third, areas in the ventral striatum and midbrain were more active during testing than during restudying, which could reflect activity that supports prefrontal selection processes during memory retrieval as well as motivation and reward circuits that strengthen memory retention. To conclude, the present study improves insight into the neural correlates of testing effects; it thereby adds to explanations of behaviorally established testing effects and further encourages the use of tests in educational practice.

3.5 REFERENCES

Alario, F. X., Chainay, H., Lehericy, S., & Cohen, L. (2006). The role of the supplementary motor area (SMA) in word production. *Brain Research, 1076*(1), 129-143. https://doi.org/10.1016/j.brainres.2005.11.104

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

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex, 19*(12), 2767-2796. https://doi.org/10.1093/cercor/bhp055

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

Blumenfeld, R. S., & Ranganath, C. (2007). Prefrontal cortex and long-term memory encoding: An integrative review of findings from neuropsychology and neuroimaging. *Neuroscientist, 13*(3), 280-291. https://doi.org/10.1177/1073858407299290

Boehler, C. N., Hopf, J.-M., Krebs, R. M., Stoppel, C. M., Schoenfeld, M. A., Heinze, H.-J., & Noesselt, T. (2011). Task-load-dependent activation of dopaminergic midbrain areas in the absence of reward. *The Journal of Neuroscience, 31*(13), 4955-4961. https://doi.org/10.1523/jneurosci.4845-10.2011

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences, 1124*, 1-38. https://doi.org/10.1196/annals.1440.011

Cabeza, R., Ciaramelli, E., Olson, I. R., & Moscovitch, M. (2008). The parietal cortex and episodic memory: an attentional account. *Nature Reviews. Neuroscience, 9*(8), 613-625. https://doi.org/10.1038/nrn2459

Camara, E., Rodriguez-Fornells, A., Ye, Z., & Münte, T. F. (2009). Reward networks in the brain as captured by connectivity measures. *Frontiers in Neuroscience, 3*(3), 350-362. https://doi.org/10.3389/neuro.01.034.2009

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*(6), 1563-1569. https://doi.org/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*(6), 1547-1552. https://doi.org/10.1037/a0024140

Carpenter, S. K., & Delosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory and Cognition, 34*(2), 268-276. https://doi.org/10.3758/BF03193405

Ciaramelli, E., Grady, C., Levine, B., Ween, J., & Moscovitch, M. (2010). Top-down and bottom-up attention to memory are dissociated in posterior parietal cortex: Neuroimaging and neuropsychological evidence. *Journal of Neuroscience, 30*(14), 4943-4956. https://doi.org/10.1523/jneurosci.1209-09.2010

Crescentini, C., Shallice, T., & Macaluso, E. (2010). Item retrieval and competition in noun and verb generation: an fMRI study. *Journal of Cognitive Neuroscience, 22*(6), 1140-1157. https://doi.org/10.1162/jocn.2009.21255

Danker, J. F., Gunn, P., & Anderson, J. R. (2008). A rational account of memory predicts left prefrontal activation during controlled retrieval. *Cerebral Cortex, 18*(11), 2674-2685. https://doi.org/10.1093/cercor/bhn027

Daselaar, S. M., Prince, S. E., Dennis, N. A., Hayes, S. M., Kim, H., & Cabeza, R. (2009). Posterior midline and ventral parietal activity is associated with retrieval success and encoding failure. *Frontiers in Human Neuroscience, 3*(3), 350-362. https://doi.org/10.3389/neuro.09.013.2009

Eriksson, J., Kalpouzos, G., & Nyberg, L. (2011). Rewiring the brain with repeated retrieval: A parametric fMRI study of the testing effect. *Neuroscience Letters, 505*(1), 36-40. https://doi.org/10.1016/j.neulet.2011.08.061

Gabrieli, J. D. E., Desmond, J. E., Demb, J. B., Wagner, A. D., Stone, M. V., Vaidya, C. J., & Glover, G. H. (1996). Functional magnetic resonance imaging of semantic memory processes in the frontal lobes. *Psychological Science, 7*(5), 278-283. https://doi.org/10.1111/j.1467-9280.1996.tb00374.x

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends in Cognitive Sciences, 9*(9), 416-423. https://doi.org/10.1016/j.tics.2005.07.004

Hashimoto, T., Usui, N., Taira, M., & Kojima, S. (2011). Neural enhancement and attenuation induced by repetitive recall. *Neurobiology of Learning and Memory, 96*(2), 143-149. https://doi.org/10.1016/j.nlm.2011.03.008

Hayasaka, S., & Nichols, T. E. (2003). Validating cluster size inference: random field and permutation methods. *Neuroimage, 20*(4), 2343-2356. https://doi.org/10.1016/j.neuroimage.2003.08.003

Jamal, N. I., Piche, A. W., Napoliello, E. M., Perfetti, C. A., & Eden, G. F. (2012). Neural basis of single-word reading in Spanish–English bilinguals. *Human Brain Mapping, 33*(1), 235-245. https://doi.org/10.1002/hbm.21208

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*(6018), 772-775. https://doi.org/10.1126/science.1199327

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966-968.

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

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

Kelly, A., & Garavan, H. (2005). Human functional neuroimaging of brain changes associated with practice. *Cerebral Cortex, 15*(8), 1089. https://doi.org/10.1093/cercor/bhi005

Kim, A. S. N. (2011). Neural activity that predicts subsequent memory and forgetting: A meta-analysis of 74 fMRI studies. *Neuroimage, 54*(3), 2446-2461. https://doi.org/10.1016/j.neuroimage.2010.09.045

Kim, A. S. N., Daselaar, S. M., & Cabeza, R. (2010). Overlapping brain activity between episodic memory encoding and retrieval: Roles of the task-positive and task-negative networks. *Neuroimage, 49*(1), 1045-1054. https://doi.org/10.1016/j.neuroimage.2009.07.058

Kim, H. (2010). Dissociating the roles of the default-mode, dorsal, and ventral networks in episodic memory retrieval. *Neuroimage, 50*(4), 1648-1657. https://doi.org/10.1016/j.neuroimage.2010.01.051

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics:(de) constructing the N400. *Nature Reviews. Neuroscience, 9*(12), 920-933. https://doi.org/10.1038/nrn2532

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., & Macrae, C. N. (2007). Wandering minds: The default network and stimulus-independent thought. *Science, 315*(5810), 393-395. https://doi.org/10.1126/science.1131295

Mechelli, A., Humphreys, G. W., Mayall, K., Olson, A., & Price, C. J. (2000). Differential effects of word length and visual contrast in the fusiform and lingual gyri during reading. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 267*(1455), 1909-1913. https://doi.org/10.1098/rspb.2000.1229

Petersson, K. M., Elfgren, C., & Ingvar, M. (1999). Dynamic changes in the functional anatomy of the human brain during recall of abstract designs related to practice. *Neuropsychologia, 37*(5), 567-587. https://doi.org/10.1016/S0028-3932(98)00152-3

Pugh, K. R., Sandak, R., Frost, S. J., Moore, D., & Mencl, W. E. (2005). Examining reading development and reading disability in English language learners: Potential contributions from functional neuroimaging. *Learning Disabilities Research & Practice, 20*(1), 24-30. https://doi.org/10.1111/j.1540-5826.2005.00117.x

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*(4), 437-447. https://doi.org/10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2010). Why Testing Improves Memory: Mediator Effectiveness Hypothesis. *Science, 330*(6002), 335. https://doi.org/10.1126/science.1191465

Race, E. A., Kuhl, B. A., Badre, D., & Wagner, A. D. (2009). The dynamic interplay between cognitive control and memory. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (4 ed., pp. 705-724). Cambridge, MA: MIT Press.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20-27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term memory. *Psychological Science, 17*(3), 249-255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rugg, M. D., & Vilberg, K. L. (2012). Brain networks underlying episodic memory retrieval. *Current Opinion in Neurobiology*. https://doi.org/10.1016/j.conb.2012.11.005

Schott, B. H., Wüstenberg, T., Wimber, M., Fenker, D. B., Zierhut, K. C., Seidenbecher, C. I., . . . Richardson-Klavehn, A. (2013). The relationship between level of processing and hippocampal–cortical functional connectivity during episodic memory formation in humans. *Human Brain Mapping, 34*(2), 407-424. https://doi.org/10.1002/hbm.21435

Shohamy, D., & Adcock, R. A. (2010). Dopamine and adaptive memory. *Trends in Cognitive Sciences, 14*(10), 464-472.

Thomas, R. C., & McDaniel, M. A. (2012). Testing and feedback effects on front-end control over later retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Advance online publication. https://doi.org/10.1037/a0028886

Uncapher, M. R., & Wagner, A. D. (2009). Posterior parietal cortex and episodic encoding: Insights from fMRI subsequent memory effects and dual-attention theory. *Neurobiology of Learning and Memory, 91*(2), 139-154. https://doi.org/10.1016/j.nlm.2008.10.011

Vannini, P., O'Brien, J., O'Keefe, K., Pihlajamäki, M., LaViolette, P., & Sperling, R. (2011). What goes down must come up: Role of the posteromedial cortices in encoding and retrieval. *Cerebral Cortex, 21*(1), 22-34. https://doi.org/10.1093/cercor/bhq051

Vilberg, K. L., & Rugg, M. D. (2008). Memory retrieval and the parietal cortex: A review of evidence from a dual-process perspective. *Neuropsychologia, 46*(7), 1787-1799. https://doi.org/10.1016/j.neuropsychologia.2008.01.004

Vilberg, K. L., & Rugg, M. D. (2009). Left parietal cortex is modulated by amount of recollected verbal information. *Neuroreport, 20*(14), 1295-1299. https://doi.org/10.1097/WNR.0b013e3283306798

Wagner, A. D., Schacter, D. L., Rotte, M., Koutstaal, W., Maril, A., Dale, A. M., . . . Buckner, R. L. (1998). Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science, 281*(5380), 1188-1191. https://doi.org/10.1126/science.281.5380.1188

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

Whitney, C., Kirk, M., O'Sullivan, J., Lambon Ralph, M. A., & Jefferies, E. (2011). The neural organization of semantic control: TMS evidence for a distributed network in left inferior frontal and posterior middle temporal gyrus. *Cerebral Cortex, 21*, 1066-1075. https://doi.org/10.1093/cercor/bhq180

Wimber, M., Schott, B. H., Wendler, F., Seidenbecher, C. I., Behnisch, G., Macharadze, T., . . . Richardson-Klavehn, A. (2011). Prefrontal dopamine and the dynamic control of human long-term memory. *Translational Psychiatry, 1*, e15. https://doi.org/10.1038/tp.2011.15

Wittmann, B. C., Schiltz, K., Boehler, C. N., & Düzel, E. (2008). Mesolimbic interaction of emotional valence and reward improves memory formation. *Neuropsychologia, 46*(4), 1000-1008. https://doi.org/10.1126/science.281.5380.1188

Wittmann, B. C., Schott, B. H., Guderian, S., Frey, J. U., Heinze, H. J., & Düzel, E. (2005). Reward-related FMRI activation of dopaminergic midbrain is associated with enhanced hippocampus-dependent long-term memory formation. *Neuron, 45*(3), 459-467. https://doi.org/10.1016/j.neuron.2005.01.010

Zhuang, J., Randall, B., Stamatakis, E. A., Marslen-Wilson, W. D., & Tyler, L. K. (2011). The interaction of lexical semantics and cohort competition in spoken word recognition: An fMRI study. *Journal of Cognitive Neuroscience, 23*(12), 3778-3790. https://doi.org/10.1162/jocn_a_00046

NEUROCOGNITIVE MECHANISMS OF THE TESTING EFFECT: A REVIEW

This chapter is based on: van den Broek*, G. S. E., Takashima*, A., Wiklund-Hörnqvist, C., Karlsson Wirebring, L., Segers, E., Verhoeven, L., & Nyberg, L. (2016). Neurocognitive mechanisms of the "testing effect": A review. *Trends in Neuroscience and Education, 5*(2), 52–66. https://doi.org/10.1016/j.tine.2016.05.001 *equal contributions

Abstract. Memory retrieval is an active process that can alter the content and accessibility of stored memories. Of potential relevance for educational practice are findings that memory retrieval fosters better retention than mere studying. This so-called *testing effect* has been demonstrated for different materials and populations, but there is limited consensus on the neurocognitive mechanisms involved. In this review, we relate cognitive accounts of the testing effect to findings from recent brain-imaging studies to identify neurocognitive factors that could explain the testing effect. Results indicate that testing facilitates later performance through several processes, including effects on semantic memory representations, the selective strengthening of relevant associations and inhibition of irrelevant associations, as well as potentiation of subsequent learning.

4.1 MEMORY RETRIEVAL AS AN ACTIVE PROCESS: THE TESTING EFFECT

Memory is typically viewed as a three-step process that begins with the encoding of information, followed by storage and later retrieval of fixed, stable memories. However, this view is incomplete. Retrieval is not a simple read-out process but an *active* process that can change the content and accessibility of memories (Dudai, 2004; Winocur & Moscovitch, 2011). Of particular interest for educational practice is that prompting retrieval with practice-tests enhances the retention of to-be-learned information over time, as shown in studies on the so-called *testing effect*: "taking a test enhances later performance on the material relative to rereading it or to having no re-exposure at all" (Roediger & Butler, 2011, p. 20). Surprisingly, given the plethora of empirical studies demonstrating the testing effect (see Box 4.1), there is still limited knowledge of the specific neurocognitive mechanisms involved. In this review, we relate existing cognitive accounts of the testing effect to findings from recent brain-imaging studies in order to gain a better understanding of the beneficial effects of memory retrieval on the long-term retention of information. In addition to studies on the testing effect, available neuroimaging data for the closely related phenomenon of test-potentiated encoding will also be discussed.

Box 4.1 – Benefits of memory retrieval: a robust phenomenon

The testing effect is a well-investigated phenomenon in cognitive psychology. For a comprehensive review of behavioral studies, readers are referred to published literature overviews (Roediger & Butler, 2011; Roediger & Karpicke, 2006a; Rowland, 2014). Here, we provide a brief introduction to the effect to show that its robustness across different populations, study designs and materials makes it relevant for educational practice.

A typical behavioral testing effect study includes a baseline exposure, followed by either a practice-test or further restudying of the materials, and later a final test to measure learning outcomes (see Figure 4.1A). For example, in a study by Roediger and Karpicke (2006b), students read two prose passages which covered scientific topics, and then restudied one passage and took a practice-test[1] of the other. Learning was assessed five minutes, two days or one week later. Restudying led to better immediate results but practice-testing led to better results on the delayed final tests. This is a common finding

1 In this article, we use the term "practice-testing" when we describe experimental paradigms, to distinguish testing in the practice phase in which learners engage in retrieval, from the final (performance) test used to measure the outcomes of practice. See also Figure 4.1A.

in testing effect studies, which often show that the benefits of practice-tests are stronger when the final test is given after a delay rather than immediately after practice (for further information see Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011; Toppino & Cohen, 2009; van den Broek, Segers, Takashima, & Verhoeven, 2014/Chapter 2).

The testing effect holds in authentic educational settings using course materials

The testing effect has been replicated across different laboratories and also been documented to reliably improve learning outside the laboratory. Studies have demonstrated the testing effect with course materials (Carpenter, 2012; Carpenter, Pashler, & Cepeda, 2009; Leeming, 2002; Lyle & Crawford, 2011; McDaniel, Anderson, Derbish, & Morrisette, 2007; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; Wiklund-Hörnqvist, Jonsson, & Nyberg, 2014) and real university exams (Lyle & Crawford, 2011), using on-line testing (McDaniel et al., 2007), in-class testing (Leeming, 2002; Wiklund-Hörnqvist et al., 2014), and classroom response systems ('clickers') (McDaniel et al., 2013).

The testing effect holds when compared to other pedagogical methods and for different materials

Testing is more beneficial than pedagogical methods such as mind mapping (Karpicke & Blunt, 2011) and group discussions (Stenlund, Jönsson, & Jonsson, 2016), and a better tool for self-study than techniques like reading and highlighting text (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). The effect was documented with different materials, including materials about geography (Rohrer, Taylor, & Sholar, 2010), statistics (Lyle & Crawford, 2011), and medical education (Kromann, Jensen, & Ringsted, 2009).

The testing effect generates transfer of learning

Testing enhances the transfer of learning from specific practice questions to new problems (Butler, 2010; Carpenter & Kelly, 2012; Kang, McDaniel, & Pashler, 2011; Rohrer et al., 2010), and enhances re-learning of information (de Jonge, Tabbers, & Rikers, 2014).

The testing effect is beneficial for different populations

The testing effect has been demonstrated in different age groups, ranging from children (Carpenter et al., 2009; McDaniel et al., 2013; Rohrer et al., 2010) to older adults (Meyer & Logan, 2013). Recently, comparable testing effects have been demonstrated for individuals suffering from traumatic brain injury and healthy control participants (Pastötter, Weber, & Bäuml, 2013).

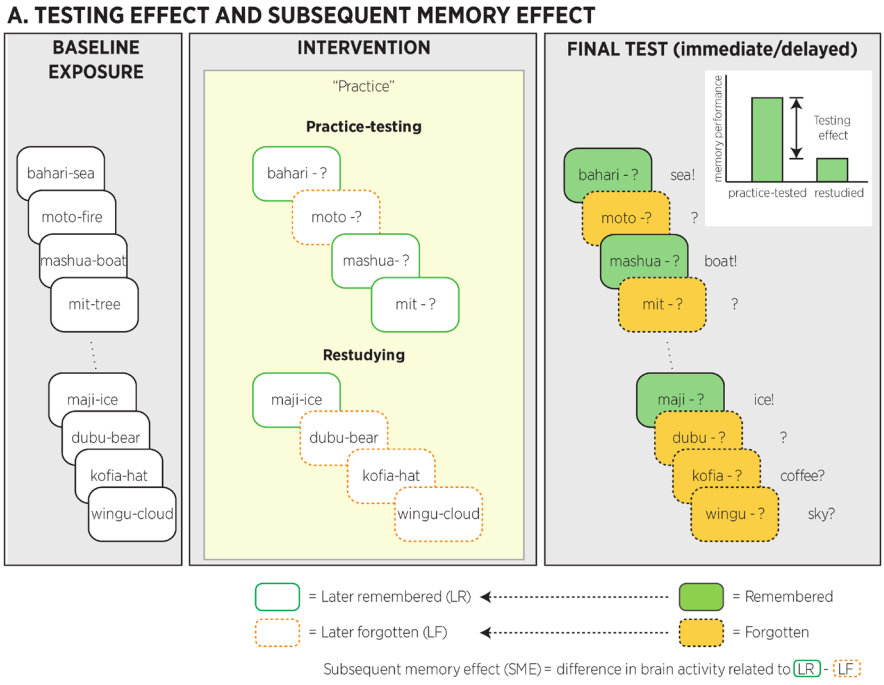


Figure 4.1 A Schematic view of a typical testing effect paradigm, with subsequent memory contrast. The typical set-up of a testing effect experiment is that participants undergo (1) a baseline exposure before (2) the critical intervention period in which they practice the items through restudying (encoding of the complete information) or practice-testing (retrieval of part of the information from memory). There can be several rounds of practice with repeated restudying or practice-testing. (3) To measure the testing effect on memory performance, a final test is administered either immediately after the intervention or after a delay. Functional magnetic resonance imaging (fMRI) analyses: Testing effect fMRI studies have measured changes in brain activation both during the intervention and the final test period. The most common contrasts are between restudied and practice-tested items, and between later remembered and later forgotten items using the so-called subsequent memory effect (SME).

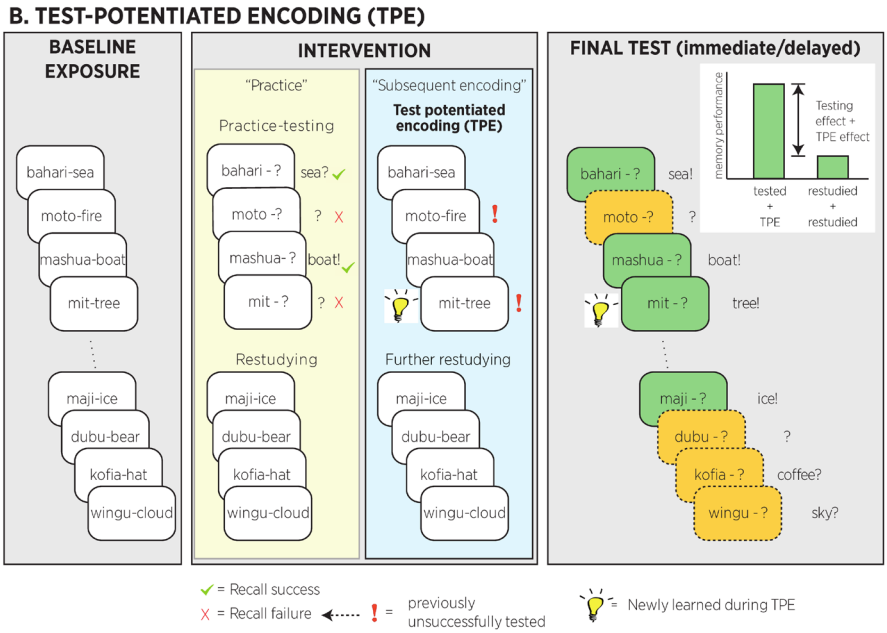


Figure 4.1 B Test-potentiated encoding (TPE) paradigm.
In TPE studies, the intervention period includes practice (practice-testing and restudying) followed by "test-potentiated encoding" to observe the effect of prior practice-testing on subsequent encoding. Brain activations during test-potentiated encoding are often contrasted between items that were previously successfully and unsuccessfully tested, and can also be related to performance on the final test using SMEs as explained in Fig. 1A.

4.2 COGNITIVE PROCESSES UNDERLYING THE TESTING EFFECT

Different ideas have been put forward regarding the cognitive processes underlying the testing effect (Roediger & Butler, 2011; Roediger & Karpicke, 2006a, 2006b). Many of these explanations focus on the way in which testing affects memory representations of the to-be-learned materials. Because most studies on testing effects use verbal materials (e.g., vocabulary or word-pairs), these memory representations are typically conceptualized as (parts of) semantic networks, in which activation spreads among related pieces of information (Rumelhart, McClelland, & PDP Research Group, 1986). Testing is thought to enhance the accessibility of target information by changing the connections within semantic networks, for example, between the representations of two words that are encoded as a word pair (Carpenter, 2009; Carpenter & Delosh, 2006; Karpicke & Zaromb, 2010; Thomas & McDaniel, 2013).

Broadly speaking, two different theories exist about the nature of changes in semantic networks. On the one hand, elaboration accounts suggest that semantic networks become richer through testing because additional associations and alternative retrieval routes are formed (Carpenter, 2009; Carpenter & Delosh, 2006). On the other hand, search-set restriction accounts hold that testing reduces the number of associations that are activated in response to retrieval cues because cue-target associations are selectively strengthened and irrelevant representations are suppressed (Karpicke & Zaromb, 2010; Thomas & McDaniel, 2013).

Carpenter and colleagues introduced the elaboration account of testing based on the assumption that mental elaboration during the search for the correct answer to a test question extends the semantic network of the tested information by creating or strengthening connections with related concepts (Carpenter, 2009; Carpenter & Delosh, 2006). These changes in semantic associations are thought to facilitate later recall by providing additional retrieval routes. Support for such accounts comes from studies showing that practice-tests enhance not only memory for presented information, but also for related semantic information that learners generate to associate cue and target information. For example, participants who studied word-pairs like *Mother:Child*, showed better target recall ("*Child"*) in response to related semantic mediators like "*Father"* after practice-testing (*Mother:* _____) than after restudying (*Mother=Child*) (Carpenter, 2011). In short, representations are thought to get increasingly elaborate with practice-testing so that target information can later be activated through different alternative retrieval routes.

The search-set restriction accounts focus more on the selective nature of retrieval processes during testing, in particular, on the way in which the activation and selection of target information among competing (incorrect) responses influence future retrieval.

One theory is that cue-target associations become selectively strengthened such that the memory search hones in on target information while competing associations are suppressed over the course of repeated testing (Karpicke & Zaromb, 2010; Thomas & McDaniel, 2013). In other words, testing is thought to refine memory representations to selectively strengthen the target response (Karpicke & Blunt, 2011; Karpicke & Smith, 2012; Karpicke & Zaromb, 2010; Thomas & McDaniel, 2013). These ideas have also been linked to the literature on retrieval-induced forgetting. For example, repeatedly retrieving "pineapple" to the cue "fruit- p....?" facilitates the response "pineapple" but inhibits the alternative response "pear" (Murayama, Miyatsu, Buchli, & Storm, 2014; Storm & Levy, 2012). Selective retrieval, thus, seems to strengthen target responses while inhibiting related but undesired responses.

Recently, Karpicke, Lehman, and Aue (2014) presented a possible mechanism that could underlie the selection processes during repeated testing. According to their "episodic context account", items become associated with the episodic context in which they are studied. During retrieval, the context from earlier presentations is re-activated and becomes integrated with current contextual information. This refines the context representation associated with an item because those contextual features that serve as effective retrieval cues are strengthened the most. As a result, the search set of candidates that are activated increasingly zooms in on the target response, while competing responses are suppressed.

The elaboration and search-set restriction accounts are both compatible with another popular account in the cognitive literature, namely that practice-testing is a more effortful process than restudying (Karpicke et al., 2014). The amount of effort during practice-testing has been related to the size of testing effects, with more difficult tests producing better outcomes than easier tests (Pyc & Rawson, 2009, 2010). Mental effort is thus likely to be important for testing, but the definition and interpretation of the term "effort" is complex. Roediger and Butler called it "an index of the amount of reprocessing of the memory trace that occurs during retrieval" (2011, p. 5). Hence, the term is only vaguely defined and more effort could reflect both more elaboration and an increasing number of available retrieval routes (following the elaboration account), or higher selection demands and more suppression of competing incorrect responses (following the search-set restriction account). Whichever process underlies the effort during testing, both accounts predict that practice-testing changes cue-target associations in such a way that future memory retrieval is facilitated. With repetition, demands on effortful retrieval processes are thought to decrease because the target information becomes increasingly accessible through changes in the available retrieval routes. This facilitation of memory retrieval is then thought to lead to better long-term performance because later performance tests (final test) involves similar retrieval processes as those engaged during the practice-testing, and

because a greater overlap of the cognitive processes involved in a final performance test with the cognitive processes engaged in practice-testing is thought to enhance performance (an idea known as *transfer-appropriate processing* (Lockhart, 2002)).

In addition to direct benefits of testing, indirect effects of testing on subsequent learning are also highly relevant for educational practice. Testing enhances the efficiency of subsequent encoding in comparison to pure restudy conditions (Grimaldi & Karpicke, 2012; Izawa, 1971; Richland, Kornell, & Kao, 2009), which is known as *test-potentiated encoding* (TPE). Explanations of TPE broadly fall into two categories. First, testing could enable the learner to better discriminate between information that is successfully retrieved on the test (and thus likely already well learned) and information that is not retrieved (and thus needs further studying) (e.g., Roediger & Karpicke, 2006a, 2006b). In this way, testing enables a more efficient focus of attention during subsequent encoding, and the refinement of learning strategies. For example, mnemonic mediators that link cue and target information could become refined over the course of repeated testing (Pyc & Rawson, 2012). A second explanation is that the testing context is reactivated during TPE when learners are reminded of prior testing during the encoding opportunity. This could enrich the memory representation and create additional retrieval cues when the different testing contexts become integrated (Nelson, Arnold, Gilmore, & McDermott, 2013; Vestergren & Nyberg, 2014).

In summary, cognitive explanations of the testing effects involve three lines: 1) memory representations change due to elaboration of relevant and/or suppression of irrelevant associations, 2) testing is an effortful process that becomes facilitated by repetition, and 3) testing leads to more efficient subsequent learning (TPE). In the following, we will relate these ideas to available neuroimaging studies.

4.3 NEURAL CORRELATES OF THE TESTING EFFECT

Although the accounts outlined above cannot be directly translated into predictions about brain activity, results from functional magnetic resonance imaging (fMRI) studies can be related to the broad cognitive processes addressed in the accounts (see Box 4.2 and Figure 4.2 for more information).

To date, ten fMRI studies have investigated beneficial effects of practice-testing for learning (listed in the Appendix). Four of these studies measured classic testing effects by comparing performance after restudying and practice-testing[2] and related differences in brain activation during restudying and practice-testing to later performance (Rosner, Elman, & Shimamura, 2013; van den Broek, Takashima, Segers, Fernández, & Verhoeven, 2013/Chapter 3; Vannest et al., 2012; Wing, Marsh, & Cabeza, 2013)[3]. Four other studies investigated the consequences of practice-testing by measuring changes in brain activation during repeated practice-testing only, or *after* practice on a final test (Eriksson, Kalpouzos, & Nyberg, 2011; Karlsson Wirebring et al., 2015; Keresztes, Kaiser, Kovács, & Racsmány, 2014). Two of these studies focused on TPE (Nelson et al., 2013; Vestergren & Nyberg, 2014). Liu, Liang, Li, and Reder (2014) also analyzed brain activations during restudying directly following practice-testing, and measured brain activation patterns during interleaved restudying and practice-testing. A schematic overview of the experimental paradigms employed in the reviewed studies and the terms that are used in this article to describe them are presented in Figure 4.1. More detailed descriptions of the paradigms of the studies can be found in the Appendix. In the following, we review the neuroimaging studies in light of available cognitive accounts to summarize how reports of brain activity related to practice-testing can tentatively inform ideas from the cognitive literature on testing effects. In the course of reviewing, some other literature related to, but not directly focusing on testing effects, will also be discussed.

2 These include two studies (Vannest et al., 2012; Rosner et al., 2014) in which participants either read a pair of semantically related words or generated a word when presented with a cue word and the initial letter of a semantically associated word (e.g., salt - p*****). We do not distinguish between such retrieval of pre-existing semantic associations and the retrieval of recently learned associations as employed in the other testing effect studies in this review, because both involve the activation of target information in response to a cue, focus processing on cue-target relations, and can be influenced in a similar way by certain experimental manipulations (cf. Peterson & Mulligan, 2013). Therefore, we refer to the "read" and "generate" conditions as "restudying" and "practice-testing" in the main text. However, because some authors argue that the generation of semantic associations is qualitatively different from the retrieval of recently learned episodic associations (e.g., Karpicke & Zaromb, 2010), the type of retrieval is mentioned in the text if results differ between the two types of studies.

3 The van den Broek et al. (2013) study has been included in this thesis as Chapter 3.

Box 4.2. Candidate brain regions of potential importance for the testing effect

Cognitive accounts of the testing effect suggest that practice-testing compared to restudying leads to a) changed *memory representations*, and b) facilitated *selective retrieval*. Here we discuss which brain regions are likely to support these cognitive processes. All anatomical areas that are mentioned can be found in Figure 4.2.

Semantic memory representations

It is well established that patterns of brain activity during memory retrieval partly overlap with those seen during initial encoding (Nyberg, Habib, McIntosh, & Tulving, 2000; Nyberg et al., 2001; Wheeler, Petersen, & Buckner, 2000). Thus, visually encoded material will evoke visual processing areas, and auditory memory will evoke auditory processing areas at retrieval, even when memory is probed with a different sensory input. For encoding and retrieval of semantic memory, as used in all fMRI studies on testing effects, it is likely that brain areas related to semantic processing will be involved. Semantic representations contain multi-sensory information widely distributed across cortical and subcortical regions of the brain (Binder & Desai, 2011; Binder, Desai, Graves, & Conant, 2009; Martin, 2007). Candidate brain regions for semantic representations include posterior brain areas such as the lateral temporal cortex, and the inferior parietal lobe (IPL) including the supramarginal (SMG) and the angular (AG) gyri (see Figure 4.2). These are involved in the storage and conceptual integration of semantic representations (Binder & Desai, 2011; Price, 2012). The inferior (ITG) and the middle (MTG), as well as the anterior temporal lobe, are thought to integrate different sensory inputs into multi-modal representations and are therefore expected to be active during concept retrieval (Binder et al., 2009; Patterson, Nestor, & Rogers, 2007; Whitney, Jefferies, & Kircher, 2011). The AG in the IPL is a higher-order associative area thought to integrate different components of semantic concepts into a coherent meaning (Binder & Desai, 2011; Binder et al., 2009; Binder et al., 2003; Graves, Desai, Humphries, Seidenberg, & Binder, 2010; Patterson et al., 2007; Whitney et al., 2011).

Selective retrieval of semantic memory

In order to retrieve target representations from memory, processes related to control and monitoring are necessary. The ventral lateral prefrontal cortex (VLPFC) has been discussed as one key brain region underlying cognitive control during memory retrieval (Badre & Wagner, 2007; Race, Kuhl, Badre, & Wagner, 2009), and is thought to direct attention to goal-relevant information and inhibit irrelevant information during selective retrieval (Blumenfeld & Ranganath, 2007; Storm & Levy, 2012). The repeated selection of target information and inhibition of competing information during testing that is thought to facilitate later retrieval, are likely mediated by this area. This is supported by the involvement of the VLPFC during selective retrieval that weakens competing memories: VLPFC activity increases during retrieval when there is competition between several responses and decreases when a specific memory trace is selectively strengthened (Wimber et al., 2008).

The anterior cingulate cortex (ACC) and the dorsal lateral prefrontal cortex (DLPFC) are other brain areas that are activated when demands on cognitive control are high and are suggested to be a part of the attention-control network. The ACC is thought to detect conflicts in information proces-sing and play a role in outcome evaluation and decision making (Botvinick, 2007). The DLPFC is thought to play a role in directing attention to task-relevant representations (Kuhl, Dudukovic, Kahn, & Wagner, 2007), selectively mediate the resolution of response conflicts (Badre & Wagner, 2004), and to become deactivated with practice as a function of decreased demands on selection (Kuhl et al., 2007).

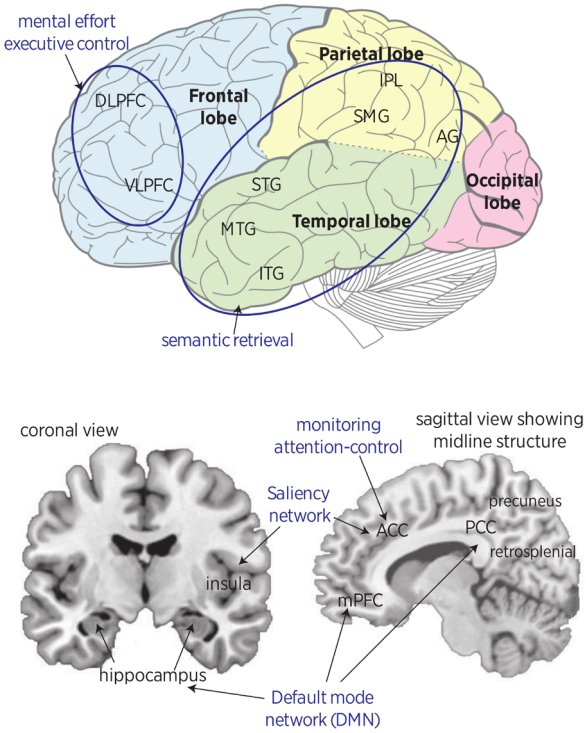


Figure 4.2 Key anatomical areas of the brain for the testing effect.
The upper panel shows a lateral view of the brain. The temporal lobe (superior temporal gyrus STG; middle temporal gyrus MTG; inferior temporal gyrus ITG), and the inferior parietal lobe (IPL) including the supramarginal (SMG) and the angular (AG) gyri, have often been found to increase in activation during semantic tasks (see Box 4.2). The dorsal lateral prefrontal cortex (DLPFC) and the ventral lateral prefrontal cortex (VLPFC) have been found to increase in activation when tasks require mental effort and executive control, such as selective memory retrieval (see Box 4.2). The lower panel shows coronal (left) and sagittal midline (right) sections of the brain. The anterior cingulate cortex (ACC) has been related to attention control and conflict monitoring. Furthermore, this area together with the anterior insula is a part of the Saliency network (see Box 4.3). The medial prefrontal cortex (mPFC), the posterior cingulate cortex (PCC), the retrosplenial, the precuneus and the hippocampus along with the inferior parietal lobe (IPL) and lateral temporal cortices have been associated with the Default mode network (DMN; see Box 4.3). The figure in the upper panel was adapted from Carter (1918, public domain). The figure in the lower panel is based on a template provided in MRI-cron (Rorden & Brett, 2000), from „Enhancement of MR Images Using Registration for Signal Averaging“, by C.J. Holmes, R. Hoge, L. Collins, R. Woods, A.W. Toga, & A.C. Evans, Journal of Computer Assisted Tomography, 22(2), 324-333. Copyright 1993–2009 by Louis Collins, McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University. Adapted with permission.

4.3.1 HOW DOES TESTING AFFECT MEMORY REPRESENTATIONS?

The cognitive account that memory traces are semantically elaborated through testing predicts greater activation of semantic representations during practice-testing than restudying. If this is the case, we would expect greater involvement of brain areas related to semantic memory retrieval, such as the left temporo-parietal areas (see Box 4.2 and Figure 4.2 for candidate brain structures) during practice-testing than restudying. Furthermore, the activation in these areas would be expected to increase with repetition. If, on the other hand, search-set restriction underlies testing effects, then one would expect that retrieval leads to less activation with each repetition.

The four studies that directly compared brain activations during practice-testing and restudying (see Appendix) reported *lower* activation in putative semantic memory storage areas during practice-testing relative to restudying. Wing and colleagues Wing et al. (2013) had participants practice English word-pairs in an MRI scanner. After baseline exposure, half of the pairs were practiced through testing (cued recall of the second word) and half through restudying (re-exposure to the complete pair). Practice-testing led to significantly better memory performance than restudying one day after practice. However, brain activity was lower during practice-testing than restudying in the left superior temporal gyrus (STG) and in the bilateral middle temporal gyrus (MTG). Similarly, van den Broek et al. (2013) observed less activation in the right MTG during practice-testing than restudying. Moreover, brain activity was lower during practice-testing than restudying in the inferior parietal lobe (IPL), including the supramarginal gyrus (SMG), and the angular gyrus (AG) in three studies (Rosner et al., 2013; van den Broek et al., 2013; Wing et al., 2013), although in one study, in which participants retrieved semantically associated words from prior knowledge, activation in bilateral SMG/AG was *higher* during practice-testing than during restudying (Vannest et al., 2012).

In order to understand which processes contribute to the testing effect, the relation between brain activity during practice and later performance is critical. In brain imaging studies, differences in brain activation during studying of items that are later remembered in contrast to later forgotten items are often taken as evidence for successful encoding and termed *subsequent memory effect* (SME; Sanquist, Rohrbaugh, Syndulko, & Lindsley, 1980; see also Figure 4.1A). Within the putative semantic memory representation areas, increased activation predictive of later successful memory performance was found in the bilateral AG and SMG (van den Broek et al., 2013), the right (van den Broek et al., 2013) or the left (Wing et al., 2013) inferior temporal gyrus (ITG), the left MTG (Liu et al., 2014; van den Broek et al., 2013), the left STG (Liu et al., 2014; Vannest et al., 2012) and the right STG (Liu

et al., 2014; van den Broek et al., 2013; Wing et al., 2013). However, these effects differed between practice-testing and restudying. Whereas activity was consistently *higher* during practice-testing of items that were later remembered than for items that were later forgotten (Karlsson Wirebring et al., 2015; Liu et al., 2014; van den Broek et al., 2013; Vannest et al., 2012; Wing et al., 2013), no difference (van den Broek et al., 2013; Vannest et al., 2012) or even a reversed effect (Wing et al., 2013) was found during restudying. Taken together, these results suggest that the engagement of semantic memory storage areas during practice-testing is different from that during restudying. Tentatively, activity in these areas appears to be higher during restudying than practice-testing, but only predictive of later performance when measured during practice-testing.

In a different type of study relevant to the question how practice-testing affects memory representations, Keresztes et al. (2014) reported brain activity on an immediate and a delayed final test after restudying or practice-testing. They found that activity in the IPL decreased from the immediate to the delayed test for restudied items but not for items that were tested during practice. Since one hypothesis concerning the level of activation in the IPL is that it might reflect the strength or richness of retrieved representations (Reas & Brewer, 2012; Rugg & Vilberg, 2013), the comparably stable activation level over time for tested items might indicate that practice-testing made memory traces more resistant to decay than restudying.

More direct investigations of changes in memory representations over time have recently become available from two publications that apply multivariate analysis techniques to the study of repeated retrieval (Karlsson Wirebring et al., 2015; Wimber, Alink, Charest, Kriegeskorte, & Anderson, 2015). Karlsson Wirebring et al. (2015) used *representational similarity analysis* (RSA; Kriegeskorte, Mur, & Bandettini, 2008) to correlate patterns of brain activation between multiple trials to determine under which conditions brain activations are more similar or dissimilar from each other. The study focused on successful repeated retrieval across three consecutive practice-tests, and related patterns of brain activations during practice-testing to performance at final test one week later (see Table 4.1 in the appendix). Remembered items elicited higher activity in different areas of the brain, including the left lateral temporal cortex and bilateral posterior parietal cortices one week after practice. Notably, the same right parietal region identified as important for retrieval success one week after practice showed higher BOLD activity already at the day of practice for items that were later successfully remembered. An RSA in this region revealed that activation patterns were less correlated over the three consecutive tests for items later remembered compared to those forgotten. This can be interpreted as a sign of more altered or elaborated semantic representations during repeated practice-testing for items later remembered, which tentatively supports the idea that semantic elaboration is one key mechanism fostering long-term retention after repeated practice-testing.

So far, there has not been a study that compared changes in brain activations during repeated testing to changes during repeated studying. However, Xue et al. (2010) investigated brain activations during repeated studying and it is interesting to note that unlike Karlsson Wirebring et al. (2015), they reported that a *greater* similarity in patterns of neural activation across study trials predicted better memory performance. Xue et al. concluded that items for which repeated studying leads to a consistent neural representation are better remembered. The different conclusions from these two studies suggest that the effect of variations of the amount of semantic activation across repetitions might be different during testing (Karlsson Wirebring et al., 2015) and studying (Xue et al., 2010). The role of semantic elaboration in explaining the testing effect thus needs further exploration.

The second report of fMRI measurements during repeated retrieval is available from a study that focused on retrieval-induced forgetting (Wimber et al., 2015). The authors employed what they call a *canonical template tracking method* to determine the activation state of target memories and competing information during testing. Participants initially learned to associate one cue word (e.g., *sand*) with two different pictures (e.g., *sand – Marilyn Monroe, sand – hat*) and were then repeatedly instructed to retrieve only one of the pictures (e.g., *Marilyn Monroe*). The authors investigated the overlap between brain activations during this selective retrieval and template brain responses to the target (*Marilyn Monroe*) and competitor (*hat*) information in order to infer to what extent the target and competitor information were activated. Results suggest that while participants initially tended to activate both target and competitor, the competitor was progressively suppressed over the course of repeated testing. Moreover, target information was reinstated increasingly over repetitions. Although this study does not directly address testing effects, it is informative for the evaluation of search-set restriction accounts of the testing effect because it shows that selective retrieval can lead to the suppression of competing information while strengthening target information. However, it is unclear how such competitor suppression is related to the retention of target information because the study focused largely on competitor suppression rather than target enhancement.

In summary, neural responses in brain areas related to semantic memory retrieval lead to mixed conclusions about the competing cognitive models of testing effects. Overall, the available studies indicate a different role for temporo-parietal regions during practice-testing and restudying. Tentatively, activity in these areas appears to be higher during restudying than practice-testing, but only predictive of later performance when measured during practice-testing. Practice-testing, if successful, seems to strengthen the neural representations of target information in temporo-parietal regions. Restudying on the other hand seems to evoke semantic information that is less relevant for learning, possibly related to mind-wandering

(see also Box 4.3). The question of whether semantic networks become elaborated or restricted over the course of repeated practice-testing remains open. The results from Karlsson Wirebring et al. (2015) suggest that successful repeated retrieval that fosters long term retention might be characterized by semantic elaboration, as indicated by reduced pattern similarity within the parietal lobe for items subsequently remembered. This supports the idea that semantic elaboration might underlie the benefits of testing (Carpenter, 2009; Carpenter & Delosh, 2006). However, in an earlier study (Xue et al., 2010), reduced pattern similarity during repeated studying predicted *worse* rather than better outcomes. Variations in semantic activation during repeated practice may thus not always be beneficial, and may have different effects during practice-testing and restudying. Regarding search-set restriction accounts, one study on retrieval-induced forgetting (Wimber et al., 2015) showed that information that directly competes with relevant cue-target associations becomes increasingly suppressed over the course of repeated testing.

4.3.2 THE ROLE OF MENTAL EFFORT DURING RETRIEVAL

The cognitive accounts that testing is an effortful process predict an involvement of brain regions that support controlled, selective memory retrieval. Executive control over effortful retrieval is often associated with the involvement of the prefrontal cortex, especially the ventral and dorsal lateral prefrontal cortex (VLPFC, DLPFC). Furthermore, the DLPFC and the anterior cingulate cortex (ACC) are often related to attention, control, and conflict monitoring processes (see Box 4.2 and Figure 4.2). Indeed, the four studies that analyzed brain activity during practice-testing and restudying (Rosner et al., 2013; van den Broek et al., 2013; Vannest et al., 2012; Wing et al., 2013) consistently reported higher activation in the VLPFC and the ACC during practice-testing compared to restudying.

Concerning the association between brain activation during practice and later performance at final test, different patterns of results were reported for the VLPFC and the ACC. Enhanced hippocampal functional connectivity with the VLPFC during practice-testing but not restudying predicted later performance (Wing et al., 2013), which could reflect an interaction of executive control processes through the VLPFC with core memory processing areas in the hippocampus (Scoville & Milner, 1957; Squire & Zola-Morgan, 1991) during practice-testing. There were no reports that higher VLPFC activation itself predicted better performance, and in one of the studies in which participants generated semantic associates, VLPFC activation during practice-testing even predicted later forgetting of words (Vannest et al., 2012). Further in line with this, Karlsson Wirebring et al. (2015) reported a decrease in activity of the left DLPFC over the course of repeated successful practice-testing which was predictive of better later memory performance. Similarly, in the previously

mentioned retrieval-induced forgetting study by Wimber et al. (2015), activity in the left and right VLPFC decreased over the course of repeated selective retrieval. Moreover, activations predicted how much competing information was suppressed but not how much target information was enhanced: more activation in the left and right VLPFC during the retrieval of specific memories predicted stronger suppression of that memory's competitors. At the same time, activations decreased over the course of repeated selective retrieval, and the stronger the decrease was, the more competing information was suppressed (Wimber et al., 2015). Overall, the available results are in line with the idea that activations in the VLPFC reflect the need to engage control mechanisms to select target information among competing memories (Kuhl et al., 2007), and that this process occurs during practice-testing more than restudying. Selection processes seem to become facilitated after repeated practice-testing, resulting in reduced recruitment of effortful control processes.

Cognitive accounts that testing facilitates later retrieval predict that brain activity during a later memory test (final test) should change as a function of prior testing. Indeed, VLPFC activity on a final test one day after combined restudying/practice-testing was lower the more often items had been successfully tested during prior practice (Eriksson et al., 2011), possibly reflecting that prior testing made the retrieval on the final test less demanding and reduced the need for competitor suppression. Behavioral reports of faster reaction times at the final test for previously tested compared to previously restudied items further support this interpretation (Keresztes et al., 2014; van den Broek et al., 2014; van den Broek et al., 2013).

The interpretation of ACC involvement in testing effects is not straightforward. Higher ACC activity during practice-testing than restudying predicted a larger behavioral testing effect in one study (Rosner et al., 2013), as participants who benefited more from practice-testing compared to restudying showed larger increases in ACC activation during practice-testing than restudying. Higher ACC activation measured during practice-testing was also more predictive of (better) performance on the final test than ACC activation measured during restudying (Wing et al., 2013). Although these results point to a positive effect of ACC involvement during practice-testing, one study in which highly associated words were generated reported that participants who showed higher ACC activity during practice-testing than restudying tended to perform *worse* on the final test (Vannest et al., 2012). A possible explanation for these contradictory outcomes is that ACC activation during practice could reflect the detection of competition between target responses and related competitors. Such conflict could, on the one hand, correlate with effortful, beneficial retrieval processes during practice-testing and thereby predict better performance at final test. On the other hand, high levels of conflict could also be a consequence of higher item difficulty, and therefore be related to worse performance.

Unlike activity in the VLPFC, ACC activity did not decrease but *increase* as a function of prior testing. In one study, the strength of this effect predicted performance five months later (Eriksson et al., 2011): Participants who showed a larger ACC response to previously tested items thus performed better later on. In a different study in which participants were scanned immediately and a week after testing/restudy practice, practice-testing led to better performance than restudying on the delayed final test but not on the immediate final test (Keresztes et al., 2014) (a typical result in testing effect studies, see Box 4.1). Activation in several brain areas, including the ACC, showed a similar interaction between time and condition as they were more active during the delayed final test of previously practice-tested than restudied materials, but not during the immediate final test.

In summary, activations in the VLPFC that are likely to reflect cognitive control processes required for the selection of target information among competing memories (Kuhl et al., 2007; Kuhl, Kahn, Dudukovic, & Wagner, 2008; Wimber et al., 2015; Wimber et al., 2008) were enhanced during practice-testing compared to restudying. Higher demands on this putative executive-control system did not, however, predict better learning. Instead, evidence suggests that a reduction of activation may correlate with the facilitation of retrieval processes, as competing information becomes less activated and demands on inhibition processes are reduced over the course of repeated testing. Results with respect to ACC involvement are mixed and cannot fully be explained by the idea that ACC activation is involved in conflict detection. This area has been related to long-term benefits of testing in two studies, and might play a role in consolidation processes that enhance retention (Eriksson et al., 2011; Keresztes et al., 2014).

4.3.3 NEURAL CORRELATES OF TEST-POTENTIATED ENCODING (TPE)

Three fMRI experiments (Liu et al., 2014; Nelson et al., 2013; Vestergren & Nyberg, 2014) addressed TPE. In two experiments (Nelson et al., 2013; Vestergren & Nyberg, 2014), participants encoded word-pairs and practiced the word-pairs between encoding episodes by practice-testing or restudying (see Figure 4.1). In the third study, all word-pairs were tested and then immediately restudied (Liu et al., 2014). For all TPE experiments, the neural response of interest is that during subsequent encoding *after* testing.

Some TPE accounts predict that previously unsuccessfully tested items receive extra attention during subsequent encoding (Pyc & Rawson, 2010). Vestergren and Nyberg (2014) specifically addressed this idea by comparing brain activity during the encoding of previously unsuccessfully tested items to successfully tested and restudied items. In their study, no differences in brain activity were found that were specific to re-encoding of *unsuccessfully* tested items, although several areas were

more activated overall after practice-testing than restudying. This finding suggests that TPE might affect all previously tested items regardless of testing success.

Both Liu et al. (2014) and Vestergren and Nyberg (2014) investigated which brain activations correlated with encoding success during TPE of previously unsuccessfully tested items. Liu et al. found that the left putamen and the caudate, the left hippocampus and the left PFC were more active during restudying of previously unsuccessfully tested items when these items were subsequently remembered than when they were not remembered. The authors suggested that the striatal (caudate, putamen) activity reflects changes in memory representations in response to negative feedback from the failed testing trial. Regarding the striatal involvement, van den Broek et al. (2013) also reported activity increases during successful practice-testing compared to restudying. They attributed this effect to the highlighting of motivationally significant information through the dopaminergic system. Vestergren and Nyberg (2014) found that activity in the anterior insula reflected successful encoding and related this to the possible involvement of a saliency detecting network (see Box 4.3 and Figure 4.2).

A consistent result across the two experiments that included comparisons of activations during encoding of previously tested and restudied items was higher activation in the VLPFC for previously tested items than un-tested items (Nelson et al., 2013; Vestergren & Nyberg, 2014). Vestergren and Nyberg speculated that this increased VLPFC activity along with activity in the anterior insula and the ACC could reflect deep processing due to the involvement of a saliency network in the brain that detects tested items as relevant and reduces distraction and mind-wandering during practice (see Box 4.3). Moreover, these authors found higher hippocampal activity for previously practice-tested than for restudied items, which suggests that practice-tested items underwent further re-encoding. Although Liu et al. (2014) did not compare practice-tested and restudied items, they reported (marginally) higher activation in the left hippocampus and the left PFC for word pairs that benefited from re-encoding after unsuccessful prior testing. Activity in these areas was higher for successful re-encoding relative to unsuccessful re-encoding after a failed test, but not after a successful test. These results are in line with cognitive accounts of TPE that suggest that items receive extra attention after retrieval failure (e.g., Pyc & Rawson, 2010).

Among areas found to be more active during encoding of previously tested than during encoding of previously restudied words, in one study (Nelson et al., 2013) analyses were focused on the left IPL/AG, an area that had shown retrieval-related activity in an earlier meta-analysis (Nelson et al., 2010). Nelson et al. (2013) found that activity in the IPL/AG correlated with the amount of new learning during TPE. Moreover, the authors plotted the average time course of brain activity for TPE trials from the IPL/AG showing signal fluctuation of a typical TPE trial. Based on a comparison

of this plot to the results of a previous meta-analysis on memory recognition, the authors argued that the time course of brain activity during the re-encoding of previously tested words (i.e., TPE) resembled that of successful recognition during memory retrieval paradigms. The activity of re-encoding of previously restudied pairs, in contrast, resembled that of seeing unknown items. From this finding, the authors concluded that there is increased retrieval-like activation in the IPL/AG during TPE. This could be due to the reinstatement of prior testing experiences, in line with the idea that prior testing becomes incorporated into the memory representation. This interpretation is somewhat in line with the "episodic context account" (Karpicke et al., 2014).

In sum, the three studies that investigated TPE revealed that after unsuccessful testing, activity in the insula (Vestergren & Nyberg, 2014) and the PFC, as well as the parietal cortex and the hippocampus (Liu et al., 2014) correlated with successful TPE. Only one study directly compared previously successfully and unsuccessfully tested items during subsequent encoding (Vestergren & Nyberg, 2014) and concluded that TPE might affect all previously tested items regardless of prior testing success.

4.4. TOWARDS A NEUROCOGNITIVE ACCOUNT OF THE TESTING EFFECT

In the behavioral literature, the testing effect has been explained in terms of changes in semantic representations through increasing elaboration or restriction of the search-set to relevant associations, effortful retrieval processes that become easier with repetition, and the potentiation of subsequent encoding. Can these cognitive accounts be informed by the neuroimaging studies published to date?

A relatively consistent finding across the testing effect fMRI studies is that even though the engagement of temporo-parietal semantic memory storage areas, such as the IPL and the MTG, was greater during restudying compared to practice-testing (Rosner et al., 2013; van den Broek et al., 2013; Vannest et al., 2012; Wing et al., 2013), activation increases only predicted later performance when measured during practice-testing and not when measured during restudying (van den Broek et al., 2013; Vannest et al., 2012; Wing et al., 2013). Engagement of these areas could reflect how testing alters memory representations such that they become more resistant to decay. A simple interpretation of the observation that temporo-parietal storage areas were less active during practice-testing than during restudying would be that testing does not generally enhance elaboration in comparison to restudying. However, activations in these areas could reflect both relevant and irrelevant semantic information processing and reduced activations during practice-testing could also reflect a beneficial focus of attention on relevant information (van den Broek et al.,

2013). Testing might direct elaboration to relevant associations that improve retention more than unfocused elaboration that takes place during restudy. Support for the notion that testing elaborates memory representations comes from the finding that successful repeated retrieval that fostered long-term retention was characterized by higher BOLD signal but lower pattern similarity in the parietal cortex (Karlsson Wirebring et al., 2015). However, there is also support for an involvement of selection processes during testing. During repeated selective retrieval, target information seems to become increasingly activated, whereas competing information becomes increasingly suppressed (Wimber et al., 2015). These results raise the question of whether an alternative cognitive model is needed that can accommodate both elaboration and selection processes. For example, elaboration during testing could be selective and focus on associations that strengthen the cue-target link, whereas associations that compete with the target response are suppressed.

The fMRI findings related to selective retrieval and inhibition processes are mixed. There is evidence that the VLPFC and the ACC are more involved during practice-testing than restudying (Rosner et al., 2013; van den Broek et al., 2013; Vannest et al., 2012; Wing et al., 2013), and several studies demonstrated a link between activity in the VLPFC and the ACC and later memory performance at final test, but the direction of this link differed among studies. Moreover, different effects of prior testing on VLPFC and ACC activations during the final test suggest a differential role of these two areas during testing. With regard to the VLPFC, one explanation could be that the VLPFC does not directly influence cue-target associations, but has an effect through interactions with the core memory system in the hippocampus (Wing et al., 2013). The VLPFC may act during the first instances of testing when selection is demanded during retrieval but with repeated successful retrieval, the VLPFC engagement might decrease. This claim is supported by the relation between changes in brain activity over the course of repeated retrieval and later memory performance (Karlsson Wirebring et al., 2015; Wimber et al., 2015). With regard to the ACC, activation increases after prior testing are difficult to interpret under the assumption that the ACC detects conflicting response options during retrieval. An alternative interpretation of the role of the ACC during the final test is that it reflects the amount of attention evoked by the tested materials in a similar way as during TPE (see Box 4.3 on the ACC as a part of the saliency network). By this view, attention is higher during practice-testing than restudying (Wing et al., 2013), more attention is paid to previously tested items than previously restudied items (Vestergren & Nyberg, 2014), and through this heightened attention there is a higher chance for these items to become well-consolidated for better retention (Eriksson et al., 2011).

Imaging studies on the testing effect have also revealed patterns of activation that cannot easily be linked to existing cognitive accounts. Certain structures of the

brain were not covered in this review, such as the posterior cingulate cortex (PCC) and the thalamus. At least some of these areas are likely to also play a functional role in testing, and future findings could reveal additional explanations of the testing effect. As a case in point, we outline in Box 4.3 how the testing effect could be explained in terms of reduced mind-wandering or saliency detection.

Results from TPE studies show how testing influences subsequent learning. Depending on the timing of the re-encoding occasion, encoding success was predicted by the involvement of different areas. Immediate feedback after a failed practice-test evoked extra brain activity in the left PFC when items were later remembered (compared to forgotten) on the final performance test (Liu et al., 2014). When re-encoding occurred after a whole round of practice-testing, activation in the anterior insula correlated with successful re-encoding (Vestergren & Nyberg, 2014). Analyses of patterns of brain activation suggest that practice-testing could lead to the engagement of retrieval processes in the IPL/AG regions during subsequent encoding when subjects are reminded of prior testing, and thereby form an alternative learning context that enriches memory representations (Nelson et al., 2013). Alternatively, or in addition, testing could increase attention via the engagement of saliency networks through the VLPFC/anterior insula and the ACC activity (see Box 4.3).

Box 4.3. The role of Default-Mode and Saliency networks in testing practice

Imaging studies have revealed patterns of activation that may stimulate ways to think about the testing effect that go beyond existing cognitive accounts. The role of default mode and saliency networks is an example of such results (see Figure 4.2 for candidate brain regions).

The default mode network / resting state network

Imaging studies have identified the *Default Mode Network* (DMN), a set of brain regions that are coactive during rest but become relatively deactivated during demanding tasks (Gusnard & Raichle, 2001; Raichle & Snyder, 2007). Brain regions commonly implicated in the DMN include the ventro- and dorso-medial prefrontal cortex, the hippocampus, the posterior midline structures (posterior cingulate; retrosplenial, precuneus), as well as the inferior parietal lobe (IPL) and lateral temporal cortices. The DMN was initially identified during rest, and it has been interpreted as reflecting internally driven thoughts that are, at least partly, task-unrelated. Indeed, unsuccessful memory encoding is accompanied by increased DMN activation (Kim, 2011), alluding to the idea that increased DMN activity reflects lapses in the focus of attention towards the encoding task. However, DMN activity has also been found when tasks involve self-referential thought such as (autobiographical) episodic memory retrieval (Buckner, Andrews-Hanna, & Schacter, 2008; Svoboda, McKinnon, & Levine, 2006).

Testing effect fMRI studies often reported higher activation in the DMN during restudying compared to practice-testing (van den Broek et al., 2013; Vannest et al., 2012;

Wing et al., 2013). This may indicate that testing reduces mind-wandering and distraction in comparison to restudying and, as a consequence, enhances attention to the task, which is in accordance with behavioral studies showing positive effects of interim tests (Szpunar, Jing, & Schacter, 2014; Szpunar, Khan, & Schacter, 2013). At the same time, a somewhat puzzling result is that increased activity in lateral temporal and parietal areas of DMN during practice-testing (but not during restudying) predicted successful retention (van den Broek et al., 2013; Vannest et al., 2012; Wing et al., 2013). A possible reason for this is that testing requires retrieval, which involves task-related self-referential thoughts that engage parts of the DMN.

Saliency network

Activity in the DMN is negatively correlated with activity in a network that becomes active when attention must be directed to external stimuli (Corbetta & Shulman, 2002) instead of internal thoughts (Fox et al., 2005; Greicius, Krasnow, Reiss, & Menon, 2003). This *saliency network* (including as important nodes the anterior insula and the anterior cingulate cortex; ACC) might play a crucial role in responding to salient cues in the environment that require attention and reducing internally directed thought (Menon & Uddin, 2010; Sridharan, Levitin, & Menon, 2008). During practice-testing more than restudying, and also as a function of the amount of prior testing, the saliency network is reported to become activated (Eriksson et al., 2011; van den Broek et al., 2013; Vannest et al., 2012; Vestergren & Nyberg, 2014; Wing et al., 2013). Thus, testing possibly increases the learners' attention to the materials through the saliency network. This could be an additional explanation of the testing effect revealed by neuroimaging studies, thus adding to the accounts suggested in the cognitive literature.

4.5 FUTURE PERSPECTIVES AND CONCLUSION

Several cognitive theories exist that make predictions about the nature of the testing effect and its constituent processes. However, there is no consensus yet about how to best explain the effect. This review adds a neurocognitive perspective to the literature by summarizing the evidence available from fMRI studies on the testing effect.

It is not trivial to link neuroimaging results to the behavioral literature and a number of limitations need to be taken into account. Cognitive theories are usually not constrained by the way the human brain works, and can have a rather abstract level of description. This makes one-to-one mapping between cognitive accounts and neural responses challenging. In addition, typical paradigms in neuroimaging and in behavioral research differ, which can make it more difficult to compare results. On the one hand, these differences in paradigms are due to methodological constraints of imaging studies, such as the required high numbers of comparable trials and the types of responses that participants can make in the scanner. On the other hand,

imaging studies have different paradigms to allow analyses that are not possible with behavioral data, such as tracking changes in patterns of brain activity over the course of repetitions or test-potentiated encoding. It also allows us to categorize the encoding trials retrospectively to later-remembered and later-forgotten trials (i.e., SME) and observe brain responses that are predictive of memory outcome later at test. As in other areas of educational neuroscience, continued exploration and testing of the translation between and incorporation of theories from cognitive and neuroscientific fields are called for.

The number of imaging studies of beneficial effects of practice-testing on learning outcomes is still comparably small. For this review, we identified ten fMRI studies that related practice-testing to later performance or explicitly focused on the testing effect or TPE. These studies employed a variety of approaches, measuring brain activations during practice-testing and restudying, over the course of repeated practice-testing, during subsequent encoding after practice-tests, and during final performance tests. Each of these approaches allows the evaluation of different predictions from the cognitive literature, but the number of studies of each approach is still small. More research is therefore needed before more definite conclusions can be drawn. In addition, a number of limitations of the literature base should be noted. First, two of the reviewed studies did not report a behavioral testing effect (Nelson et al., 2013; Vestergren & Nyberg, 2014). This did not influence the analysis of the imaging data because brain activations were related to prior retrieval success and later performance on a by-item basis, but a behavioral effect would confirm that the chosen paradigms caused TPE. Second, two studies (Karlsson Wirebring et al., 2015; Liu et al., 2014) did not include a restudy comparison condition, so it is unclear to what extent the report(ed neural activations during practice-testing also occur during other forms of practice such as restudying. This however, is not necessarily problematic for cognitive theories that predict that quantitatively *more* elaboration or mental effort is involved during practice-testing compared to restudying, rather than qualitatively different processes. Third, we reviewed two studies that implemented retrieval of pre-existing semantic associations rather than recently learned associations (Rosner et al., 2013; Vannest et al., 2012). Although similarities exist between these two forms of retrieval (Peterson & Mulligan, 2013), it would be informative to take into account the nature of the associations that are activated during practice-tests in future studies. The testing effect has been found with many different materials (Box 4.1) but all testing effect fMRI studies so far used visual word-pairs. It is an open question if similar effects are obtained when participants study different materials (e.g., auditory, non-verbal).

Regarding the fMRI methods employed, the focus of the available imaging studies on testing effects has been largely on activity changes in specific (more

or less isolated) nodes of the brain. We suggest that in future studies, the neural mechanisms underlying the testing effect may better be seen by examining patterns of interactions across several brain areas. Connectivity analyses of the interactions between brain regions may help develop a network perspective on the mechanisms of testing, as illustrated by the results of hippocampal connectivity by Wing et al. (2013). Extending such analyses to other brain regions may reveal interactions in a broader network of areas underlying the testing effect. Furthermore, changes in memory representations as a consequence of practice-testing may be better documented by changes in *patterns* of neural activation over the course of and after practice rather than net activation differences during practice-testing and restudying. Techniques like multivariate pattern analysis/RSA (Kriegeskorte, Goebel, & Bandettini, 2006; Kriegeskorte et al., 2008), have only recently been used to investigate testing effects in this way (Karlsson Wirebring et al., 2015). More research along these lines might further improve insight into the mechanisms of testing effects.

Imaging studies offer a unique way to test predictions from cognitive theories and may suggest new ways to think about well-known behavioral phenomena like the testing effect. This review of recent fMRI studies on the testing effect informs the literature about its neurocognitive substrates by highlighting several different processes that might be important for testing effects. First, the available data support the idea that practice-testing indeed engages the memory representational areas in the posterior cortices in a different way than restudying does, possibly in a more focused way that stabilizes the relevant memory trace. Examining variation in patterns of brain activity during successful repeated retrieval in relation to subsequent memory, produced support for the semantic elaboration view. However, more studies are needed to establish how these results fit with studies that show the suppression of competing information during selective retrieval. An alternative cognitive model might be needed that can accommodate both elaboration and selection processes, such as selective elaboration. Second, there is evidence for effortful, controlled retrieval processes reflected in the engagement of the prefrontal cortex during practice-testing, which reduces over the course of repeated practice-testing, although the link of this reduction with later performance at final test needs further investigation. Third, TPE is not restricted to materials that were previously not recalled and could involve the reactivation of testing experiences and extra attention to previously tested information. Finally, neuroimaging studies point at other possible mechanisms that have not been covered by cognitive accounts yet, such as enhanced attention through the engagement of motivation or saliency networks or a reduction of mind-wandering.

4.6 REFERENCES

Badre, D., & Wagner, A. D. (2004). Selection, integration, and conflict monitoring: Assessing the nature and generality of prefrontal cognitive control mechanisms. *Neuron, 41*(3), 473-487. https://doi.org/10.1016/S0896-6273(03)00851-1

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

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences, 15*(11), 527-536. https://doi.org/10.1016/j.tics.2011.10.001

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex, 19*(12), 2767-2796. https://doi.org/10.1093/cercor/bhp055

Binder, J. R., McKiernan, K. A., Parsons, M. E., Westbury, C. F., Possing, E. T., Kaufman, J. N., & Buchanan, L. (2003). Neural correlates of lexical access during visual word recognition. *Journal of Cognitive Neuroscience, 15*(3), 372-393. https://doi.org/10.1162/089892903321593108

Blumenfeld, R. S., & Ranganath, C. (2007). Prefrontal cortex and long-term memory encoding: An integrative review of findings from neuropsychology and neuroimaging. *The Neuroscientist, 13*(3), 280-291. https://doi.org/10.1177/1073858407299290

Botvinick, M. (2007). Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. *Cognitive, Affective, & Behavioral Neuroscience, 7*(4), 356-366. https://doi.org/10.3758/cabn.7.4.356

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain‘s default network. *Annals of the New York Academy of Sciences, 1124*(1), 1-38. https://doi.org/10.1196/annals.1440.011

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(5), 1118-1133. https://doi.org/10.1037/a0019902

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*(6), 1563-1569. https://doi.org/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*(6), 1547-1552. https://doi.org/10.1037/a0024140

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*(5), 279-283. https://doi.org/10.1177/0963721412452728

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*(2), 268-276. https://doi.org/10.3758/bf03193405

Carpenter, S. K., & Kelly, J. W. (2012). Tests enhance retention and transfer of spatial learning. *Psychonomic Bulletin & Review, 19*(3), 443-448. https://doi.org/10.3758/s13423-012-0221-2

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students‘ retention of U.S. history facts. *Applied Cognitive Psychology, 23*(6), 760-771. https://doi.org/10.1002/acp.1507

Carter, H.V. (1918). Principal fissures and lobes of the cerebrum viewed laterally. Figure 728 from H. Gray (Ed.), *Anatomy of the human body* (20th edition, revised by Warren H. Lewis). Philadelphia: Lea & Febiger. Image edited by O. Räisänen for wikimedia [Public domain]. Retrieved from https://commons.wikimedia.org/wiki/File:Lobes_of_the_brain_NL.svg

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3*(3), 201-215. https://doi.org/10.1038/nrn755

de Jonge, M., Tabbers, H. K., & Rikers, R. M. J. P. (2014). Retention beyond the threshold: Test-enhanced relearning of forgotten information. *Journal of Cognitive Psychology, 26*(1), 58-64. https://doi.org/10.1080/20445911.2013.858721

Dudai, Y. (2004). The neurobiology of consolidations, or, how stable is the engram? *Annual Review of Psychology, 55*, 51-86. https://doi.org/10.1146/annurev.psych.55.090902.142050

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4-58. https://doi.org/10.1177/1529100612453266

Eriksson, J., Kalpouzos, G., & Nyberg, L. (2011). Rewiring the brain with repeated retrieval: A parametric fMRI study of the testing effect. *Neuroscience Letters, 505*(1), 36-40. https://doi.org/10.1016/j.neulet.2011.08.061

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America, 102*(27), 9673-9678. https://doi.org/10.1073/pnas.0504136102

Graves, W. W., Desai, R., Humphries, C., Seidenberg, M. S., & Binder, J. R. (2010). Neural systems for reading aloud: A multiparametric approach. *Cerebral Cortex, 20*(8), 1799-1815. https://doi.org/10.1093/cercor/bhp245

Greicius, M. D., Krasnow, B., Reiss, A. L., & Menon, V. (2003). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences, 100*(1), 253-258. https://doi.org/10.1073/pnas.0135058100

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition, 40*(4), 505-513. https://doi.org/10.3758/s13421-011-0174-0

Gusnard, D. A., & Raichle, M. E. (2001). Searching for a baseline: Functional imaging and the resting human brain. *Nature Reviews Neuroscience, 2*(10), 685-694. https://doi.org/10.1038/35094500

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*(4), 801-812. https://doi.org/10.1037/a0023219

Holmes, C. J., Hoge, R., Collins, L., Woods, R., Toga, A. W., & Evans, A. C. (n.d.). Enhancement of MR images using registration for signal averaging. *Journal of Computer Assisted Tomography*, *22*(2), 324–333.

Izawa, C. (1971). The test trial potentiating model. *Journal of Mathematical Psychology, 8*(2), 200-224. https://doi.org/10.1016/0022-2496(71)90012-5

Kang, S. K., McDaniel, M., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review, 18*(5), 998-1005. https://doi.org/10.3758/s13423-011-0113-x

Karlsson Wirebring, L., Wiklund-Hörnqvist, C., Eriksson, J., Andersson, M., Jonsson, B., & Nyberg, L. (2015). Lesser neural pattern similarity across repeated tests is associated with better long-term memory retention. *The Journal of Neuroscience, 35*(26), 9595-9602. https://doi.org/10.1523/jneurosci.3550-14.2015

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*(6018), 772-775. https://doi.org/10.1126/science.1199327

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 61, pp. 237-284). San Diego, CA: Elsevier Academic Press.

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

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

Keresztes, A., Kaiser, D., Kovács, G., & Racsmány, M. (2014). Testing promotes long-term learning via stabilizing activation patterns in a large network of brain areas. *Cerebral Cortex, 24*(11), 3025-3035. https://doi.org/10.1093/cercor/bht158

Kim, H. (2011). Neural activity that predicts subsequent memory and forgetting: A meta-analysis of 74 fMRI studies. *Neuroimage, 54*(3), 2446-2461. https://doi.org/10.1016/j.neuroimage.2010.09.045

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65*(2), 85-97. https://doi.org/10.1016/j.jml.2011.04.002

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America, 103*(10), 3863-3868. https://doi.org/10.1073/pnas.0600244103

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

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical Education, 43*(1), 21-27. https://doi.org/10.1111/j.1365-2923.2008.03245.x

Kuhl, B. A., Dudukovic, N. M., Kahn, I., & Wagner, A. D. (2007). Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nature Neuroscience, 10*(7), 908-914. https://doi.org/10.1038/nn1918

Kuhl, B. A., Kahn, I., Dudukovic, N. M., & Wagner, A. D. (2008). Overcoming suppression in order to remember: Contributions from anterior cingulate and ventrolateral prefrontal cortex. *Cognitive, Affective, & Behavioural Neuroscience, 8*(2), 211-221. https://doi.org/ 10.3758/CABN.8.2.211

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29*(3), 210-212. https://doi.org/10.1207/s15328023top2903_06

Liu, X. L., Liang, P., Li, K., & Reder, L. M. (2014). Uncovering the neural mechanisms underlying learning from tests. *PLoS ONE, 9*(3), e92025. https://doi.org/10.1371/journal.pone.0092025

Lockhart, R. S. (2002). Levels of processing, transfer-appropriate processing, and the concept of robust encoding. *Memory, 10*(5-6), 397-403. https://doi.org/10.1080/09658210244000225

Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology, 38*(2), 94-97. https://doi.org/10.1177/0098628311401587

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology, 58*, 25-45. https://doi.org/10.1146/annurev.psych.57.102904.190143

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*(4-5), 494-513. https://doi.org/10.1080/09541440701326154

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27*(3), 360-372. https://doi.org/10.1002/acp.2914

Menon, V., & Uddin, L. (2010). Saliency, switching, attention and control: A network model of insula function. *Brain Structure and Function, 214*(5-6), 655-667. https://doi.org/10.1007/s00429-010-0262-0

Meyer, A. N. D., & Logan, J. M. (2013). Taking the testing effect beyond the college freshman: Benefits for lifelong learning. *Psychology and Aging, 28*(1), 142-147. https://doi.org/10.1037/a0030890

Murayama, K., Miyatsu, T., Buchli, D., & Storm, B. C. (2014). Forgetting as a consequence of retrieval: A meta-analytic review of retrieval-induced forgetting. *Psychological Bulletin, 140*(5), 1383-1409. https://doi.org/10.1037/a0037505

Nelson, S. M., Arnold, K. M., Gilmore, A. W., & McDermott, K. B. (2013). Neural signatures of test-potentiated learning in parietal cortex. *The Journal of Neuroscience, 33*(29), 11754-11762. https://doi.org/10.1523/jneurosci.0960-13.2013

Nelson, S. M., Cohen, A. L., Power, J. D., Wig, G. S., Miezin, F. M., Wheeler, M. E., . . . Petersen, S. E. (2010). A parcellation scheme for human left lateral parietal cortex. *Neuron, 67*(1), 156-170. https://doi.org/10.1016/j.neuron.2010.05.025

Nyberg, L., Habib, R., McIntosh, A. R., & Tulving, E. (2000). Reactivation of encoding-related brain activity during memory retrieval. *Proceedings of the National Academy of Sciences, 97*(20), 11120-11124. https://doi.org/10.1073/pnas.97.20.11120

Nyberg, L., Petersson, K. M., Nilsson, L.-G., Sandblom, J., Åberg, C., & Ingvar, M. (2001). Reactivation of motor brain areas during explicit memory for actions. *Neuroimage, 14*(2), 521-528. https://doi.org/10.1006/nimg.2001.0801

Pastötter, B., Weber, J., & Bäuml, K.-H. T. (2013). Using testing to improve learning after severe traumatic brain injury. *Neuropsychology, 27*(2), 280-285. https://doi.org/10.1037/a0031797

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience, 8*(12), 976-987. https://doi.org/10.1038/nrn2277

Peterson, D. J., & Mulligan, N. W. (2013). The negative testing effect and multifactor account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(4), 1287-1293. https://doi.org/10.1037/a0031337

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage, 62*(2), 816-847. https://doi.org/10.1016/j.neuroimage.2012.04.062

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*(4), 437-447. https://doi.org/10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*(6002), 335. https://doi.org/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*(3), 737-746. https://doi.org/10.1037/a0026166

Race, E., Kuhl, B. A., Badre, D., & Wagner, A. D. (2009). The dynamic interplay between cognitive control and memory. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (4th ed., pp. 705-724). Cambridge, MA: MIT Press.

Raichle, M. E., & Snyder, A. Z. (2007). A default mode of brain function: A brief history of an evolving idea. *Neuroimage, 37*(4), 1083-1090. https://doi.org/10.1016/j.neuroimage.2007.02.041

Reas, E. T., & Brewer, J. B. (2012). Retrieval search and strength evoke dissociable brain activity during episodic memory recall. *Journal of Cognitive Neuroscience, 25*(2), 219-233. https://doi.org/10.1162/jocn_a_00335

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied, 15*(3), 243-257. https://doi.org/10.1037/a0016496

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20-27. https://doi.org/10.1016/j.tics.2010.09.003

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

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249-255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 233-239. https://doi.org/10.1037/a0017678

Rorden, C., & Brett, M. (2000). Stereotaxic display of brain lesions. *Behavioural Neurology, 12*(4), 191–200. https://doi.org/10.1155/2000/421719

Rosner, Z. A., Elman, J. A., & Shimamura, A. P. (2013). The generation effect: Activating broad neural circuits during memory encoding. *Cortex, 49*(7), 1901-1909. https://doi.org/10.1016/j.cortex.2012.09.009

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

Rugg, M. D., & Vilberg, K. L. (2013). Brain networks underlying episodic memory retrieval. *Current Opinion in Neurobiology, 23*(2), 255-260. https://doi.org/10.1016/j.conb.2012.11.005

Rumelhart, D. E., McClelland, J. L., & PDP Research Group (1986). *Parallel Distributed Processing* (Vol. 1). Cambridge The MIT Press.

Sanquist, T. F., Rohrbaugh, J. W., Syndulko, K., & Lindsley, D. B. (1980). Electrocortical signs of levels of processing: Perceptual analysis and recognition memory. *Psychophysiology, 17*(6), 568-576.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery & Psychiatry, 20*(1), 11-21.

Squire, L. R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. *Science, 253*(5026), 1380-1386.

Sridharan, D., Levitin, D. J., & Menon, V. (2008). A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proceedings of the National Academy of Sciences, 105*(34), 12569-12574. https://doi.org/10.1073/pnas.0800005105

Stenlund, T., Jönsson, F. U., & Jonsson, B. (2016). Group discussions and test-enhanced learning: Individual learning outcomes and personality characteristics. *Educational Psychology*, 1-15. https://doi.org/10.1080/01443410.2016.1143087

Storm, B., & Levy, B. (2012). A progress report on the inhibitory account of retrieval-induced forgetting. *Memory & Cognition, 40*(6), 827-843. https://doi.org/10.3758/s13421-012-0211-7

Svoboda, E., McKinnon, M. C., & Levine, B. (2006). The functional neuroanatomy of autobiographical memory: A meta-analysis. *Neuropsychologia, 44*(12), 2189-2208. https://doi.org/10.1016/j.neuropsychologia.2006.05.023

Szpunar, K. K., Jing, H. G., & Schacter, D. L. (2014). Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education. *Journal of Applied Research in Memory and Cognition, 3*(3), 161-164. https://doi.org/10.1016/j.jarmac.2014.02.001

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences of the United States of America, 110*(16), 6313-6317. https://doi.org/10.1073/pnas.1221764110

Thomas, R. C., & McDaniel, M. A. (2013). Testing and feedback effects on front-end control over later retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(2), 437-450. https://doi.org/10.1037/a0028886

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval. *Experimental Psychology, 56*(4), 252-257. https://doi.org/10.1027/1618-3169.56.4.252

van den Broek, G. S. E., Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? Insights from immediate and delayed retrieval speed. *Memory, 22*(7), 803-812. https://doi.org/10.1080/09658211.2013.831455

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

Vannest, J., Eaton, K. P., Henkel, D., Siegel, M., Tsevat, R. K., Allendorfer, J. B., . . . Szaflarski, J. P. (2012). Cortical correlates of self-generation in verbal paired associate learning. *Brain Research, 1437*(0), 104-114. https://doi.org/10.1016/j.brainres.2011.12.020

Vestergren, P., & Nyberg, L. (2014). Testing alters brain activity during subsequent restudy: Evidence for test-potentiated encoding. *Trends in Neuroscience and Education, 3*(2), 69-80. https://doi.org/10.1016/j.tine.2013.11.001

Wheeler, M. E., Petersen, S. E., & Buckner, R. L. (2000). Memory's echo: Vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences, 97*(20), 11125-11129.

Whitney, C., Jefferies, E., & Kircher, T. (2011). Heterogeneity of the left temporal lobe in semantic representation and control: Priming multiple versus single meanings of ambiguous words. *Cerebral Cortex, 21*(4), 831-844. https://doi.org/10.1093/cercor/bhq148

Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology, 55*(1), 10-16. https://doi.org/10.1111/sjop.12093

Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., & Anderson, M. C. (2015). Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nature Neuroscience, 18*(4), 582-589. https://doi.org/10.1038/nn.3973

Wimber, M., Bäuml, K.-H., Bergström, Z., Markopoulos, G., Heinze, H.-J., & Richardson-Klavehn, A. (2008). Neural markers of inhibition in human memory retrieval. *The Journal of Neuroscience, 28*(50), 13419-13427. https://doi.org/10.1523/jneurosci.1916-08.2008

Wing, E. A., Marsh, E. J., & Cabeza, R. (2013). Neural correlates of retrieval-based memory enhancement: An fMRI study of the testing effect. *Neuropsychologia, 51*(12), 2360-2370. https://doi.org/10.1016/j.neuropsychologia.2013.04.004

Winocur, G., & Moscovitch, M. (2011). Memory transformation and systems consolidation. *Journal of the International Neuropsychological Society, 17*(5), 766-780. https://doi.org/10.1017/S1355617711000683

Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J. A., & Poldrack, R. A. (2010). Greater neural pattern similarity across repetitions is associated with better memory. *Science, 330*(6000), 97-101. https://doi.org/10.1126/science.1193125

4.7 APPENDIX

Table 4.1 Paradigms, behavioral testing effect and key brain activations related to beneficial effects of practice-testing in the 10 reviewed studies.

Study	Overview of experiment	
Studies focusing on brain activity during the intervention		
Karlsson Wirebring et al., 2015	Baseline exposure	Intentional encoding of 60 Swahili-Swedish word-pairs (ten consecutive presentations)
	Intervention	Each word-pair was repeatedly tested three times. The Swahili word was used as a probe, and participants indicated whether they knew the Swedish word, believed they knew or did not know. Immediately after, they chose among four alternatives the second letter of the word.
	Final test	After 7 days, cued recall of the translation of the Swahili words as during the intervention
Rosner et al., 2014	Baseline exposure	None (participants had prior knowledge of the 100 pairs of semantically related words that were practiced)
	Intervention	Half of the pairs were generated (=tested) once ("garbage = w_st_"), and half were read (=restudied) once ("garbage = waste"); i.e. the task was to either retrieve a target word that was semantically related to the cue from memory or to read the complete word pair.
	Final test	Immediately after the intervention, a recognition test with confidence rating included all target words and 100 lures.
van den Broek et al., 2013 (Chapter 3)	Baseline exposure	Intentional encoding of 100 Swahili words with translations through writing task and repeated exposure with judgments of learning
	Intervention	Half of the pairs were tested three times ("mit –translate!"), and half were restudied ("wingu - cloud") three times;
	Final test	After 7 days, cued recall of the translation of the Swahili words

	Behavioral testing effect	Key brain regions involved

	n/a (no comparison with a study condition)	During the intervention:
		• SME during repeated correct retrieval:
		  • LR > LF: right SPL
		• RSA analysis in the right SPL during repeated correct retrieval:
		  • LR > LF: Lower pattern similarity for LR items compared to LF items
		  • Repetition x subsequent memory interaction: a monotonic decrease in the left DLPFC over the course of repeated testing for LR items, but not for LF items
		During the final test:
		• Retrieval success effect (remembered > forgotten items) in bil. PPC (right SPL/AG), bil. ITG, bil. IFG, right HC and PHG, left putamen (no differences for the reversed contrast)

	Yes, at immediate test higher correct recognition rate for tested (87%) than for restudied words (65%)	• Activation difference
		  • T > RS for LR condition: VLPFC, bil.VLPFC, LOC, ITG, IPS, PrC and ACC.
		  • RS > T: none
		• Correlation between behavioural and neural testing effect: Participants who showed a stronger behavioral testing effect tended to show a larger difference in neural activity during testing and restudying of LR words in paracingulate, frontal pole, left ACC, and right SFG
		• Brain activity related to successful retrieval during the final test did not differ between previously T and RS

	Yes, at delayed test: Higher cued recall and shorter response times for tested (58%) than for restudied words (49%)	• Activation difference
		  • T > RS: bil.VLPFC/insula, bil. striatum
		  • RS > T: bil. IPL, right MTG
		  • LR > LF for T but not RS: left MTG, left IPL(AG/SMG)
		  • LF > LR for T but not RS: left calcarine, left SMA

Study	Overview of experiment	
Vannest et al., 2012	Baseline exposure	None (participants had prior knowledge of the 60 pairs of semantically or phonologically related words that were practiced)
	Intervention	Half of the pairs were generated (=tested; "salt - p*****") once and half were read (=restudied; "salt – pepper") once. Similar to Rosner et al. participants either retrieved associated word from memory or read the word pair that was presented on the screen. Participants knew that a memory test would follow.
	Final test	Immediately after the intervention; forced-choice task in which one word was presented and the associated word was selected among two lures
Wing et al., 2013	**Baseline exposure**	Rating semantic relatedness of 192 weakly related noun pairs
	Intervention	Half of the pairs were tested ("TUSK -?") once and half were restudied ("TUSK – HORN") once.
	Final test	After 1 day, surprise cued recall test by presenting the first word as cue for recall of the second word *Note: This study alternated between baseline exposure and intervention. It consisted of 12 blocks in which 16 words were first presented for baseline exposure and then for one testing or restudy trial.*
Studies focusing on brain activity during test-potentiated encoding		
Nelson et al., 2013	**Baseline exposure**	Intentional encoding of 126 weakly associated word pairs
	Intervention	Interim testing ("crater - ?") or restudying ("crater – lake") followed by subsequent (test-potentiated) encoding (crater – lake).
	Final test	After 1 day, cued-recall combined with detection of new words, by presenting the first words intermixed with 42 new words to prompt the cued recall of the second word (or recognition as new)

Behavioral testing effect	Key brain regions involved

Yes, at immediate test better multiple choice performance for tested (80%) than for restudied words (72%)	• Activation difference
	  • T > RS: bil.VLPFC/MFG/insula, ACC/MedFG, bil. caudate, bil. AG/SMG/IOG/MOG/SOG,
	  • RS > T: Left M/SFG, right MFG, bil. insula, bil. IPL, right PrC, right lingual, left cuneus
	  • LR > LF for T but not RS: Left STG/SMG, left insula, right MedFG
	  • LF > LR for T but not RS: Right insula/VLPFC
	• Correlations between neural testing effect and recognition performance at the final test across participants:
	  • Degree of T > RS in Left M/STG predicted better recognition of tested items.
	  • Degree of T > RS in Right M/SFG, bil. ACC, left PCC, right insula, right paracentral lobule predicted *worse* recognition of tested items.
	  • Degree of RS > T in bil. cuneus and right PrC predicted better recognition of restudied items.
	  • Degree of RS > T in Right M/MedFG, left postcentral, right cingulate, left MTG predicted *worse* recognition of restudied items.

Yes, at delayed test higher cued recall for tested (63%) than for restudied words (51%)	• Activation difference
	  • T > RS: bil. VLPFC/insula, ACC, left ITG, left PrC, left PHG, left MOG
	  • RS > T: bil. MFG, bil. MTG, bil. IPL, right PrC
	  • LR > LF: SMedFG
	• Activity interaction
	  • LR > LF for T and reverse for RS: Left I/MTG, bil. hippocampus
	  • Interaction effect (no direction info): ACC, right STG, left insula/claustrum
	• Connectivity with hippocampus interaction
	  • LR > LF for T only: PCC, vmPFC, left VLPFC,
	  • LR > LF vs T > RS (no direction reported): Right MFG, right MedFG, bil. STG, right insula/STG, right MTG, left ITG, left PHG, right SMG, bil. postcentral, right PrC

No
At delayed test, no difference between tested and restudied only word pairs
	• Activation difference
	  • baseline exposure > test-potentiated encoding: Left VLPFC,
	  • baseline exposure < test-potentiated encoding: Left IPL/AG, PrC, MCC
	  • Previously T > RS at test-potentiated encoding: Left IPL/AG, PrC, MCC
	• Activity level at test-potentiated encoding for items not recalled previously in left IPL correlated with the amount of new learning: left IPL/AG
	• Overlap of time course of brain activation in left IPL/AG during test-potentiated encoding with time course of activation during successful recognition found in previous meta-analysis

Study	Overview of experiment	
Vestergren et al., 2014	Baseline exposure	24h before the intervention, intentional encoding of 120 Swahili words with translations during 5 presentations. Immediately preceding the intervention one more presentation of all word pairs.
	Intervention	Testing (wingu _____) or restudying (wingu – cloud) followed by subsequent (test-potentiated) encoding (wingu – cloud).
	Final test	Immediately after intervention, cued recall of the translation of the Swahili words, followed by a multiple choice task to select the translation among 3 lures
Liu et al., 2014	**Baseline exposure**	Intentional encoding of 45 high-frequency unrelated Chinese-Chinese word pairs
	Intervention	One testing trial (T1) per word pair immediately followed by restudy. Cue words were shown with a prompt to recall the second word before selecting it among all possible targets. After each test trial, the complete pair was shown for 3 seconds for restudying.
	Final test	Immediately after the intervention, cued recall of the target words (T2). The delay between the intervention and the final test of single items was approximately 20 minutes.
Studies focusing on brain activity after the intervention		
Eriksson et al., 2011	Baseline exposure	Intentional encoding of 40 Swahili words with translations (this was combined with intervention)
	Intervention	Alternating restudying and testing until a word was tested successfully, then only testing continued and restudying stopped for that word for at least 4 cycles
	Final test	After 1 day, cued recall of the translation of the Swahili words (fMRI). The same test was given again after 7 days and 5 months (no fMRI)

Behavioral testing effect	Key brain regions involved

No
At immediate test, no difference between tested and restudied only word pairs
	• Activation difference during test-potentiated encoding
	  • previously T > RS: bil. VLPFC/insula, left hippocampus
	  • previously RS > T: Left MCC, bil. SMG, bil. PrC, bil. MTG
	  • Previously unsuccessfully recalled T: recalled later > not recalled later: anterior insula
	• No general activation difference during test-potentiated encoding between items that were previously tested unsuccesfully and items that were previously tested successfully

N/A
All trials were both tested and subsequently restudied
 	
There was an increase in recall accuracy from T1 to T2
	• LR > LF for successful test trials at T1 (brain activity during T1 dependent on performance at T2)
	  • ROI analysis: left PFC, right PFC, right PPC, and left hippocampus. Marginally significant left PPC and right hippocampus
	  • whole brain analyses:
	  left SFG, MFG, IFG, right IFG; left IPL, SMG, MTG; left STG; right STG, MTG
	• Significant correlation between activation for LR items during T1 and performance at T2:
	  • ROI analysis: in right PFC (r = .64, p = .022) and right PPC (r = .57, p = .012)
	• LR > LF for *unsuccessful* test trials at T1 (brain activity during T1 dependent on performance at T2)
	  • MFG, Precuneus, Cingulate gyrus
	• SME for baseline exposure
	  • PFC, left PPC, and bil. hippocampus, left
	• Test-potentiated learning during restudy after a failed test: trials that became correct on T2 > trials that were again incorrect on T2
	  • ROI: Marginally significant differences in left hippocampus and left PFC
	  • Whole brain analysis: Putamen/caudate

Yes
Delayed test: Positive correlation between number of successful practice tests and memory performance 1 week later
	• Activation difference during final test for items previously
	  • more retrieved < less retrieved items: Right SPL, right VLPFC
	  • more retrieved > less retrieved items: ACC
	  • more repetition of successful retrieval led to higher ACC activation
	• Participants who benefited from testing during intervention (more retrieval of items at intervention led to heightened ACC activation during final test) were the ones who had better memory performance 5 months later

Study	Overview of experiment	
Keresztes et al, 2013	Baseline exposure	Intentional encoding of 60 Swahili words with translations
	Intervention	6 rounds which each consisted of 1 block of 30 words being tested and one block of 30 words being restudied, and one (subsequent) encoding block of all words
	Final tests	Immediately (half of the participants) and 7 days (other half of the participants) after the intervention. Cued recall of the translation of the Swahili words.

Note. The overview of the experiment refers to the same baseline, intervention, and final test phases as depicted in Figure 1. Phases printed in bold in the second column were conducted in the MR scanner. Abbreviations: T = testing; RS = restudying; T > RS = contrast between testing trials and restudy trials; LR = later remembered (see Figure 1); LF = later forgotten; RSA = representational similarity analysis; ROI = region of interest. Abbreviations of anatomical regions in alphabetical order: ACC = anterior cingulate cortex; AG = angular gyrus; bil. = bilateral; IFG = inferior frontal gyrus; IPL = inferior parietal lobe; IPS = inferior parietal sulcus; ITG = inferior temporal gyrus; LOC = lateral occipital cortex; MCC = middle cingulate cortex; MedFG = medial frontal gyrus; MFG = middle frontal gyrus; orb. FG = orbital frontal gyrus; MTG = middle temporal gyrus; PHG = parahippocampal gyrus; PPC= posterior parietal cortices; PrC = precuneus; SFG = superior frontal gyrus; SMedFG = superior medial frontal; SMG = supramarginal gyrus; SPL = superior parietal lobe; STG = superior temporal gyrus; VLPFC = ventrolateral prefrontal cortex

Behavioral testing effect	Key brain regions involved
Yes, at delayed test higher cued recall for tested (50%) than for restudied words (39%) At immediate test, no difference between tested and restudied words.	• Analyses were focused at a set of brain areas (regions of interest) activated by a working memory task. • 2 way interaction Condition x Testing Moment: • Activation difference RS > T after 20min, T > RS after 1 week: bil. DLPFC, bil. insula, bil. IPL, right middle orb. FG, right SPL, left fusiform, right thalamus • This pattern of activations resembled the pattern found for behavioural effects (RS = T after 20min, T > RS after 1 week) • Interaction appeared driven by different changes over time: all areas decreased in activation over time after RS, no areas decreased over time after T

EFFECTS OF ELABORATE FEEDBACK DURING RETRIEVAL PRACTICE: COSTS AND BENEFITS OF RETRIEVAL PROMPTS

An article based on this chapter is in preparation as:
van den Broek, G.S.E., Segers, E., van Rijn, H., Takashima, A., Verhoeven, L. Effects of Elaborate Feedback during Practice Tests: Costs and Benefits of Retrieval Prompts.

Abstract. Retrieval practice enhances the retention of information over time, especially in combination with feedback. This study focuses on the effect of hints during feedback provided after retrieval failures. In three classroom experiments, high-school students practiced vocabulary words with an adaptive spaced retrieval program with either standard show-answer feedback or hints feedback. Show-answer feedback consisted of the display of the correct answer; hints feedback gave students a second chance to retrieve the correct answer from memory using orthographic (Experiment 1), mnemonic (Experiment 2), or cross-language hints (Experiment 3). During practice, hints feedback led to a shift in the distribution of the available practice time from further repetitions to longer feedback processing after errors. However, only mnemonic hints reduced (repeated) errors during practice. There was no overall effect of feedback on learning outcomes measured with a recall test several days after learning. However, feedback influenced later recall when the hints from practice were available on the test: compared to the show-answer condition, students needed significantly more orthographic recall prompts on the final test after practice with orthographic hints (Exp. 1) and showed better recall on a later test with mnemonic hints after practice with mnemonic hints (Exp. 2). These results indicate limited transfer of practice with hints to later recall without hints. Overall, three experiments did not produce convincing evidence that hints feedback is preferable over show-answer feedback during retrieval practice. The common preconception that hints feedback is beneficial for learning may not hold in realistic learning settings and under time constraints.

5.1 INTRODUCTION

Imagine two high-school students, Ann and Bob, who are practicing Latin vocabulary. Ann holds a list of vocabulary words and asks Bob to translate them from memory. "What does vestis mean?" Bob cannot remember the translation, so Ann says: "Think of the word vest!" And suddenly Bob remembers: "Oh, right! Vestis is clothing!" The two students in this fictional example practice the retrieval of words from memory – a practice strategy that a plethora of research has shown to be beneficial for long-term retention (Adesope, Trevisan, & Sundararajan, 2017; Roediger & Karpicke, 2006; Rowland, 2014). Retrieval practice is particularly effective with feedback that allows learners to correct errors and re-exposes them to information that they cannot recall (Finley, Benjamin, Hays, Bjork, & Kornell, 2011; Kornell, Bjork, & Garcia, 2011). Different feedback formats exist for this purpose: *show-answer feedback*, for example, presents the correct answer for restudy; more *elaborate feedback* presents additional explanations or requires the learner to make a new response. It is not clear which form of feedback is most efficient. Previous feedback research has shown that elaborate feedback can lead to better learning outcomes than show-answer feedback (Shute, 2008; van der Kleij, Eggen, Timmers, & Veldkamp, 2012). However, previous studies have not controlled for time on task (e.g., Finn & Metcalfe, 2010; Hall, Adams, & Tardibuono, 1968) and feedback formats have varied widely across studies so that it is unclear which particular elaborations are beneficial. The present study addresses this gap in the literature. We investigated one specific element of elaborate feedback, namely hints that create an extra opportunity for memory retrieval, while controlling for time on task. In three experiments, we compared how well students learned from retrieval practice with simple show-answer feedback and from retrieval practice with elaborate feedback with hints that created an extra opportunity for students to retrieve the correct answer from memory.

Empirical research regarding the creation of retrieval opportunities during the feedback phase is scarce. This is surprising given the substantial evidence from prior research that practicing memory retrieval is more beneficial than restudying information (for recent meta-analyses, see Adesope et al., 2017; Rowland, 2014). Moreover, providing learners with *scaffolds*, assistance to perform tasks that they cannot complete on their own, has a long tradition in education (Wood, Bruner, & Ross, 1976). Different research fields thus suggest that it should be more beneficial to give learners hints to retrieve an answer from memory than to just show learners the correct answer for restudy. Yet, only few feedback studies have included hints to help learners respond again (for reviews of feedback research, see Narciss & Huth, 2004; Shute, 2008). As a case in point, the most recent meta-analysis on feedback

interventions (van der Kleij, Feskens, & Eggen, 2015) identified 23 studies that compared show-answer feedback and more elaborate feedback. Of these, only 5 studies included elaborate feedback that prompted learners to make a new response. Most of these studies provided instructions about the application of complex skills such as mathematical operations or reading comprehension (e.g., Murphy, 2007, 2010, Narciss & Huth, 2014, in van der Kleij et al., 2015). Only one reviewed study, Hall, Adams and Tardibuono (1968), focused at hints that led to an attempt to retrieve information from memory.

This study by Hall et al. (1968) compared learners' retention of geographical facts after retrieval practice at the computer with two different forms of feedback. In one group, learners who made an error saw orthographic hints and tried again to type in the correct answer. In the other group, learners copied the correct response. The group who received hints took on average about 25% longer (100 minutes instead of 75 minutes) to study to criterion than the group who typed over the correct response. Nevertheless, recall accuracy on tests immediately and fourteen days after practice was the same in the two conditions. This suggests that the two feedback conditions led to similar learning results even though practice with hints feedback stimulated retrieval and led to longer overall study times.

A few studies have also reported the effect of hints specifically on the retention of items that learners initially could not retrieve correctly (Finn & Metcalfe, 2010; Kornell, Klein, & Rawson, 2015; Kornell & Vaughn, 2016). These studies led to different conclusions. On the one hand, Finn and Metcalfe (2010) found that participants who could not answer general knowledge questions remembered the correct response better when they constructed the response from a word fragment (*hints feedback*) than when they saw and typed over the correct response (*copy-answer-feedback*). This feedback effect was found on recall tests half an hour and a day after practice (but not on immediate tests). On the other hand, Kornell and Vaughn (2016) and Kornell et al. (2015) did not find benefits of hints feedback on immediate or delayed tests. They report four experiments in which participants first encoded weakly-associated word pairs and then attempted to recall the second word of each pair. After an incorrect response, participants either received copy-answer-feedback or hints feedback. Two experiments did not show any differences in learning outcomes between the two feedback conditions (Kornell et al 2015, Exp. 3a; Kornell and Vaughn, 2016); one experiment with a very large sample showed a small benefit of hints feedback (d = 0.13) (Kornell et al., 2015, Exp. 3b), and another experiment showed a benefit of show-answer feedback (Kornell et al., 2015, Exp. 5). Based on these results, the authors concluded that learning outcomes were roughly equivalent in the copy-answer-feedback condition and the hints feedback condition, and that it

did not matter whether learners were simply given the correct answer as feedback after a failed recall attempt or retrieved the answer from memory using hints (Kornell et al., 2015; Kornell & Vaughn, 2016).

In sum, the limited number of previous studies on hints feedback produced mixed results, but the overall image that emerges from these studies is that there is little, if any, evidence that hints feedback that creates an extra retrieval opportunity is more beneficial than show-answer feedback (Hall et al., 1968; Kornell et al., 2015; Kornell & Vaughn, 2016). The application of these results to realistic learning situations is hampered by design characteristics of the studies, however. First, all experiments except Hall et al. (1968) included only a single presentation per item. This is relevant because it is not realistic that learners engage in only a single retrieval. Repeated retrieval practice leads to better learning outcomes than a single retrieval (Pyc & Rawson, 2009; Rawson & Dunlosky, 2011), and repetition could change feedback effects. Elaborate feedback could, for example, influence subsequent repetitions of an item if it prevents repeated errors. Second, the lack of clear evidence in favor of orthographic hints feedback over show-answer feedback (Hall et al., 1968; Kornell et al., 2015; Kornell & Vaughn, 2016) could be due to the type of hints that were used. Kornell et al. (2015) and Kornell and Vaughn (2016) used orthographic hints that were constructed to almost always lead to a correct response (e.g., *wine – vine__r)*. Retrieval practice is, however, more beneficial, when it requires an effortful mental search for the correct answer (Pyc & Rawson, 2009). In addition, all previous studies that we found used orthographic hints. A well-known principle in learning research holds that semantic processing leads to better retention than processing of physical features such as orthography or pronunciation (Craik & Lockhart, 1972; Craik & Tulving, 1975). According to this *depth-of-processing* framework, hints that lead to more effortful retrieval and deeper, semantic processing might enhance retention more than the previously used orthographic hints. Thus, prior research leaves open whether it is beneficial for learning to provide hints feedback that stimulates effortful retrieval and semantic processing.

An additional characteristic to take into account in the study of hints feedback is time on task. Previous studies used a fixed amount of trials but such a design potentially favors hints feedback. By definition, hints feedback increases processing times after errors compared to standard feedback. This means that, with a fixed number of trials, the total exposure to the materials is longer in the hints condition. In contrast, when the total study time is limited, processing elaborate feedback costs time that cannot be spent on other forms of practice, such as further repetitions. Distributing time to feedback processing instead of further repetitions can in some cases reduce overall learning outcomes (Hays, Kornell, & Bjork, 2010). It is thus an open question whether providing hints feedback to create an additional retrieval opportunity is useful in a

realistic, time-constrained learning situation: If students have only a limited amount of time to practice – can this time better be spent on retrieval practice with simple show-answer feedback or on retrieval practice with hints feedback? The present study was set up to answer this question.

5.1.1 THE PRESENT STUDY

The central research question of this study was whether retrieval practice with hints feedback is more efficient than retrieval practice with show-answer feedback, measured in terms of performance on a recall test several days after practice. We investigated this question in three separate experiments, using three different types of hints. Students practiced vocabulary words in a foreign language by repeatedly translating the words from memory, a common learning activity for high school students. Scheduling of the repetitions was adaptive to learner performance and controlled with a learning system that modeled the memory strength per word based on the history of practice (for an extensive discussion of the system, see Sense, Behrens, Meijer, & van Rijn, 2016). Such adaptive scheduling produces better performance than repetitions in a random order or non-adaptive spacing strategies (e.g., Pavlik & Anderson, 2008) and has been applied successfully in foreign language courses (Lindsey, Shroyer, Pashler, & Mozer, 2014).

The main differences between the present study and earlier studies on hints feedback were as follows. First, we manipulated feedback in a realistic learning situation in a classroom setting, with students who engaged in repeated, spaced retrieval of vocabulary words. Second, the total study time was controlled. Students practiced for 15 minutes per condition to investigate which form of practice – retrieval with standard show-answer feedback or retrieval with hints feedback – most efficiently uses a limited amount of study time. Third, we conducted three separate experiments with different types of hints.

We used different types of hints in the present study in order to be able to draw broader conclusions about the effect of hints feedback compared to show-answer feedback than is possible with the existing studies with only orthographic hints. In Experiment 1, students received feedback with orthographic hints that consisted of word fragments similar to the hints used in previous studies. We chose fragments that were unlikely to be completed without prior exposure, in order to trigger effortful retrieval (see Carpenter & Delosh, 2006). In Experiments 2 and 3, the hints stimulated more semantic processing of the vocabulary words in order to trigger deeper processing, based on ideas on depth-of-processing (Craik & Lockhart, 1972; Craik & Tulving, 1975). The hints in Experiment 2 stimulated mental imagery to associate the presented vocabulary words to their meaning with so-called keywords (Atkinson, 1975), a mnemonic technique that works well in vocabulary learning (overview in

Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). The hints in Experiment 3 contained cognates of the to-be learned vocabulary words that students already knew. Cognates are words from different languages with a similar meaning and word form, such as "exit" (English) and "exitus" (Latin) (Dijkstra, Grainger, & van Heuven, 1999). The hints drew learners' attention to the overlap between the languages they knew, which is thought to facilitate vocabulary acquisition (Helms-Park & Dronjic, 2015; White & Horst, 2012).

In sum, this article reports the results of three experiments to answer the overarching question whether retrieval practice with orthographic, mnemonic or cross-language hints feedback leads to better learning outcomes than retrieval practice with show-answer feedback. This comparison was done in classroom studies using a realistic learning task and, unlike in previous studies, the total study time was equal in the two feedback conditions.

5.2 EXPERIMENT 1

Experiment 1 was conducted to compare how efficiently students learn from retrieval practice with feedback with orthographic hints or with show-answer feedback. The main outcome measure was the recall performance on a test seven days after practice. Performance on this test was used to measure how well the students remembered the meaning of the practiced words over time. We did not formulate a directional hypothesis about the effect of feedback on later recall due to the mixed results in previous research with hints feedback (Finn & Metcalfe, 2010; Hall et al., 1968; Kornell et al., 2015; Kornell & Vaughn, 2016). On the one hand, there is a large literature showing that retrieval practice enhances retention compared to restudy (e.g., Rowland, 2014), which suggests that hints that stimulate retrieval are beneficial. On the other hand, hints feedback changes the use of the available study time, and longer feedback processing that reduces the number of repetitions can impair learning outcomes (Hays et al., 2010).

In addition to feedback effects on later recall, we report measures to describe how the hints feedback changed the use of the available practice time. The reported measures are the (average) number of practiced words, the number of trials during practice, the number of errors per word, and the chance that students learned from errors, which was the average chance that students responded correctly to a word if they made an error on the previous presentation of a word. The most interesting measure among these is the number of practiced words, which depended directly on the number of trials but also on learner performance (i.e., the number and speed of *successful* retrieval trials), and is a good overall measure of the rate of acquisition

during practice. We expected that, due to the trade-off between spending time on hints feedback and spending time on additional repetitions, hints feedback would reduce the number of repetitions during practice. This could reduce the number of words practiced overall. However, hints feedback could also reduce the number of errors during practice if it increased the chance that students learned from errors. This in turn could increase the number of words in practice.

5.2.1 METHOD

5.2.1.1 PARTICIPANTS. A total of 108 students from Dutch high schools took part in the experiment. The data of 85 students (64.71 % female, M_{age} = 14.21 years, SD_{age} = 0.77) were analyzed. Of the 23 discarded datasets, 21 students had incomplete data because they were absent during one of the two sessions or experienced technical problems that led to incomplete or repeated practice blocks; one student did not provide consent to use his data, and one student was excluded because the log files showed that he had not followed instructions during the training. All students were in grade 8 or 9 in high-school tracks that prepare students for later admission to (applied) universities (the Netherlands has a tracked educational system after grade 6).

5.2.1.2 STIMULI. Seventy-two English words were selected from vocabulary lists from the last chapters of two schoolbooks of grade 9, which had not yet been covered in class. This was confirmed by low translation performance for unpracticed experimental words on the final test (*M* = 0.16, *SD* = 0.13) and subjective ratings that the students made about the number of words they already knew before practice (*M* = 0.19, *SD* = 0.20)[1]. Per practice block, up to 36 words were practiced, depending on each student's performance (using an adaptive learning system, see section *Retrieval practice*).

5.2.1.3 DESIGN AND EXPERIMENTAL CONTROL. The experiment had a within-subject design with Feedback Condition (Hints feedback or Show-answer feedback) as independent variable. There was a practice block of 15 minutes for each of the feedback conditions; the order of the two practice blocks was counterbalanced across participants. Words were randomly assigned to the two practice blocks for each student.

The main dependent variable of interest was the number of words that were recalled on the final test, and the need for recall prompts on this test. In addition, a number of measures to describe the practice phase were analysed. More detailed information is given in the next sections.

1 We report the proportion of words that were known rather than the absolute number of words in order to account for differences in the number of practiced words between students. The average subjective rating was calculated after excluding four students who reported already knowing more words than they had practiced.

5.2.1.4 RETRIEVAL PRACTICE. There was a practice block of 15 minutes for each of the two feedback conditions. Practice consisted of one initial study trial per word, followed by several retrieval trials. During the initial study trial, the English word was presented together with the Dutch translation and students retyped the translation. During the subsequent retrieval trials, only the English word was shown and students had to recall and type in the translation (see Figure 5.1).

An adaptive learning system was used to determine the order in which items were presented for each student, using a mathematical model to continuously estimate the accessibility in memory of each practiced word (a proxy for memory strength) based on the number, timing, accuracy and speed of previous retrievals during the study session (for a detailed description, see Sense et al., 2016). Briefly summarized, the purpose of this learning system is to maximize spacing of repetitions of each word while ensuring a high rate of retrieval success (Sense et al., 2016). This is achieved through step-wise addition of words to practice, and increasing spacing between repetitions. Such an approach leads to higher learning outcomes than common flashcard techniques and non-adaptive spacing models (Pavlik & Anderson, 2008; van Rijn, van Maanen, & van Woudenberg, 2009).

For the present study, the learning system (Sense et al., 2016) was used both to determine when to add more words to practice and when to repeat the words in practice. Normally, the system ensures high retrieval success during practice with around 70% of retrieval trials answered correctly. For the present experiments, the model parameters were adjusted to increase the delay between repetitions of words. This made the retrieval more difficult and elicited a higher number of errors, and thus more feedback moments in the limited practice time. With the changed settings, students answered on average 60% of the retrieval trials correctly in Experiment 1. Each word appeared on average about 6 times during the 15 minutes practice block (see Table 5.1 for descriptive statistics per condition). The delay between repetitions of the same word increased over the course of practice but summarized across all trials, words were on average repeated 83 seconds (median 79.0s) after the previous presentation.

Feedback. Feedback was given on all retrieval trials. In case of a correct answer, the word "correct" was displayed for 600 milliseconds. In case of an empty or incorrect response, corrective feedback was shown that differed between the two experimental conditions. In the show-answer feedback condition, the word and its translation were presented for four seconds with the instruction to "try to remember". In the hints feedback condition, *orthographic hints* were shown, that is, the first and the last grapheme of the response were shown with an instruction to try again (see Figure 5.1). The student could then submit another response, which was followed by the word "correct" in case of a correct response or by the show-answer feedback in case of an error.

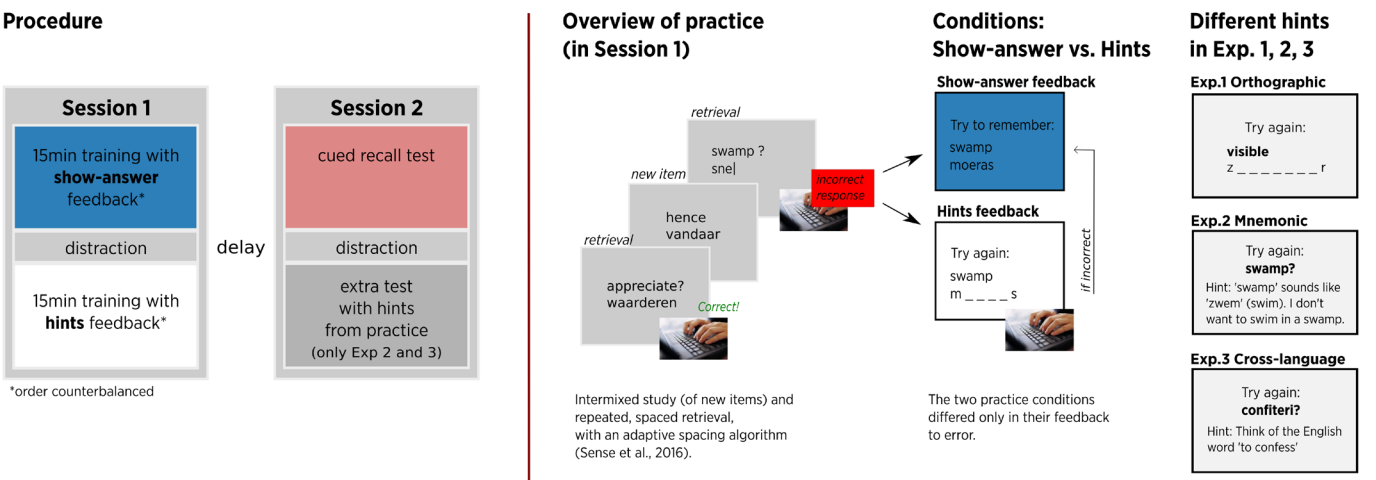


Figure 5.1 Overview of Experimental Procedure with Feedback Conditions. *Left:* Overview of the two sessions of the experiment with retrieval practice with experimental manipulation of the two different feedback conditions in Session 1, and performance measures with different recall tests in Session 2 several days later. *Middle*: The vocabulary practice consisted of intermixed studying of items with translations (when a new word was added to practice) and repeated, spaced retrieval. Shown are three English vocabulary items used in Experiment 1. In Experiment 1 and 2, students translated English vocabulary items into Dutch; in Experiment 3, students translated Latin items into German. The two feedback conditions differed only when students made an error: the show-answer feedback immediately revealed the correct answer; the hints feedback gave students a prompt to find the answer. *Right:* Examples of orthographic, mnemonic, and cross-language hints feedback. The instructions were in Dutch (Exp. 1, 2) or German (Exp. 3); they were translated for the figure.

Descriptives of the practice phase. Different descriptives of the practice phase have been included in Table 5.1. For the statistical analyses of feedback effects, we focused on the number of trials during practice, the number of words practiced in the 15 minutes practice time in each of the two conditions, the average number of errors made per word, and the chance that students learned from errors. This last measure was calculated as the probability that a word was translated correctly if it had been translated incorrectly on the previous trial, which was aggregated per learner across all (incorrect) practice trials.

5.2.1.5 TEST OF LEARNING OUTCOMES. Learning outcomes were measured seven days after practice with a translation test in which the English words were presented and students typed in the Dutch translation. The test was split into two blocks of 50 words, with a short rest break of about one minute between the blocks. All 72 experimental words were shown one by one, in a randomized order. The test thus included both the words that students had practiced and any experimental words that had not been added to practice in the limited amount of study time. On average, the students saw 36.6 (*SD* = 13.6) unpracticed words on the test. In addition to the experimental words, the test presented 28 easy control words that were selected from the vocabulary lists of the beginners' edition of the students' schoolbook series. These words were included in the test to ensure that there were at least some trials that the students could answer easily, and as an extra source of information to control whether students filled out the test conscientiously. The students all recalled at least 67.9% of the control words (*M* = 0.91, *SD* = 0.09), suggesting that they complied with the given instructions.

Recall prompts. In case of an incorrect response on the test, the same orthographic hints were shown as during practice in the orthographic-hint condition and students submitted another response. At the beginning of the test, students were instructed to try to translate the words as much as possible without resorting to the hints.

Scoring. Responses on the translation test were categorized as either correct or incorrect, with obvious spelling errors (e.g., *verdreit* instead of *verdriet*) being counted as correct. The number of correctly translated words was then calculated separately for the words practiced with hints feedback and for words practiced with show-answer feedback, as well as for the easy control words and the unpracticed experimental words.

Overall recall and need for prompts. The overall number of words translated correctly (short: *overall recall*) on the test was the main measure to describe learning outcomes. This measure is the total number of words that students translated correctly either directly (when presented with just the vocabulary word) or on the second attempt with orthographic recall prompts. To describe how much students

relied on the recall prompts, we also report the proportion of words which students translated only on a second attempt with prompts (short: *need for prompts*).

5.2.1.6 ADDITIONAL MEASURES. Students filled in pen and paper questionnaires at different moments during both sessions. These questionnaires were mainly used as distractors to introduce short breaks between the computer tasks and to obtain basic demographic information. We also obtained measures regarding students' prior knowledge and vocabulary learning strategies, which are not reported here because they are not directly related to the research questions.

5.2.1.7 PROCEDURE. The experiment consisted of two sessions (see Figure 5.1), which were conducted in a classroom setting during students' regular English lessons. The students worked individually at their computers. During the whole experiment, one or two researchers and the students' English teacher were present to ensure a quiet working atmosphere. Session 1 took 50 minutes. The session started with a brief group instruction, in which students were informed that they would practice vocabulary words with a computer program that adjusted practice to each student's learning rate. They then filled in a short pen and paper questionnaire and afterwards started the first practice block of 15 minutes by opening a link in the web browser. After the first practice block, students filled in another short pen and paper questionnaire. Then they underwent a second practice block of 15 minutes. After the second practice block, a third questionnaire was administered and students were told that the researchers would come back for a second practice session seven days later. In Session 2, the students first completed a sustained attention test, which took five minutes and is not reported here (Smilek, Carriere, & Cheyne, 2010). Afterwards, they took the recall test and completed a final questionnaire. The remainder of the second session was spent with debriefing.

5.2.1.8 STATISTICAL ANALYSES. We tested the effect of the within-subject factor Feedback Condition (Hints feedback or Show-answer feedback) on several dependent variables with two-sided t-tests for paired samples. The dependent variables describing learning outcomes on the final test were the total number of words that were recalled correctly and the need for prompts during the recall test (measured as the proportion of correct responses which students provided only after receiving recall prompts). The dependent variables describing the practice phase were the number of trials, the number of practiced words, the number of errors made during practice, and the chance that students learned from errors (i.e., the average chance that a word was translated correctly if it had not been translated correctly the previous time it was presented to the learner). Exact p-values are reported; to control for the number of statistical tests, an adjusted alpha value of 0.05 / 6 = 0.008 was used to determine significance and tests with p between 0.008 and 0.05 are reported as "numerical difference". We report as effect size Cohen's d corrected for

the correlation between paired observations (using Formula (3) in Dunlap, Cortina, Vaslow, & Burke, 1996, p.171).

In addition to classic t-tests, two-sided Bayesian paired t-tests (with a default Cauchy prior width of r = 0.707) were used to quantify the evidence for or against the null hypothesis, using the JASP software (Version 0.8.0.0, JASP Team, 2016). To increase readability, we always report the Bayes factor for the alternative hypothesis (BF_{10}). Values of BF_{10} smaller than 1 indicate evidence in favor of the null hypothesis; a BF_{10} larger than 1 indicates evidence in favor of the alternative hypothesis. A BF_{10} of 10 indicates, for example, that the observed data are 10 times more likely under the alternative hypothesis that there is a difference between the conditions, than under the null hypothesis that no difference exists. A BF_{10} of 0.2 indicates that the data are 0.2^{-1} = 5 times more likely under the null hypothesis than under the alternative hypothesis. We used a verbal classification scheme as proposed by Jeffreys (1961, in Wetzels & Wagenmakers, 2012) to interpret the evidence as "anecdotal" (1 < BF < 3), "moderate" (3 < BF < 10), "strong" (10 < BF < 30), or "very strong" (BF > 30). In case of a BF between 0 and 1, the inverse of the BF was calculated before applying this classification scheme.

5.2.2 RESULTS

Descriptive statistics about the practice phase have been included in Table 5.1; descriptive statistics about recall performance on the final test have been included in Table 5.2.

5.2.2.1 FEEDBACK EFFECTS ON LATER RECALL. On the test seven days after practice, overall recall was not significantly different for words from the practice block with show-answer feedback and for words from the practice block with hints feedback, $t(84) = 1.19$, $p = .24$, $d = 0.08$. A Bayes Factor BF_{10} of 0.24 (BF_{01} = 4.24) indicated moderate evidence for the null hypothesis. Overall recall on the test was calculated as the sum of the number of words that were recalled directly and the number of words that were recalled only on a second attempt with orthographic prompts. Further analyses revealed that students recalled a larger proportion of the words from the show-answer condition than from the hints condition directly on the first attempt, whereas they more often used orthographic prompts to recall the words from the hints condition, $t(80) = -2.90$, $p = .005$, $d = 0.40$[2]. Bayesian t-tests indicated that the evidence for this difference was moderate (BF_{10} = 5.84).

2 Degrees of freedom are 80 instead of 84 because four students recalled 0 words from at least one of the two practice blocks, which made it impossible to calculate the proportion of words that were translated with prompts.

5.2.2.2 FEEDBACK EFFECTS DURING THE PRACTICE PHASE. The number of trials that students went through in the 15 minutes of practice time and the number of words they practiced, were significantly higher in the show-answer condition than in the hints condition, $t(84) = 6.25$, $p < .001$, $d = 0.48$, and $t(84) = 4.22$, $p < .001$, $d = 0.36$. On average, students practiced 2.8 more words in the show-answer condition than in the hints condition; see Table 5.1 for descriptive statistics. Bayes factors (BF_{10}) of 722435 and 304 indicated very strong evidence for these differences between conditions. The number of errors that students made during practice was not significantly different between the two conditions, $t(84) = -1.39$, $p = .17$, $d = 0.15$, nor was the chance that students corrected their errors during practice (i.e., the chance that, after an error, students correctly responded on the next presentation of the same word), $t(84) = 1.84$, $p = .07$, $d = 0.18$. Bayes factors BF_{10} of 0.30 and 0.60 ($BF_{01} = 3.33$ and $BF_{01} = 1.69$) indicated moderate and anecdotal evidence for the null hypotheses.

Table 5.1 Descriptives of the Practice Phase with Show-Answer Feedback or Hints Feedback

	Experiment 1 *N* = 85	
Dependent variable	**Show-answer condition**	**Orthographic hints condition**
Number of words practiced in 15 minutes	*M* = 19.34*** *SD* = 8.13	*M* = 16.52 *SD* = 7.22
Total number of trials per practice block	*M* = 111.04*** *SD* = 27.05	*M* = 97.91 *SD* = 28.00
Number of incorrect responses per word	*M* = 2.25 *SD* = 1.71	*M* = 2.49 *SD* = 1.55
Chance for correct response at next presentation of a word, after an error	*M* = 0.48 *SD* = 0.19	*M* = 0.45 *SD* = 0.21
Proportion of correct responses	*M* = 0.62° *SD* = 0.17	*M* = 0.58 *SD* = 0.17
Proportion of correct responses to hints-feedback during practice	n/a	*M* = 0.45 *SD* = 0.21
Average time until next presentation of word (in seconds)	*M* = 82.33 *SD* = 15.59	*M* = 83.48 *SD* = 18.62
Number of presentations per word (incl. 1 study trial)	*M* = 6.34 *SD* = 1.79	*M* = 6.43 *SD* = 1.67

Note. The table provides descriptive statistics of the practice phase. All measures were aggregated at participant level and then at group level. The number of errors per word was first aggregated at word level, then at participant level. The asterisks indicate the p-value obtained when comparing the scores in the two feedback conditions with paired t-tests, with °$p < .05$, °°$p < .01$, or ***$p < .008$. Due to the number of statistical tests, only p-values below 0.008 were considered significant.

	Experiment 2 *N* = 90		**Experiment 3** *N* = 74	
	Show-answer condition	**Mnemonic hints condition**	**Show-answer condition**	**Cross-language hints condition**
	M = 18.31 *SD* = 7.73	*M* = 20.24*** *SD* = 7.41	*M* = 21.39° *SD* = 7.95	*M* = 19.46 *SD* = 7.15
	M = 105.60 *SD* = 26.43	*M* = 102.02 *SD* = 26.61	*M* = 109.36*** *SD* = 23.12	*M* = 96.28 *SD* = 22.83
	M = 2.21*** *SD* = 1.52	*M* = 1.39 *SD* = 1.00	*M* = 1.41 *SD* = 1.13	*M* = 1.36 *SD* = 1.14
	M = 0.50 *SD* = 0.19	*M* = 0.64*** *SD* = 0.22	*M* = 0.65*** *SD* = 0.20	*M* = 0.59 *SD* = 0.22
	M = 0.62 *SD* = 0.16	*M* = 0.71*** *SD* = 0.15	*M* = 0.73 *SD* = 0.13	*M* = 0.71 *SD* = 0.16
	n/a	*M* = 0.53 *SD* = 0.18	n/a	*M* = 0.69 *SD* = 0.21
	M = 82.03 *SD* = 16.49	*M* = 88.16°° *SD* = 16.27	*M* = 88.74 *SD* = 18.00	*M* = 91.82 *SD* = 22.22
	M = 6.31*** *SD* = 1.70	*M* = 5.35 *SD* = 1.17	*M* = 5.57 *SD* = 1.61	*M* = 5.30 *SD* = 1.26

Table 5.2 Learning Outcomes after Retrieval Practice with Show-Answer Feedback and Hints Feedback

	Experiment 1	
Dependent variable	**Show-answer condition**	**Orthographic hints condition**
Number of words recalled overall (on first attempt or on second attempt with orthographic prompts)	*M* = 7.17 *SD* = 6.74	*M* = 6.67 *SD* = 6.34
Proportion of words recalled only on second attempt with orthographic prompts	*M* = 0.25 *SD* = 0.30	*M* = 0.39*** *SD* = 0.32
Number of words recalled on test with mnemonic hints	n/a	n/a

Note. The table provides descriptive statistics of the learning outcomes on the final test after practice. Overall recall denotes the number of words that were recalled either directly on the first attempt, when students saw a vocabulary word and attempted to type in the translation in their native language, or on the second attempt, after students saw the first and the last grapheme of the correct answer as a prompt. In addition, Experiment 2 and 3 contained a separate test on which the hints from practice were presented together with the vocabulary words. The asterisks indicate the *p*-value obtained when comparing the scores in the two conditions with paired t-tests, marking significantly higher scores at °$p < .05$, °°$p < .01$, or ***$p < .008$. Due to the number of statistical tests, only p-values below 0.008 were considered significant.

5.2.3 DISCUSSION

Experiment 1 was conducted to compare the effects of retrieval practice with feedback with orthographic hints and retrieval practice with show-answer feedback. The main outcome variable of interest was the performance on a test seven days after learning; in addition, the effect of feedback during practice was analyzed. Overall recall performance on the final test was not significantly different when the show-answer feedback and orthographic hints feedback condition were compared. Students' test taking behavior indicated, however, that students translated a larger proportion of the words from the show-answer condition directly on the first attempt. For the words from the hints condition, students more often used recall prompts. Measures from the practice phase showed a higher number of trials during practice and a higher number of practiced words in the show-answer condition than in the hints condition, whereas errors during practice were not influenced by the feedback condition.

The finding that students relied more on recall prompts to translate the words from the hints condition than the show-answer condition, was unexpected. One possible explanation is that the hints feedback led to overall lower memory strength

Experiment 2		Experiment 3	
Show-answer condition	**Mnemonic hints condition**	**Show-answer condition**	**Cross-language hints condition**
M = 7.23	*M* = 7.56	*M* = 12.82	*M* = 11.20
SD = 6.51	*SD* = 5.92	*SD* = 6.77	*SD* = 6.57
M = 0.20	*M* = 0.25°	*M* = 0.32	*M* = 0.36
SD = 0.22	*SD* = 0.22	*SD* = 0.19	*SD* = 0.22
M = 7.53	*M* = 9.89***	*M* = 17.26	*M* = 15.63
SD = 6.67	*SD* = 6.62	*SD* = 7.82	*SD* = 7.27

than the show-answer feedback. More words from the hint condition may therefore have fallen below the recall threshold for direct (unprompted) recall on the test (see also the discussion of recall thresholds in the "bifurcation model" in Chapter 3). An alternative explanation is that students may have been unable to transfer what they practiced *with* hints to the test *without* hints. During practice, students likely used the hints feedback to find the correct translation by figuring out which of the practiced translations fit the presented word fragment, and this does not necessarily require the association of the foreign vocabulary word to the translation. However, such an association is needed for later recall of the word (e.g., Deconinck, Boers, & Eyckmans, 2015). The limited benefits of the hints feedback could thus be an example of context-dependent memory where later recall (here: recall of the translation) becomes dependent on cues that were available during practice (here: orthographic hints) (Smith & Handy, 2014, 2016).

The fact that students went through fewer repetitions and practiced fewer words in the hints condition than in the show-answer condition, is in line with earlier experiments in which spending time on hints feedback took time away from

further repetitions (Hays et al., 2010). The hints feedback by definition led to longer processing times on error trials, so that less repetitions fit into the 15 minutes of practice and fewer words were introduced. Surprisingly, although students spent extra time processing feedback after errors in the hints condition, they were not more likely to learn from feedback. The chance for a correct response on the next repetition after an error was similar in the two conditions, and so was the average number of errors per word. This suggests that responding to the hints feedback had few benefits, even during practice.

The lack of feedback effects on learning outcomes is in line with the majority of previous studies on hints feedback, which found no differences between orthographic hints and copy-answer feedback (Hall et al., 1968; Kornell et al., 2015; Kornell & Vaughn, 2016). For some of the earlier studies, the lack of effects could have been due to very easy hints (Kornell et al., 2015; Kornell & Vaughn, 2016). However, even with orthographic hints that required effortful processing, Experiment 1 did not show any clear benefits of hints feedback compared to show-answer feedback. This result contradicts benefits of orthographic hints feedback reported by Finn and Metcalfe (2010). There are several possible explanations for this. One is that Finn and Metcalfe presented feedback when participants could not answer general knowledge questions based on their prior knowledge. Thus, participants received feedback on information that they may have never encountered before and the hints may not have facilitated retrieval but rather been a means for students to encode the correct answer for the first time by constructing it from orthographic cues. Kornell et al. (2015) showed in control experiments that constructing a word from hints led to better encoding of new words compared to copying the words, whereas the same hints were not more beneficial as feedback compared to copy-answer feedback when presented after a failed retrieval. A second possible reason for differences between the present results and those reported by Finn and Metcalfe (2010) is that they did not control for time on task. The total study time was therefore likely longer in the hints feedback condition than in the show-answer condition, whereas it was equal in the two conditions in the present study.

Taken together, the results of Experiment 1 suggest that when feedback is manipulated in the context of spaced, repeated retrieval practice and the total study time is controlled, there are no clear benefits of responding to errors with orthographic hints that create an extra retrieval opportunity rather than with standard show-answer feedback. Learners might even perform better after practice with show-answer feedback because shorter feedback processing times allow them to go through more repetitions. Moreover, students' test taking behavior suggested that there is only limited transfer from practice with hints to recall tests without hints.

5.3 EXPERIMENT 2

Orthographic hints like those used in Experiment 1 can be easily automatically generated for computer-assisted learning and have been used as retrieval cues in previous memory studies (e.g., Finley et al., 2011). However, as we argued above, orthographic hints may focus learners' attention too much on the spelling of the response (here: the translation) rather than the association between the cue or question and the response (here: the vocabulary word and the translation). A more efficient way to strengthen this association might be to focus processing on semantic aspects of the to-be learned information with *keyword mediators* (Atkinson, 1975). Keyword mediators are an effective technique to encode the link between a vocabulary word and its meaning (e.g., Beaton, Gruneberg, Hyde, Shufflebottom, & Sykes, 2005; see Dunlosky et al., 2013 for a critical review). The technique involves two steps. First, the learner chooses a *keyword,* which is a known word that sounds similar to the new vocabulary. Next, the learner makes a meaningful association between the keyword and the translation, usually by forming a mental image. For example, to remember the meaning of the word "sorrow", a learner could choose the keyword "Zorro" and then create a mental image of Zorro feeling sorrow. Learners can generate keywords themselves but benefit in a comparable way from keywords generated by others (Shapiro & Waters, 2005). Therefore, we included experimenter-generated keywords in the hints feedback in Experiment 2.

Experiment 2 had a similar setup as Experiment 1 but the students now received feedback with mnemonic hints instead of orthographic hints, and took an extra recall test at the end of Session 2. On this test, the vocabulary words were presented together with the mnemonic hints from practice. The extra test with mnemonic hints allowed us to investigate whether the hints feedback during practice led to benefits only on a test with the same hints from practice, as the results of Experiment 1 suggest. Alternatively, mnemonic hints that enforce the link between vocabulary words and translations, could also enhance recall on a test without hints.

5.3.1 METHOD

The research design, materials and procedure in Experiment 2 were identical to those in Experiment 1, with a few exceptions outlined hereafter. The most important difference between the experiments was that we used mnemonic hints instead of orthographic hints in Experiment 2.

5.3.1.1 PARTICIPANTS. A total of 120 Dutch high school students took part in the experiment. The data of 90 students (56.7 % female, M_{age} = 13.96 years, SD_{age} = 0.76) were analyzed; the other students had incomplete datasets because they were absent during one of the lessons or experienced technical problems that led to

incomplete or repeated study blocks. The students who participated in Experiment 2 had not participated in Experiment 1.

5.3.1.2 DESIGN AND EXPERIMENTAL CONTROL. All students first practiced in the show-answer condition and then in the hints condition in order to prevent differential transfer from the block with mnemonic hints to the block without hints. This is common in studies on mnemonic techniques (e.g., Fritz, Morris, Acton, Voelkel, & Etkind, 2007).

5.3.1.3 RETRIEVAL PRACTICE.

Feedback. In case of an incorrect response, students in the hints feedback condition were presented with a mnemonic hint. This hint consisted of a keyword and a sentence that linked the keyword to the Dutch translation. For example, if a student failed to fill in the translation of the word "vain", the hint was: "*vain..* fijn*. Als je er altijd fijn wilt uitzien, dan ben je _____"* [Engl. approximately: "*vain .. pretty. If you always want to look pretty, you are _____"*].

5.3.1.4 TEST OF LEARNING OUTCOMES. The students first took the same translation test as in Experiment 1, to measure overall recall and need for prompts. Afterwards, the students took an additional separate test during which the words were presented one by one together with the mnemonic hint used in the practice phase of the hints condition. Performance on this test is called *recall with mnemonic prompts.*

5.3.1.5 PROCEDURE. The procedure was identical with Experiment 1, but the students did no sustained attention task at the beginning of Session 2 and did an extra recall test with mnemonic prompts at the end of Session 2. Between the first recall test and the extra recall test with mnemonic prompts, students filled in a short questionnaire about the effort they invested on the test and their English grades.

5.3.2 RESULTS

5.3.2.1 FEEDBACK EFFECTS ON LATER RECALL. The overall number of words that students recalled was not significantly different in the two conditions, $t(89) = -0.65$, $p = .52$, $d = 0.05$, with a Bayesian t-test indicating moderate support for the null hypothesis that no difference exists between the conditions, $BF_{10} = 0.14$ ($BF_{01} = 7.14$). The need for recall prompts on the test did not differ significantly between conditions, $t(84) = -2.30$, $p = .02$, $d = 0.28$ (alpha corrected for multiple comparisons = .008), although numerically, students used more orthographic prompts to recall the words from the hints condition. Bayesian analyses indicated that the evidence concerning the (lack of) difference between the conditions on the need for recall prompts was inconclusive, $BF_{10} = 1.44$.

The two feedback conditions were also compared on recall on a separate test with mnemonic prompts. This recall measure was higher in the hints condition than

in the show-answer condition, $t(87)^3$ = -4.50, p < .001, d = 0.35. A Bayesian t-test showed that the evidence for this difference between the feedback conditions was very strong, BF_{10} = 809.80.

5.3.2.2 FEEDBACK EFFECTS DURING THE PRACTICE PHASE. The number of trials that participants went through was approximately the same in the two conditions, $t(89)$ = 1.74, p = .09, d = 0.13. However, students practiced 1.9 more words in the hints condition than in the show-answer condition, $t(89)$ = -3.28, p = .001, d = 0.26. Bayesian t-tests indicated inconclusive evidence regarding the (absence of) differences in the number of trials, BF_{10} = 0.49, but strong evidence for the higher number of words practiced in the hints condition, BF_{10} = 16.3. The number of errors was also significantly lower in the hints condition (M = 1.39) than in the show-answer condition (M = 2.21), $t(89)$ = 6.42, p < .001, d = 0.60, with BF_{10}= 1720000, indicating very strong evidence. When students made an error, the chance that they translated the word correctly the next time it came up was higher when they had received hints feedback than when they had received show-answer feedback, $t(89)$ = -6.31, p < .001, d = 0.67, BF_{10} = 1070763.

5.3.3 DISCUSSION

In Experiment 2, we tested whether retrieval practice with feedback with mnemonic hints is more efficient than retrieval practice with standard show-answer feedback. Differences between the two feedback conditions were found on the recall test with hints and during practice. As in Experiment 1, the overall number of words that were recalled on the final test one week after practice was not significantly different in the two feedback conditions. The need for recall prompts was also not significantly different, although there was a trend that students used more prompts to recall the words from the hints condition. Consistent and strong evidence for a difference in recall between the conditions was only found on a separate test on which words were presented with the mnemonic hints from practice. On this test, students showed better recall for the words from the hints condition than for the words from the show-answer condition. Likely, students were better able to find the correct translation based on the mnemonic hints presented at test if they had seen those hints before during practice. This suggests that students remembered the hints from practice and recognized them on the test, but their knowledge of the hints did not enhance their recall performance on the test without hints. Benefits of practice with hints did not transfer to a recall situation without hints, as in Experiment 1.

3 The degrees of freedom differ from the other recall measures because two participants did not complete the extra test with mnemonic hints within the lesson.

During practice, the total number of trials was similar in the two conditions, but a higher number of words were practiced in the hints condition than in the show-answer condition. This was the case because students made fewer errors in the hints condition. Students were also more likely to answer correctly on the next repetition of a word when they got hints feedback after an error than when they got show-answer feedback. These results are promising because they suggest that students may have benefited from the mnemonic hints feedback and were less likely to repeat errors during practice. However, differences during practice did not lead to differences in learning outcomes measured on the final test.

A limitation of Experiment 2 is that we did not counterbalance the order in which the two conditions were presented. This was a conscious decision to avoid the possibility that students who first practiced with mnemonic hints would then exploit the keyword strategy during practice in the show-answer condition. Differential transfer has also been avoided with such a fixed presentation order in other within-subject studies on mnemonic techniques (e.g., Fritz et al., 2007). To get an estimate of possible order effects in Experiment 2, the data of Experiment 1 - in which the practice blocks were counterbalanced - were re-analyzed. These control analyses showed that learning outcomes and all measures of practice efficiency except for the chance to learn from errors, were better for words practiced in Block 2 than in Block 1 (averaged across feedback conditions). This order effect in Experiment 1 suggests that some performance measures may indeed have been enhanced in the hints condition in Experiment 2 because the mnemonic hints were always presented in Block 2[4]. These possible order benefits for the hints condition might have contributed to the lower number of errors found during practice in the hints condition compared to the show-answer condition. However, in spite of possible order benefits, later learning outcomes did not differ significantly between the two feedback conditions in Experiment 2. This further strengthens the conclusion that mnemonic hints feedback did not benefit learning.

In sum, Experiment 2 did not reveal clear benefits of mnemonic hints feedback over show-answer feedback. In the practice block with mnemonic hints feedback, students made fewer errors and practiced a higher number of words. Nevertheless,

4 We also did a second control analysis in which we included only the data of the 39 students who first practiced in the show-answer condition and then in the hints condition in Experiment 1. This is the same order that all students underwent in Experiment 2. In this selection of students in Experiment 1, none of the obtained learning outcomes or practice measures differed significantly between the feedback conditions, and Bayesian tests indicated anecdotal or moderate evidence for the null hypotheses. The same counterbalancing order thus did not show benefits of orthographic hints in Experiment 1 but did show benefits of mnemonic hints during practice in Experiment 2. This suggests that relative to show-answer feedback, mnemonic hints seem to be more beneficial than orthographic hints.

later overall recall was the same in the two conditions. This was the case even though the order in which the practice blocks were presented may have created an advantage for the hints condition. The only clear benefit of hints feedback on later recall was found on a test on which words were presented together with the same mnemonic hints from practice. This suggests that, again, possible benefits of the hints did not transfer to recall situations without hints.

5.4 EXPERIMENT 3

In Experiment 3, we presented students with cross-language hints. These hints were similar to Experiment 2 in that they contained a keyword to help the learners associate the vocabulary words to their translation. This time, however, the keyword was a cognate rather than an arbitrarily chosen word. Cognates are words from different languages that have similar phonological and/or orthographical forms and are often semantically related, like "exit" (English) and "exitus" (Latin).

Cognates tend to be recognized and learned more efficiently than noncognates (e.g., Helms-Park & Dronjic, 2015; Rogers, Webb, & Nakata, 2015). However, learners often fail to recognize cognates (Moss, 1992) and increasing learners' cognate awareness might be beneficial for word learning (e.g., White & Horst, 2012). We therefore used cross-language hints in Experiment 3, which drew students' attention to the cognate status of the to-be-learned words. The target language that students practiced in Experiment 3 was Latin and the hints contained cognates from the students' second language, English. For example, when students were trying to translate the Latin word "*exitus*" into their first language German (de. *Ausgang* [exit]), the hint was: "*Try again! Think of the English word 'exit'*!". We expected that these hints would help students associate the Latin vocabulary to their prior knowledge, and thereby enhance retention.

5.4.1 METHOD

5.4.1.1 PARTICIPANTS. A total of 88 German high school students took part in the experiment. The data of 74 students (59.5% female, M_{age} = 13.2 years, SD_{age} = 0.6) were analyzed; the other students had incomplete datasets because they were absent during one of the lessons or experienced technical problems that led to incomplete or repeated study blocks. The students were in grade 7 and 8 of a grammar school that prepared them for university education. Students had learned English at school for 5.9 years on average (*SD* = 1.17), and had learned Latin for less than three years (*M* = 2.2, *SD* = 0.6).

5.4.1.2 STIMULI. Seventy-two Latin words were selected from vocabulary lists from the last chapters of the schoolbook of grade 8, which the students had not yet studied according to their teachers. This was confirmed by a low proportion of unpracticed experimental words that were correctly translated on the final test (*M* = 0.18, *SD* = 0.14) and subjective ratings that the students made about the number of words they already knew (*M* = 0.06, *SD* = 0.07).

5.4.1.3 DIFFERENCES BETWEEN EXPERIMENT 2 AND EXPERIMENT 3.

Design and experimental control. In Experiment 3, the order of the two practice blocks (hints feedback or show-answer condition) was counterbalanced across participants.

Feedback during retrieval practice. Students in the hints feedback condition were presented with a cross-language hint when they made an error. The target language that students practiced in Experiment 3 was Latin and the hints contained cognates from the students' second language, English. For example, when students were trying to translate the Latin word "*procedere*" (de. fortfahren [proceed]), the hint was: "*Try again! Think of the English word 'to proceed' (vorgehen, fortfahren)!*". To ensure that students understood the English keywords, the hints contained either German translations of the keyword or a brief phrase from which its meaning could be derived. For example, for the word *honestus* (de. ehrlich, Engl. honest], the hint was:*"Try again! Think of the English word 'honest' (e.g., 'an honest answer' = 'eine ehrliche Antwort)*"! The hints were designed in such a way that students could not just type over the only German translation in the hint but had to think about the meaning of the keyword. The keywords were chosen from a database of etymological relations between Latin and English, published by linguists and language teachers (Gerbrandy et al., 2014).

Test of learning outcomes. The delayed test took place three days after learning instead of a week later, as in Experiment 1 and 2. During Session 2, students first took the same recall test as in Experiment 1 and 2, and then a test that presented the vocabulary words with the cross-language hints from practice.

5.4.2 RESULTS

5.4.2.1 FEEDBACK EFFECTS ON LATER RECALL. The overall number of words that students recalled on the final test, was not significantly different in the show-answer condition and in the hints condition, $t(73) = 2.60$, $p = .01$, $d = 0.24$ (alpha = 0.008), but was numerically higher in the show-answer condition. The proportion of words for which students needed recall prompts on the test was not significantly different in the two conditions, $t(73) = -1.37$, $p = .18$, $d = 0.19$, $BF_{10} = 0.31$. On the separate test with the cross-language hints from practice, students did not perform significantly different for words from the show-answer condition and the hints

condition, $t(72) = 2.46$, $p = .016$, $d = 0.22$[5], but again showed numerically higher results in the show-answer condition. Bayesian t-tests indicated anecdotal evidence for differences between the conditions on overall recall and recall on the test with mnemonic hints (BF_{10} respectively 2.9 and 2.14).

5.4.2.2 FEEDBACK EFFECTS DURING THE PRACTICE PHASE. On average, students went through significantly more repetitions in the show-answer condition than in the hints condition, $t(73) = 5.73$, $p < .001$, $d = 0.60$, $BF_{10} = 65493$ (very strong evidence). Students numerically practiced more words in the show-answer condition than in the cross-language hints condition, $t(73) = 2.52$, $p = .01$, $d = 0.25$, $BF_{10} = 2.4$, but this difference was not significant at alpha = 0.008 and Bayesian analyses indicated only anecdotal evidence. The number of errors that students made per word, was relatively low ($M_{Hints} = 1.36$ and $M_{ShowAns} = 1.41$), and not significantly different between conditions, $t(73) = -0.35$, $p = .73$, $d = 0.04$, $BF_{10} = 0.14$ (moderate evidence for H_0). The chance that students gave a correct response on the next repetition of a word after a previous error, was higher in the show-answer condition than in the cross-language hints condition, $t(73) = 2.75$, $p = 0.0074$, $d = 0.30$, $BF_{10} = 4.19$.

5.4.3 DISCUSSION

In Experiment 3, cross-language hints were used to compare learning outcomes and practice behavior of retrieval practice with hints feedback and with show-answer feedback. We found no significant difference in learning outcomes between the two feedback conditions. On the recall tests three days after practice, overall recall and recall with the cross-language hints from practice were not significantly different between the two conditions. Performance was even numerically higher in the show-answer condition than in the hints condition, but effects were not significant after correction for multiple comparisons. Test taking behavior did not differ significantly between conditions; students needed recall prompts on the final test about equally often for words practiced with cross-language hints feedback and for words practiced with show-answer feedback.

During practice, students went through significantly more repetitions in the show-answer condition than in the cross-language hints condition, as in Experiment 1 and 2. Surprisingly, students were also more likely to respond correctly on the next repetition of a word if they had received show-answer feedback after an error than if they had received cross-language hints. A possible explanation for this could be that students changed their response behavior during practice when they received hints feedback and more readily submitted incorrect responses because they knew

5 Degrees of freedom are different from the other two recall measures because one student did not complete the extra test with mnemonic hints.

that they would get a second chance with a hint. There was, however, no significant difference in the proportion of correct answers overall during the practice blocks with and without hints-feedback.

A possible reason why the cross-language hints in Experiment 3 did not reduce the number of errors during practice is that students made only few errors overall. Experiment 3 was conducted with a group of grammar school students who learned two foreign languages, Latin and English. This relatively homogeneous group of high-performing students made only 1.4 errors on average per word even in the show-answer condition. With such a small number of feedback moments, the hints could only have a limited effect (potentially reducing the number of errors from 1.4 to 1 per word). Moreover, the low number of errors indicates that students did not need elaborate support to learn from errors. This could explain why the cross-language hints did not reduce errors during practice even though drawing students' attention to cognates is thought to be helpful (Helms-Park & Dronjic, 2015). Finally, unlike Experiments 1 and 2, Experiment 3 did not show specific advantages of practice with hints feedback on the recall test with the hints from practice. A possible explanation for this could be that students were successful at deciphering the hints even without prior exposure.

Overall, Experiment 3 provided further evidence that during repeated retrieval practice, there are no clear benefits of responding to errors with hints that create an extra retrieval opportunity rather than with standard show-answer feedback. Nonsignificant numerical differences between the show-answer condition and the hints condition even suggest that the hints had a small negative effect on learning outcomes. A possible reason for this is that the longer feedback processing after errors in the hints condition reduced the number of repetitions possible in the limited study time, but did not reduce errors.

5.5 GENERAL DISCUSSION

In spite of a large literature on feedback effects in general (Narciss & Huth, 2004; Shute, 2008), relatively little is known about specific elaborations that could make elaborate feedback more efficient compared to simple show-answer feedback (van der Kleij et al., 2012). The present study provides information about a specific elaboration, namely memory retrieval, in the context of a realistic learning situation in the classroom. We conducted three experiments to compare how well high-school students learned from a computerized vocabulary training procedure that included either standard show-answer feedback or hints feedback in case of an error. Show-answer feedback consisted of the display of the correct answer; hints feedback

consisted of a hint to give the students a second chance to retrieve the correct answer from memory. Learning outcomes, measured as the number of words that were recalled on a test several days after learning, and practice behaviors were compared between the two feedback conditions. Overall, learning outcomes were similar in the two feedback conditions in the three experiments except when the hints from practice were available again during the recall test. During practice, hints feedback led to a shift of time from further repetitions to longer feedback processing after errors in all three experiments, but only the mnemonic hints feedback in Experiment 2 also reduced errors during practice.

Both an extensive literature on the benefits of retrieval practice (Adesope et al., 2017; Rowland, 2014) and the long-standing tradition in education to work with scaffolds (Wood et al., 1976), suggest that hints feedback is beneficial for later recall. However, we did not find better learning outcomes after practice with hints feedback compared to show-answer feedback. Neither orthographic (Exp. 1), mnemonic (Exp. 2), nor cross-language hints (Exp. 3) led to higher overall recall compared with standard show-answer feedback. Experiment 3 even showed numerically higher recall in the show-answer condition than in the cross-language hints condition. The only significant differences between the feedback conditions were found when the hints from practice were available again on the recall test. Students used significantly more orthographic recall prompts on the final test for the words from the orthographic hints condition than for the words from the show-answer condition (Exp. 1), and recall on a separate test with mnemonic prompts was higher for words from the mnemonic hints condition than for words from the show-answer condition (Exp. 2). This could be an example of transfer-appropriate processing (Morris, Bransford, & Franks, 1977), the phenomenon that memory performance depends on the match between practice and test. Specifically, recall might have become dependent on the hints if the hints feedback strengthened an association between the hints and the correct response but this association could not be reactivated when the hints were later absent (for more information on context-dependent memory see Smith & Handy, 2016). Tentatively, this could mean that effects of practice with hints do not transfer to recall situations without hints.

The lack of general benefits of hints feedback for later recall replicates the results of three prior studies on hints feedback which also found no difference in recall after practice with (orthographic) hints and show-answer feedback (Hall et al., 1968; Kornell et al., 2015; Kornell & Vaughn, 2016). For the earlier studies, this lack of effects could have been due to the use of hints that were completed too easily (Kornell et al., 2015; Kornell & Vaughn, 2016) or triggered only superficial processing of orthographic features (Hall et al., 1968). However, even with orthographic hints that required effortful processing and with carefully constructed mnemonic and

cross-language hints, the present study did not show clear benefits of hints feedback compared to show-answer feedback. Clearly, adding hints feedback to repeated retrieval practice does not in every case enhance learning outcomes. To understand this result, it is necessary to take into account the effect of hints feedback during practice under time constraints.

In the present study, we controlled the total practice time to ensure that the feedback manipulation was not confounded by longer practice times in the hints condition than in the show-answer condition (as, for example, Hall et al., 1968). As a consequence, hints feedback could influence mainly two aspects of practice: the duration of feedback processing after errors (which was by definition longer in the hints condition than in the show-answer condition), and the chance that students learned from feedback and did not repeat errors. These effects, in turn, influenced the number of (successful) retrieval trials that students could go through in the available practice time and thereby the number of words in practice (see also Hays et al., 2010). In Experiments 1 and 3, the hints feedback did not significantly reduce (repeated) errors, and therefore resulted in a significantly lower number of trials and lower number of practiced words in the hints condition than in the show-answer condition. In Experiment 2, the mnemonic hints had more positive effects, but these need to be interpreted with some caution due to possible order effects. During practice with mnemonic hints feedback, students made significantly fewer (repeated) errors, and this led to a higher number of words practiced in the hints condition than in the show-answer condition[6].

A number of characteristics of this study need to be taken into account when generalizing conclusions to other learning situations. First, the learning system that controlled the spacing of repetitions during practice (Sense et al., 2016) was changed to produce a relatively high error rate during practice, compared to its default settings. The error rates of 30 to 40% in the present study were, however, still lower

6 The differences between the two feedback conditions in the number of words practiced may have influenced later learning outcomes. Therefore, we replicated all analyses using as dependent variable the *proportion* instead of the absolute number of practiced words that were recalled on the test. These control analyses led to the same conclusions as the analyses reported before: The feedback conditions only differed significantly from each other in Experiment 1 and 2 on test trials with the hints from practice; all other recall measures were not significantly different. This was the case even though learners tend to remember a larger proportion of items from smaller study sets, which could have created an advantage for the hints condition (a "list-length effect", Gillund & Shiffrin, 1984). Thus, there is no reason to assume that a positive effect of hints feedback on retention was covered up by the (lower) number of words practiced in the hints condition. The control analyses showed that even when the number of words practiced was taken into account, no significant difference in later overall recall existed between the feedback conditions.

than the error rate in the single previous study that found positive effects of hints feedback (Finn & Metcalfe, 2010, who reported an initial error rate of 70 to 78% before feedback). We used these adjusted settings to evoke enough errors during practice to observe differences between the conditions, since the feedback conditions differed only on error trials. Feedback effects may be different if fewer errors occur during practice so that the few feedback moments draw more attention. This could, for example, motivate students to more thoroughly encode mnemonic or cross-language hints, and increase benefits of hints. Similarly, the effect of hints feedback might be different if hints are only given after certain incorrect responses, for example, only for words that were repeatedly translated incorrectly. These could be starting points for further research, although based on the present results it is questionable how large the benefits of hints feedback would be. Second, we used an effective baseline practice condition – adaptive, repeated, spaced retrieval (Delaney, Verkoeijen, & Spirgel, 2010; Pavlik & Anderson, 2008). The trade-off between longer feedback processing and further practice trials may be different when the baseline is not as effective. However, a less effective baseline is also less relevant for practical purposes. Third, the effect of hints feedback might, to some extent, depend on the retrieval task. In the present experiments, students responded to newly learned vocabulary words by typing in words from their native language. Hints might, for example, have more benefits when the translation direction is reversed and learners must retrieve a newly learned word form. In this case, hints may not be a distraction from the association between word form and translation but rather a trigger to deeply process the new word form, which students still need to encode. However, such a recall task might require more sophisticated feedback that points out spelling errors rather than the simple display of the first and last grapheme of the correct response.

5.5.1 CONCLUSION

We found no clear benefits of feedback that created an extra retrieval opportunity with hints compared to show-answer feedback. This was an unexpected finding given the large support for benefits of retrieval practice in general (Adesope et al., 2017; Rowland, 2014). Similar to Kornell et al. (2015) we have to conclude that a manipulation that otherwise enhances learning outcomes (i.e., retrieving a word from memory instead of restudying the complete word) is not automatically also a beneficial addition to feedback after a failed retrieval attempt. This finding has practical implications for the design of training conditions. More is clearly not always more: even a seemingly beneficial manipulation, such as replacing show-answer feedback with hints feedback, incurs costs because it takes time away from other forms of practice. Such time costs can be acceptable if the hints reduce further errors during practice (as the mnemonic hints in Experiment 2), but can impair learning

outcomes if practice with hints does not support later recall without hints (Experiment 1) or students do not need extra support to correct their errors (Experiment 3). Moreover, students did not automatically transfer what they practiced *with* hints to a recall situation *without* hints. It is important to consider carefully whether hints that support retrieval during practice lead to associations that can also be used when the hints are later not available anymore.

5.6 REFERENCES

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Atkinson, R. C. (1975). Mnemotechnics in second-language learning. *American Psychologist*, *30*(8), 821–828. https://doi.org/10.1037/h0077029

Beaton, A., Gruneberg, M., Hyde, C., Shufflebottom, A., & Sykes, R. (2005). Facilitation of receptive and productive foreign vocabulary learning using the keyword method: The role of image quality. *Memory*, *13*(5), 458–471. https://doi.org/10.1080/09658210444000395

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*(2), 268–276. https://doi.org/10.3758/BF03193405

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

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

Deconinck, J., Boers, F., & Eyckmans, J. (2015). "Does the form of this word fit its meaning?" The effect of learner-generated mapping elaborations on L2 word recall. *Language Teaching Research*, *21*(1), 31–53. https://doi.org/10.1177/1362168815614048

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation*, *53*, 63–147. https://doi.org/10.1016/S0079-7421(10)53003-2

Dijkstra, T., Grainger, J., & van Heuven, W. J. B. (1999). Recognition of cognates and interlingual homographs: The neglected role of phonology. *Journal of Memory and Language*, *41*(4), 496–518. https://doi.org/10.1006/jmla.1999.2654

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*(2), 170–177. https://doi.org/10.1037/1082-989X.1.2.170

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Finley, J. R., Benjamin, A. S., Hays, M. J., Bjork, R. A., & Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language*, *64*(4), 289–298. https://doi.org/10.1016/j.jml.2011.01.006

Finn, B., & Metcalfe, J. (2010). Scaffolding feedback to maximize long-term error correction. *Memory & Cognition*, *38*(7), 951–961. https://doi.org/10.3758/MC.38.7.951

Fritz, C. O., Morris, P. E., Acton, M., Voelkel, A. R., & Etkind, R. (2007). Comparing and combining retrieval practice and the keyword mnemonic for foreign vocabulary learning. *Applied Cognitive Psychology*, *21*(4), 499–526. https://doi.org/10.1002/acp.1287

Gerbrandy, P., Castricum, J., Hermsen, C., Hupperts, C., Raiijmakers, M., Risselada, R., .. et al. (2014). Janus. Connecties tussen het Latijn en moderne Europese talen. [Janus. Connections between Latin and modern European languages]. Retrieved from http://janus.humanities.uva.nl/

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1–67. https://doi.org/10.1037/0033-295X.91.1.1

Hall, K. A., Adams, M., & Tardibuono, J. (1968). Gradient- and full-response feedback in computer assisted instruction. *Journal of Educational Research*, *61*(5), 195–199.

Hays, M. J., Kornell, N., & Bjork, R. A. (2010). The costs and benefits of providing feedback during learning. *Psychonomic Bulletin & Review*, *17*(6), 797–801. https://doi.org/10.3758/PBR.17.6.797

Helms-Park, R., & Dronjic, V. (2015). Crosslinguistic lexical influence: Cognate facilitation. In R. Alonso (Ed.), *Crosslinguistic Influence in Second Language Acquisition* (Vol. 95). Bristol, UK: Multilingual Matters.

JASP Team. (2016). *JASP (Version 0.8.0.0)[Computer software]*. Retrieved from https://jasp-stats.org/

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. https://doi.org/10.1016/j.jml.2011.04.002

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 283–294. https://doi.org/10.1037/a0037850

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. *Psychology of Learning and Motivation*, *65*, 183–215. https://doi.org/10.1016/bs.plm.2016.03.003

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological Science*, *25*(3), 639–647. https://doi.org/10.1177/0956797613504302

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Moss, G. (1992). Cognate recognition: Its importance in the teaching of ESP reading courses to Spanish speakers. *English for Specific Purposes*, *11*(2), 141–158. https://doi.org/10.1016/S0889-4906(05)80005-5

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multimedia learning. In H. M. Niegeman, D. Leutner, & R. Brünken (Eds.), *Instructional Design for Multimedia Learning* (pp. 181–195). Berlin: Waxmann.

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*(2), 101–117. https://doi.org/10.1037/1076-898X.14.2.101

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*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*(3), 283–302. https://doi.org/10.1037/a0023956

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

Rogers, J., Webb, S., & Nakata, T. (2015). Do the cognacy characteristics of loanwords make them more easily learned than noncognates? *Language Teaching Research*, *19*(1), 9–27. https://doi.org/10.1177/1362168814541752

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

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An individual's rate of forgetting is stable over time but differs across materials. *Topics in Cognitive Science*, *8*(1), 305–321. https://doi.org/10.1111/tops.12183

Shapiro, A. M., & Waters, D. L. (2005). An investigation of the cognitive processes underlying the keyword method of foreign vocabulary learning. *Language Teaching Research*, *9*(2), 129–146. https://doi.org/10.1191/1362168805lr151oa

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Smilek, D., Carriere, J. S. A., & Cheyne, J. A. (2010). Failures of sustained attention in life, lab, and brain: Ecological validity of the SART. *Neuropsychologia*, *48*(9), 2564–2570. https://doi.org/10.1016/j.neuropsychologia.2010.05.002

Smith, S. M., & Handy, J. D. (2014). Effects of varied and constant environmental contexts on acquisition and retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1582–1593. https://doi.org/10.1037/xlm0000019

Smith, S. M., & Handy, J. D. (2016). The crutch of context-dependency: Effects of contextual support and constancy on acquisition and retention. *Memory*, *24*(8), 1134–1141. https://doi.org/10.1080/09658211.2015.1071852

van der Kleij, F. M., Eggen, T. J. H. M., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education*, *58*(1), 263–272. https://doi.org/10.1016/j.compedu.2011.07.020

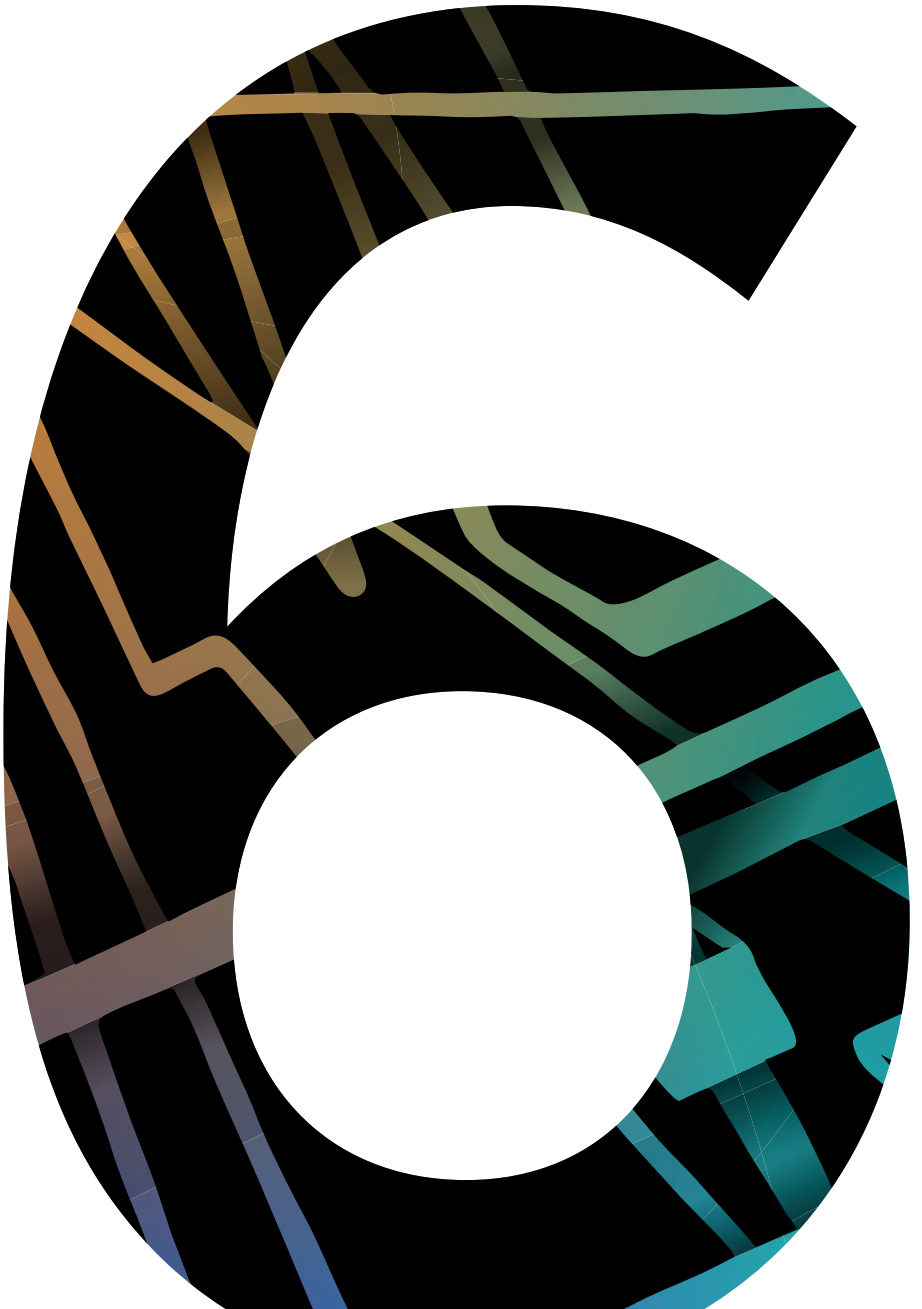
van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, *85*(4), 475–511. https://doi.org/10.3102/0034654314564881

van Rijn, H., van Maanen, L., & van Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. In A. Howes, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 9th International Conference on Cognitive Modeling* (pp. 110–115). UK: Manchester.

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*(6), 1057–1064. https://doi.org/10.3758/s13423-012-0295-x

White, J. L., & Horst, M. (2012). Cognate awareness -raising in late childhood: teachable and useful. *Language Awareness*, *21*(1–2), 181–196. https://doi.org/10.1080/09658416.2011.639885

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x



EFFECTS OF CONTEXTUAL RICHNESS ON WORD RETENTION: MEMORY RETRIEVAL VERSUS INFERENCES

An article based on this chapter has been submitted for publication, as: van den Broek, G.S.E., Takashima, A., Segers, E., & Verhoeven, L. Effects of Contextual Richness on Word Retention: Memory Retrieval versus Inferences.

Stimuli and datasets can be accessed through the Open Science Framework, at https://osf.io/eujyn/?view_only=679d9ff5e4764906be5854b977d5f844

Abstract. Learning foreign vocabulary from context typically requires multiple encounters during which the word meaning can be retrieved from memory or inferred from context. We compared the effect of memory retrieval and context-inferences on short-term and long-term word retention in three experiments. Participants studied novel foreign words and then practiced the words either in an uninformative context that required the retrieval of word meaning from memory ("I need the *funguo*.") or in an informative context from which word meaning could be inferred ("I want to unlock the door. I need the *funguo*."). The informative context facilitated word comprehension during practice. However, later recall of word form and meaning, and word recognition in a new context were better after successful retrieval practice and retrieval practice with feedback than after context-inference practice. These findings suggest benefits of retrieval (so-called *testing effects*) during contextualized vocabulary learning; the uninformative context enhanced word retention by triggering memory retrieval.

6.1 INTRODUCTION

Learning vocabulary in a foreign language is a gradual process that often requires multiple repetitions (e.g., Webb, 2007b). The way in which a word is processed during these repetitions predicts whether the word is remembered over time (e.g., Hulstijn & Laufer, 2001; Nation, 2001). To support acquisition, words are often presented in context to allow learners to infer word meaning from contextual clues. In addition to using contextual clues, learners can also understand the meaning of words by retrieving knowledge gained during previous encounters with the word from memory (Nation, 2015; Schmitt, 2008). Both of these processes, context-inferences and memory retrieval, are potentially beneficial for the long-term retention of words (e.g., Folse, 2006; Hulstijn, 1992; Nation, 2001). However, it is unclear whether the degree to which a text stimulates context-inferences or retrieval influences the long-term retention of words. The present study was therefore conducted to examine the effect of these two processes more closely. Specifically, we investigated in three experiments if word retention is better when learners can infer word meaning from rich contextual clues, or when learners must engage in the retrieval of word meaning from memory because the context is uninformative.

6.1.1 WORD LEARNING THROUGH INFERENCES FROM CONTEXT

Successful context-inferences allow readers to establish the meaning of hitherto unknown words, which is necessary to create a form-meaning association (e.g., Li, 1988; Webb, 2008). Beyond this effect on *understanding*, inferences may also have effects on word *retention*. First, the processing of a word together with relevant contextual information could create semantic associations and enhance retention compared to processing of words without context (e.g., Schouten-van Parreren, 1989). Second, the inference process could enhance word learning because of deeper (i.e., more effortful, elaborate) processing of words during inferences compared to other ways to gain access to word meaning, such as consulting a glossary (e.g., Grace, 1998; Hulstijn, 1992). This is especially likely when the inferences are difficult (Haastrup, 1991; Hu & Nassaji, 2012).

A substantial number of studies has focused on learners' *understanding* of novel words in context, for example, describing the contextual clues and comprehension strategies that facilitate inferences (Beck, McKeown, & McCaslin, 1983; Fukkink & de Glopper, 1998; Kuhn & Stahl, 1998). In comparison, less is known about the effects of context-inferences on word *retention*. Some studies reported better word retention after words had been studied in context than without context (e.g., Baleghizadeh & Shahry, 2011), but others reported no effect of contextual information or even an advantage of learning words without context (e.g., Choi, Kim, & Ryu, 2014; Prince, 1996; Webb, 2007a). Similarly contradicting results were found in studies that

compared the retention of inferred and given word meaning. Some experiments showed better word retention in an inference-condition compared to a condition in which words were presented in the same context with the word meaning given (Carpenter, Sachs, Martin, Schmidt, & Looft, 2012; Hulstijn, 1992, Exp. 5). Others found no benefits of inferences (Hulstijn, 1992, Exp. 1 and 2; Mondria, 2003), even when participants spent more time processing each word when inferring the meaning than when the meaning was provided (Mondria, 2003). Taken together, previous research has shown that readers can use inference processes to understand unknown words in a text, but there is limited evidence that exposure to contextual information and the cognitive processes involved in inferring word meaning from context also have benefits for retention.

A possible explanation for limited benefits of context-inferences for retention is that learners might insufficiently process the word *form* during inferences (Lawson & Hogben, 1996; Pressley, Levin, & McDaniel, 1987). Although learning foreign vocabulary involves the acquisition of many different aspects of word knowledge, the encoding of the novel word form, its spelling and pronunciation, and the association of this word form with meaning are crucial (Deconinck, Boers, & Eyckmans, 2015). Inferences might not always strengthen form-meaning associations. Consider this sentence for illustration: "I want to unlock the door. I need the ___.". Here, the missing word "*key*" can be guessed even when no word form is present. Contextual information can thus enable readers to infer the meaning of a word without paying attention to its orthographic or phonological characteristics (Hu & Nassaji, 2012; Hulstijn, Hollander, & Greidanus, 1996). Such a focus on semantics can lead to reduced encoding of the word form and, consequently, weak form-meaning associations (Barcroft, 2002). Thus, although inferences might involve effortful processing of word meaning, the processing of the word form might be insufficient to retain the form-meaning association (Pressley et al., 1987).

6.1.2 WORD LEARNING THROUGH RETRIEVAL

Word learning often requires multiple repetitions and during the later repetitions, readers can access word meaning not only through inferences from context but also increasingly through the retrieval of word meaning from memory (Nation, 2015). Inferences and memory retrieval are to some extent competing processes because information that readers infer from the context is not searched for and retrieved from memory, and vice versa. This trade-off is relevant for word learning because memory retrieval influences the retention of information over time (Roediger & Karpicke, 2006b).

Compared to other forms of practice, such as mind-mapping or re-reading, repeated successful memory retrieval leads to better retention over time (e.g., Karpicke & Blunt, 2011; Roediger & Karpicke, 2006a). For example, Karpicke and

Roediger (2008) showed that when learners remembered the meaning of a new foreign word, practicing the retrieval of the word meaning from memory significantly enhanced performance on a translation test one week later, compared to a restudy condition in which words were repeatedly studied with translation but not retrieved from memory. Such positive effects of memory retrieval compared to other practice conditions are also called *testing effects.* Mechanistically, the core of these testing effects is that information – in this case the word meaning - is remembered better if it is retrieved from memory through an intentional mental search that involves the recall of knowledge, than if it is presented to the learner (see also Karpicke & Zaromb, 2010).

Testing effects have been documented in numerous studies in cognitive psychology (for reviews, see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Roediger & Butler, 2011; Roediger & Karpicke, 2006b; Rowland, 2014), including multiple studies that used foreign vocabulary or rare words from the first language (*L1)* as stimuli (see Goossens, Camp, Verkoeijen, & Tabbers, 2014 for an overview). Vocabulary researchers have also acknowledged that retrieval is beneficial for word learning (e.g., Folse, 2006; Nation, 2001). Barcroft showed, for example, that retrieval practice can be incorporated in vocabulary learning by using pictures (2007) or a cloze task during reading (2015) to trigger the retrieval of words from memory. In both experiments, retrieval enhanced word learning compared to a practice condition in which the translations were presented to the learners.

Prominent applications of memory retrieval to vocabulary learning are flash cards that allow learners to test themselves (e.g., Pimsleur, 1967, in Nation, 2001) and computer assisted language learning programs with repeated, spaced translation exercises (e.g., Lindsey, Shroyer, Pashler, & Mozer, 2014; Sense, Behrens, Meijer, & van Rijn, 2016). These exercises involve explicit recall activities similar to tasks employed in psychological research (e.g., Karpicke & Roediger, 2008). However, testing effects might also be evoked more incidentally when a learner is in a situation that requires the activation of word knowledge from memory (Barcroft, 2015). For example, a reader who encounters a newly learned word of which the meaning cannot be derived from its context might try to retrieve word knowledge from memory and thereby improve the word's retention over time. In other words, the amount of contextual clues about a word's meaning might influence to what extent readers engage in memory retrieval or context-inferences. Given the positive effects of retrieval on long-term retention, this raises the question if and how contextual richness influences word retention.

6.1.3 EFFECTS OF CONTEXTUAL RICHNESS ON WORD RETENTION

So far, only a limited number of studies have experimentally tested the effect of contextual richness on word retention in a foreign language (Mondria & Wit-de Boer, 1991; Webb, 2008)[1]. Mondria and Wit-de Boer (1991) conducted a study with Dutch high school students who guessed the meaning of French words from sentences and later reviewed the words in context with translations provided. When the initial practice sentences contained information about the function of the target words, students more often guessed the word meaning correctly than when the sentences did not contain this information. However, students were less likely to recall the words' meaning on a test after learning. Extra contextual information that made it easier to guess the word meaning during practice thus reduced later recall. The authors suggested that learners may have processed the target words less thoroughly in the richer context condition.

Unlike Mondria and Wit-de Boer (1991), Webb (2008) found positive effects of contextual information on word learning. He compared word learning between two groups of Japanese advanced learners of English who were presented with target words first in an informative sentence, and then in two more sentences that were either "more informative" or "less informative" regarding word meaning (Webb, 2008, p. 236). On an extensive test immediately after learning, the recall and recognition of word meaning was better for words practiced in the more informative condition but the recall and recognition of the word forms was similar for both conditions (Webb, 2008). From this, Webb concluded that the informative context may have increased the acquisition of word meaning, but not of the word form.

An important characteristic of both described studies is their focus on the initial presentations of words. Contextual information enables readers to understand the meaning of unknown words, which is important during readers' first encounters with a word. However, contextual information might have a different effect during later repetitions of words when readers have already acquired (partial) word knowledge that can be retrieved from memory instead of inferred from context. At this stage of learning, an uninformative context could become a beneficial trigger of retrieval. A result from Webb (2008) supports the idea that encounters of words in an uninformative context indeed become beneficial during later repetitions of words:

1 There is also a limited number of correlational studies that describe the relation between the informativeness of context and word retention from reading specific texts. Zahar, Cobb, and Spada (2001), for example, analyzed which target words most readers of a text did or did not learn from reading, and found no difference in the informativeness of the context that surrounded learned and unlearned words. These results must be interpreted cautiously, however, because contextual informativeness was not experimentally manipulated.

Comparing results from two studies, Webb found that presentations of foreign words in uninformative sentences had no measureable effect on retention when they occurred after one prior presentation, but they had significant benefits after seven prior presentations. Possibly, the uninformative sentences triggered the retrieval of (aspects of) word meaning from memory, which enhances word retention but can only succeed when learners have word knowledge that they can retrieve from memory (Kornell, Bjork, & Garcia, 2011; Chapter 2: van den Broek, Segers, Takashima, & Verhoeven, 2014). After a single exposure to a word in context, retrieval likely failed, but after seven exposures to a word, learners were likely to have gained some word knowledge, such that the uninformative sentences could trigger successful retrieval and thus produce a testing effect.

6.1.4 THE PRESENT STUDY

The central research question of this study was whether repetitions of words in context enhance retention more when the context stimulates learners to retrieve word meaning from memory than when it allows learners to infer word meaning from context. To the best of our knowledge, no previous study has triggered memory retrieval in this way. This study is thus the first attempt to evoke testing effects in vocabulary learning through a manipulation of the context in which words appear.

We conducted three separate experiments. Adult participants learned the meaning of selected words from a foreign language of which they had no prior knowledge, and then further practiced these words either in an uninformative L1-context that required memory retrieval (the *Retrieval condition*, for example, "Look at the *anga*!") or an informative L1-context that facilitated meaning inference (the *Context-inference condition*, for example, "There is not a single cloud today. Look at the *anga*!"). There is substantial evidence for beneficial effects of memory retrieval on the retention of information over time (Roediger & Butler, 2011; Roediger & Karpicke, 2006b; Rowland, 2014), but only limited evidence for benefits of context-inferences. Therefore, we predicted that retrieval would enhance word retention over time in comparison to context-inferences. Although context-inferences might involve beneficial effortful semantic processing (Hulstijn, 1992; Schouten-van Parreren, 1989), we assumed that they would direct readers' attention to comprehension rather than the form-meaning association (Pressley et al., 1987) and would therefore lead to weaker retention than retrieval.

6.2 EXPERIMENT 1

The overarching hypothesis for all three experiments was that practicing words in uninformative sentences that triggered memory retrieval would lead to better word retention than practicing words in informative sentences from which word meaning could be inferred. In Experiment 1, this hypothesis was tested by manipulating contextual richness after a pre-training during which learners gained (partial) word knowledge. Such prior exposure is necessary to obtain testing effects because retrieval practice is only beneficial if learners can indeed retrieve information from memory or receive feedback after failed retrieval attempts (e.g., Kornell et al., 2011; Rowland, 2014). Otherwise, learners are not re-exposed to the to-be-learned information and cannot benefit from retrieval practice.

The effect of retrieval and context-inference practice was tested in several ways in order to establish whether the predicted testing effects generalized across different measures of word learning. First, tests were administered both immediately and seven days after practice. Testing effects sometimes only become visible over time (Toppino & Cohen, 2009). Therefore, we tentatively predicted that benefits of the Retrieval condition might be more pronounced on the delayed test than on the immediate test. Second, participants translated words both into their native language and into the foreign language. This allowed us to test whether context-inference and retrieval practice affect both receptive and productive knowledge of the Swahili vocabulary (measured here as recall of the Dutch translation and the Swahili word form, respectively). Receptive vocabulary knowledge is typically acquired more easily than productive knowledge, likely because productive knowledge involves the formation of new lexical representations, whereas receptive knowledge "requires only discriminable, but not necessarily complete, representations of the new L2 words" (Schneider, Healy, & Bourne, 2002, p. 420). One reason to include both types of recall in the present study was that recall of word meaning and form might benefit from different retrieval tasks (Nakata, 2016b). Moreover, it was not clear if retrieval of the word meaning during practice would also benefit later recall of the word form (see also Carpenter, Pashler, & Vul, 2006). Previously, Webb (2008) suggested that contextual information might be particularly beneficial for the retention of word meaning but not of the word form. Mondria and Wit-de Boer (1991), however, found that contextual information reduced the recall of word meaning. We therefore did not formulate specific hypotheses about changes in receptive and productive word knowledge. Overall, we expected that retrieval practice would lead to better word retention than context-inference practice on all outcome measures. This testing effect would be driven by those words that participants translated successfully during practice, given the importance of retrieval success for testing effects (Kornell et al., 2011; Rowland, 2014).

6.2.1 METHODS

6.2.1.1 PARTICIPANTS. Forty-five undergraduate students (M_{age}= 23.8 years, SD_{age} = 8.8, 64.4% female) from a Dutch university took part in the experiment. All participants spoke Dutch fluently (88.9 % native speakers), and none of them had prior knowledge of Swahili. In all three experiments, participants received partial course credits or a monetary compensation (€10/hour).

6.2.1.2 STIMULI. The participants studied 104 Swahili nouns with Dutch translations, which were pronounceable for Dutch speakers, such as *anga* (sky), *bustani* (garden), *kichwa* (head), *samaki* (fish).

Retrieval and Context-Inference condition. During practice, target words were presented in sentences**.** In the Retrieval condition, these sentences contained only limited information and required memory retrieval to translate the target word (e.g., "We do not have any *mkate* left."). In the Context-inference condition, an additional sentence made it possible to infer the word meaning (e.g., "I'll go to the bakery. We do not have any *mkate* left." [bread]). The practice sentences were piloted to ensure that someone without prior knowledge of the target words could still derive word meaning from the context-inference sentences but not from the retrieval sentences (see Section 6.7.1 in the Appendix for further information).

6.2.1.3 DESIGN. The study has a 2x2 within-subject design with Practice Condition (Retrieval or Context-inference) and Testing Moment (Immediate, Delayed) as within-subject factors, and as dependent variables the proportion of words that were translated correctly on tests of receptive and productive word knowledge (see 6.2.1.4, *Recall tests*). The assignment of words to the two conditions (52 words per condition) and to the immediate or delayed test (respectively 25 and 27 words from each condition) was random, and so was the order of retrieval and context-inference trials during practice.

6.2.1.4 PROCEDURE. The experiment consisted of two sessions (see Figure 6.1): Session 1 comprised an initial encoding phase ("pre-training"), followed by retrieval and context-inference practice, and the immediate test. Session 2 seven days later comprised the delayed test. Session 1 took about 2.5 hours, Session 2 took about 1 hour.

Pre-training. The purpose of the pre-training was to ensure that participants learned the meaning of the majority of the Swahili words before the practice phase. Participants intentionally studied the Swahili words together with translations in four different encoding tasks, using the same procedure as reported in Chapter 3 (detailed description in Section 3.2.3.1 *Initial Encoding*). During the third task, each word was included in spaced repetitions until participants indicated that they knew the word already. On average, participants saw the words from both conditions 5.4 times over the course of the complete pre-training (M_{Retr}= 5.4 (*SD* = 2.1), M_{Ctx}= 5.4 (*SD* = 2.1)).

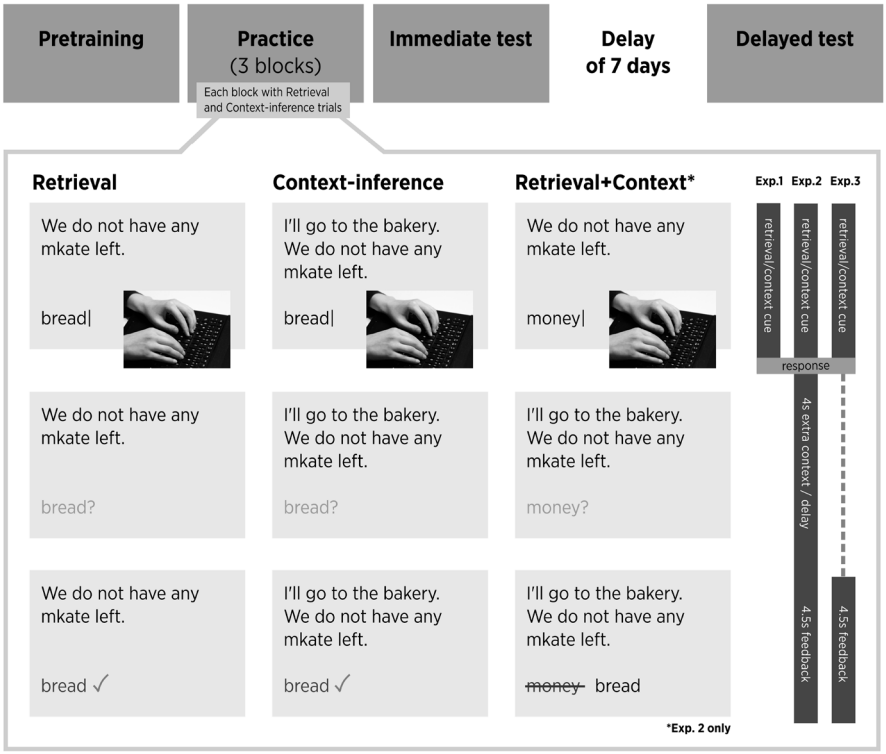


Figure 6.1 Overview of experimental procedure. In all three experiments, participants first underwent a pre-training in which Swahili words were studied together with their Dutch translations. Then participants completed the practice phase, in which a within-subject experimental manipulation was done: words were pseudorandomly assigned to the Retrieval condition, Context-inference condition, or Retrieval+Context condition (Exp.2 only). In the *Retrieval condition*, participants practiced with sentences that provided only limited information about the target word. In the *Context-inference condition*, more information was provided to allow learners to infer the meaning of the target word from context. In the *Retrieval+Context condition* in Experiment 2, participants first responded to a retrieval sentence, and were then presented with contextual information. The schema on the right indicates differences between practice trials in the three experiments: After participants responded, either the next trial began (Exp.1), or the response remained visible on the screen in grey font for 4 seconds, followed by feedback (Exp.2) or feedback was shown directly (Exp.3). The figure illustrates feedback for two correct responses (tick mark after correct word) and one incorrect responses (strikethrough response, display of correct word).

Practice with retrieval and context-inference sentences. The practice phase followed immediately after the pre-training. It consisted of three blocks. In each block, 52 words were presented in the Retrieval condition and 52 words were presented in the Context-inference condition. Sentences were presented one by one, and participants typed in the translation of the included Swahili word (see Figure 6.1). The sentences remained visible until the participants submitted a response. No feedback was provided; a fixation cross was shown for 1.5 seconds before the next sentence was presented. The same 104 sentences were presented in all three practice rounds; the order of the presentations was randomized.

Recall tests immediately and seven days after practice. A translation test was administered for 25 words from each condition directly after practice in Session 1, and for the other 27 words seven days after practice, in Session 2. Swahili words were presented one by one, and participants were asked to type in the Dutch translation (the test of *receptive word knowledge*). After a short distractor task (an iconic memory task that took about 1 minute), participants were then asked to translate the same words from Dutch to Swahili (the test of *productive word knowledge*). In Session 2, the translation test was preceded by a picture naming test. Participants were shown three complex pictures and were instructed to type in any Swahili word that described an element of the picture. Performance on this test was at floor level; the data are therefore not reported in this article. After the picture test, participants did distractor tasks for three minutes before completing the translation tests. The order in which items were tested was always randomized.

Scoring. Responses on the recall tests were categorized as either correct or incorrect, with spelling errors being counted as correct for the test of receptive knowledge (e.g., the response "*fahter*" instead of "*father*" was counted as correct). Responses during the test of productive knowledge were counted as correct when they had an edit distance of 2 or lower from the correct answer, which means that no more than two letters had to be added or removed to get to the perfect answer (e.g., "*keja*" or "*keah*", instead of "*keha*" were counted as correct).

6.2.1.5 STATISTICAL ANALYSES. We used repeated measures analyses of variance (ANOVA) in SPSS (vers. 22.0.0.1) to describe the effect of Practice Condition (Retrieval, Context-inference) and Testing Moment (Immediate, Delayed) on the proportion of words that were translated correctly, as aggregated per participant. For all analyses reported in this chapter, data sufficiently met assumptions of normality, hetereoskedasticity and sphericity. Confidence intervals of the difference scores of pairwise contrasts have been included in Figure 6.3; partial eta squared (η_p^2) is reported for omnibus tests and Cohen's *d* for pairwise comparisons (using Formula (3) in Dunlap, Cortina, Vaslow, & Burke, 1996, p.171). Because it is increasingly recommended to use mixed models in psycholinguistics (Baayen, Davidson, & Bates,

2008), all analyses reported in this paper were replicated with mixed logit models with crossed random effects for items and participants, using the glmer function in the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) in R (version 3.1.2, R Core Team, 2015). All effects found with ANOVAs on the aggregated data were also significant in the mixed models.

6.2.2 RESULTS

6.2.2.1 WORD KNOWLEDGE AFTER RETRIEVAL AND CONTEXT-INFERENCE PRACTICE. Descriptive statistics of the proportion of word forms and meanings that were translated correctly on the final memory tests immediately and seven days after learning are reported in Table 6.1. We conducted two repeated measures analyses of variance (ANOVA), with Practice Condition (Context-Inference, Retrieval) and Testing Moment (Immediate, Delayed) as within-subject-factors, and as dependent variables the proportion of words that were translated correctly from Swahili to Dutch (short: *receptive word knowledge*), and the proportion of words that were translated correctly from Dutch to Swahili (short: *productive word knowledge*).

Receptive word knowledge. There was a significant main effect of Testing Moment on receptive word knowledge, which reflected a decline in recall over time, $F(1,44) = 181.00$, $p < .001$, $d = 0.82$, but no main effect of Practice Condition, $F(1,44) = 0.01$, $p = .95$, $d = 0.01$, nor an interaction between Practice Condition and Testing Moment, $F(1,44) = 1.62$, $p = .21$, $\eta_p^2 = 0.4$.

Productive word knowledge. The pattern of results was the same for the test of productive knowledge as for the test of receptive knowledge, with a significant effect of Testing Moment due to a decline in recall over time, $F(1,44) = 109.14$, $p < .001$, $d = 0.67$, but no main effect of Practice Condition, $F(1,44) = 1.94$, $p = .17$, $d = 0.09$, nor an interaction between Practice Condition and Testing Moment, $F(1,44) = 0.27$, $p = .62$, $\eta_p^2 = .01$.

6.2.2.2 WORD KNOWLEDGE AFTER *SUCCESSFUL* RETRIEVAL AND CONTEXT-INFERENCE PRACTICE. We measured participants' performance during practice because retrieval success is a requirement for testing effects. Although participants successfully translated the majority of words during practice, they did not type in the correct translation of 18.2% of the words in the Retrieval condition, and of 2.4% of the words in the Context-inference condition. Figure 6.2 summarizes the number of correct practice responses in the two conditions, and the relation of the number of correct practice responses with receptive word knowledge after practice. There are two important patterns to note in Figure 6.2. First, as indicated by differences in the surface area of the grey and the white circles, more words were successfully translated during practice in the Context-inference condition than in the Retrieval condition. A Chi-square test of independence showed that this association between the practice

condition and the number of correct responses during practice was significant, $\chi^2(3) = 834.62$, $p < .001$. Second, for words that had been translated successfully during at least one of the three practice rounds, receptive word knowledge was higher for the Retrieval condition than for the Context-inference condition, as indicated by the higher position of the white circles than the grey circles.

Separate ANOVAs with participants' performance aggregated for only the words that had been translated correctly at least once during practice, revealed that the practice condition had a large significant effect on later recall, $F(1,44) = 30.8$, $p < .001$, $d = 0.51$, with higher accuracy on the test of receptive word knowledge after successful retrieval practice (estimated mean (M_{est} = 0.70) than after successful context-inference practice (M_{est} = 0.61, confidence intervals in Figure 6.3). Accuracy on the test of productive knowledge was also better after successful retrieval practice (M_{est} = 0.61) than after successful context-inference practice (M_{est} = 0.51), $F(1,44) = 27.27$, $p < .001$, $d = 0.46$.

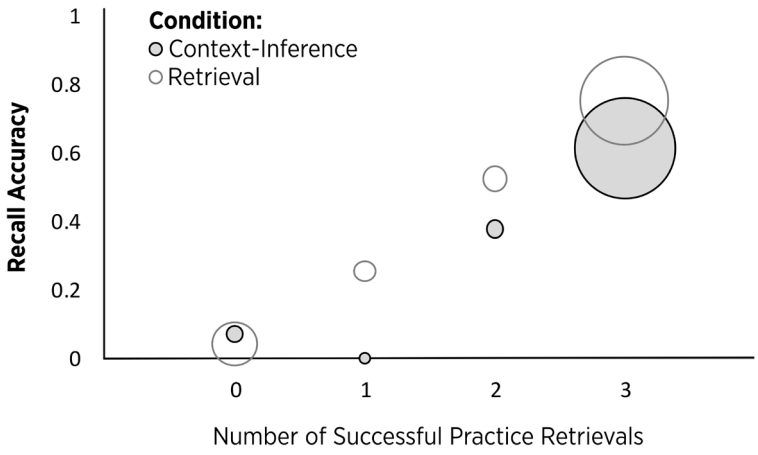


Figure 6.2 Accuracy of recall of the word meaning (receptive vocabulary knowledge) in Experiment 1 as a function of number of successful practice responses (possible values: 0, 1, 2, or 3) and Practice Condition (Context-inference or Retrieval). Data from the immediate and the delayed test are combined in this figure. The surface of the circles represents the number of items in each category; the center of the circles represents mean recall accuracy averaged across item-level observations. See Section 6.7.2 for the exact values depicted in this figure.

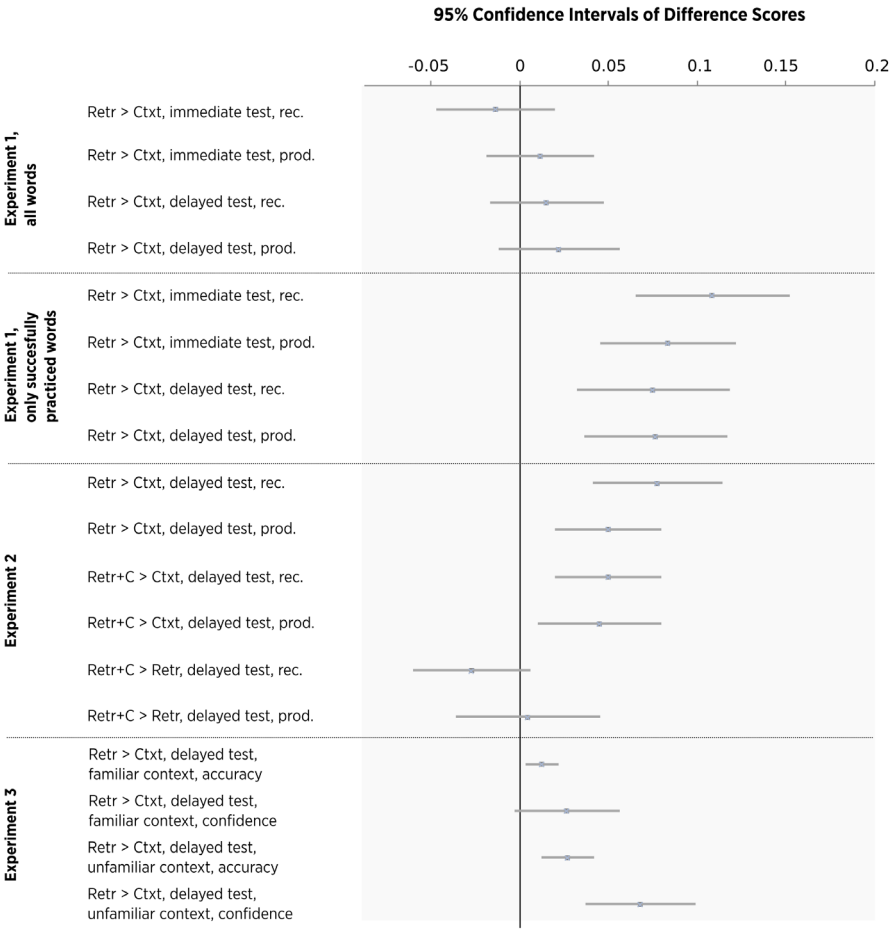


Figure 6.3 95% Confidence intervals (*CI*) of the difference scores between final test performance after practice in the Retrieval (*Retr*), Context-inference (*Ctxt*), and Retrieval+Context (*Retr+C*) conditions (direction of comparisons indicated with ">"). CI are shown for tests of receptive (*rec.*) and productive (*prod.*) vocabulary knowledge, and for accuracy and confidence on the sentence judgment test in Experiment 3. Accuracy (rec., prod., accuracy) was measured as proportion correct; confidence was measured on a rating scale. CI that do not overlap with 0 indicate significant differences with $p < .05$. All CI that overlap with 0 were not significant in the ANOVA analyses.

Table 6.1 Proportion of Correct Responses during Practice and on the Final Tests for the three Experiments

		Practice (Correctly translated during practice block)		
	Practice Condition	*Block 1*	*Block 2*	*Block 3*
Exp. 1	Retrieval	0.76 (0.20)	0.77 (0.21)	0.79 (0.20)
N = 45	Context-Inference	0.96 (0.07)	0.96 (0.07)	0.96 (0.06)
	Successful Retrieval[a]	0.92 (0.07)	0.93 (0.10)	0.95 (0.07)
	Successful Context-Inference[a]	0.98 (0.03)	0.98 (0.04)	0.99 (0.03)
Exp. 2	Retrieval	0.68 (0.18)	0.89 (0.11)	0.96 (0.06)
N = 44	Context-Inference	0.96 (0.08)	0.995 (0.02)	0.999 (0.01)
	Retrieval + Context-inference	0.68 (0.18)	0.88 (0.13)	0.95 (0.08)
Exp. 3	Retrieval	0.68 (0.18)	0.88 (0.12)	0.96 (0.08)
N =41	Context-Inference	0.96 (0.05)	0.996 (0.01)	0.999 (0.004)

Note: The table lists the proportion of correct answers (averaged across participants, standard deviations in brackets) during the three practice blocks and final test performance on tests immediately or seven days after learning (*delayed*). Proportions of correct responses on tests of receptive (*recept.*) and productive (*product.*) vocabulary knowledge are listed for Experiment 1 and 2, and accuracy (*Accur*) and confidence (*Conf.*) of ratings of words in context are listed for Experiment 3. [a]For the analysis of successful retrieval and successful context-inference practice in Experiment 1, all items were included that were correctly translated in at least one of the three practice blocks. [b]The data from the immediate test in Exp. 2 were not included in statistical analyses because the number of observations was too low (4 test items per condition), but are included here for archival purposes.

6.2.3 DISCUSSION

In Experiment 1, participants repeatedly translated Swahili words presented either in an informative context from which word meaning could be inferred or in an uninformative context that required memory retrieval. Tests immediately and seven days after practice showed no difference between the two conditions in recall accuracy. However, whereas participants during practice almost always filled in the correct translation in the Context-inference condition, participants failed to provide the correct translation in the Retrieval condition for about 18% of the words. When this difference in correct practice responses was controlled for, a benefit of retrieval practice became visible: Recall was significantly higher for words that had

	Immediate recall test		Delayed recall test		Recognition in familiar context		Recognition in unfamiliar context	
	recept. [0 – 1]	*product.* [0 – 1]	*recept.* [0 – 1]	*product.* [0 – 1]	*Accur.* [0 – 1]	*Conf.* [1 – 3]	*Accur.* [0 – 1]	*Conf.* [1 – 3]
	0.78 (0.22)	0.65 (0.21)	0.43 (0.22)	0.42 (0.21)				
	0.77 (0.21)	0.64 (0.22)	0.41 (0.22)	0.40 (0.19)				
	0.90 (0.13)	0.73 (0.16)	0.50 (0.20)	0.48 (0.19)				
	0.79 (0.20)	0.65 (0.21)	0.42 (0.22)	0.40 (0.20)				
	0.87[b] (0.19)	0.74 (0.25)	0.43 (0.20)	0.42 (0.21)				
	0.74 (0.25)	0.70 (0.28)	0.35 (0.20)	0.37 (0.21)				
	0.84 (0.22)	0.70 (0.30)	0.40 (0.20)	0.42 (0.21)				
					0.89 (0.08)	2.65 (0.24)	0.85 (0.09)	2.55 (0.32)
					0.88 (0.07)	2.63 (0.26)	0.83 (0.08)	2.51 (0.31)

been successfully practiced in the Retrieval condition than for words that had been successfully practiced in the Context-inference condition. This result was found on all sub-scores: on the immediate and the delayed test, for productive and receptive word knowledge. Benefits of retrieval practice were not more pronounced on the delayed test than on the immediate test.

One interpretation of the results of Experiment 1 is that a testing effect exists when words are practiced in context but only if the retrieval is successful (e.g., Halamish & Bjork, 2011; Kornell et al., 2011). It is plausible that participants only benefited from retrieval if they translated the target words successfully because otherwise they did not have access to the correct word meaning. However, this restriction of the

analysis to successfully retrieved items may have introduced an unwanted bias (see Karpicke, Lehman, & Aue, 2014). Specifically, the 81.8% of words that were translated correctly during retrieval practice may have been inherently easier for participants than the 97.6% of words that were translated correctly during context-inference practice. After all, during retrieval practice, participants could only translate those words from memory that they remembered from the pre-training whereas during context-inference practice they could translate practically all words by inferring their meaning. Experiment 2 was therefore conducted to replicate the comparison between retrieval and context-inference practice while controlling for item-selection effects, by ensuring that learners always had access to the correct word meaning in both conditions via the use of feedback.

An alternative interpretation of the results of Experiment 1 is that the extensive pre-training may have reduced the effect of practice. There was no baseline measurement, but translation accuracy in the first round of retrieval practice can be considered a rough estimate of the proportion of word translations which participants could recall after the pre-training, *before* practice (76%, see Table 6.1). Comparing this estimate to performance on the immediate test *after* practice (77% in the Context-inference condition, 78% in the Retrieval condition), showed no significant improvement of performance after practice in either condition (both $d < 0.10$). This result was surprising for the Context-inference condition, in which participants correctly translated almost all words repeatedly during practice and therefore could have learned additional words. The similar results before and after practice suggest that the repeated successful context-inferences had no or only minimal benefits for later recall. To rule out that this result was due to the fact that participants' performance reached a plateau through the extensive pre-training, and to make the effect of practice more measurable, the pre-training was shortened in Experiment 2.

6.3 EXPERIMENT 2

Experiment 2 was conducted to rule out that findings in Experiment 1 were driven by an item selection bias and to test again whether the Retrieval condition enhanced word retention compared to the Context-inference condition. For this purpose, we added feedback to the practice phase in Experiment 2. Feedback is beneficial for contextual word learning, especially when the contextual support is weak (Frishkoff, Collins-Thompson, Hodges, & Crossley, 2016). In particular, feedback allows learners to encode information that they cannot retrieve from memory and therefore reduces the impact of retrieval failures (e.g., Rowland & DeLosh, 2015). Adding feedback in

Experiment 2 made it unnecessary to include retrieval success in the analyses and thereby removed the potential bias through item selection.

Two other major changes were made in the paradigm in comparison to Experiment 1. First, to make the effect of practice more measurable, the pre-training was shortened in Experiment 2. Second, a third practice condition was added. In this *Retrieval+Context condition*, participants first responded to an uninformative (retrieval) sentence, and were then exposed to the contextual information from the Context-inference condition so that they could evaluate their response. In addition to the relative benefits of a context-inference condition and a retrieval condition, this allowed us to study possible additive benefits of the two conditions since both memory retrieval and context-inferences are supposedly beneficial for word learning. We expected the combined condition to further enhance performance compared to practice with only retrieval or only context-inferences.

6.3.1 METHODS

The experiment had a similar structure as Experiment 1 (see Figure 6.1), but the pre-training was shortened, a Retrieval+Context condition was added, and feedback was added to the practice trials. These and a few other, more minor changes are described in the following sections.

6.3.1.1 PARTICIPANTS. Forty-four undergraduate students (M_{age}= 18.6 years, SD_{age} = 0.8, 84% female) took part in the experiment. All participants spoke Dutch fluently (93% native speakers), and none of them had participated in Experiment 1 or reported prior knowledge of Swahili.

6.3.1.2 STIMULI. We used 102 of the 104 Swahili nouns from Experiment 1, which were distributed across the Retrieval condition, the Context-inference condition, and the Retrieval+Context condition per participant. The retrieval sentences and the context-inference sentences were identical with Experiment 1; in the newly added Retrieval+Context condition, participants first saw the retrieval sentence (e.g., "Where is the funguo?"), made a response, and were then presented with the full contextual information from the Context-inference condition (e.g., "Where is the funguo? I would like to unlock the door." See Figure 6.1).

6.3.1.3 DESIGN. We used a within-subject design with Practice Condition (Retrieval, Context-inference, Retrieval+Context) as within-subject factor and as dependent variables accuracy on delayed tests of receptive and productive word knowledge. The assignment of words to the three conditions was pseudorandom (based on the pre-training, see next section), and the order of Retrieval, Context-inference, and Retrieval+Context trials during practice was random.

6.3.1.4 OVERVIEW OF THE EXPERIMENT.

Pre-training. The pre-training was similar to Experiment 1 but the third task was shortened. During the third task, participants now saw each word only once and rated how well they already knew the word on a continuous scale from 1 (not at all) to 5 (perfectly). These ratings were used to assign words of similar difficulty to the three practice conditions for every participant, by first ranking words by rating and then randomly distributing groups of three words across the three conditions. As a result, the average ratings in the three conditions were highly similar, (M_{Retr}= 2.23 (*SD* = 1.04), M_{Ctx}= 2.24 (*SD* = 1.04), $M_{Retr+Ctx}$= 2.24 (*SD* = 1.03)).

Practice with retrieval, context, and retrieval+context sentences. As in Experiment 1, participants typed in the translation of each Swahili word upon seeing the word in a sentence. The sentences remained visible until participants submitted a response. Next, the informative context sentence was added to the display in the Retrieval +Context-inference condition and remained visible for a fixed duration of 4 seconds. To ensure that trials in the three conditions had the same length, the sentence(s) and the submitted response also remained visible on the screen for four seconds in the other two conditions, before feedback (the correct translation) was shown for 4.5 seconds in all three conditions (see Figure 6.1).

Recall tests immediately and seven days after practice. Due to the addition of a third condition, it was not possible to test sufficient items on both an immediate test and a delayed test. We therefore focused on recall performance on the delayed test in this experiment. Four words from each condition were tested on the immediate test directly after practice in Session 1 to give participants some experience with the test situation, comparable to Experiment 1; the other 30 words per condition were presented on the delayed test seven days after practice. The order of the translation tasks and distracter activities was the same as in Experiment 1 but no picture description task was done.

6.3.1.5 STATISTICAL ANALYSIS. We used only the data from the delayed test in two repeated measures ANOVA's with Practice Condition (Retrieval, Context, Retrieval+Context) as within-subject factor and accuracy on tests of receptive and productive word knowledge as dependent variables. Data from the immediate test were excluded due to the low number of test trials on that test but exploratory analyses with data from both testing moments showed the same main effect of Practice Condition on the immediate test as found on the delayed test, and no interaction of Practice Condition and Testing Moment. All analyses were replicated with a mixed model approach, as described in the method section of Experiment 1.

6.3.2 RESULTS

Descriptive statistics can be found in Table 6.1.

6.3.2.1 RECEPTIVE WORD KNOWLEDGE. There was a significant main effect of Practice Condition on receptive word knowledge, $F(2,86) = 11.15$, $p = <.001$, $\eta_p^2 = .21$. Pairwise comparisons of the three practice conditions showed that performance was lower in the Context-inference condition ($M_{est} = 0.35$, $SE = 0.03$) compared to the Retrieval condition ($M_{est} = 0.42$, $SE = 0.03$, $p < .001$, $d = 0.38$) and compared to the Retrieval+Context condition ($M_{est} = 0.40$, $SE = 0.03$, $p = .002$, $d = 0.25$), see also Figure 6.3. Numerically, performance was higher in the Retrieval condition than in the Retrieval+Context condition, but this difference did not reach significance in the ANOVA analysis, $p = .10$, $d = 0.13$. However, as in all analyses reported in this paper, results were replicated with a mixed logit model. In this analysis, the difference between the Retrieval and the Retrieval+Context condition was significant, $p < .05$, indicating higher performance in the Retrieval condition than in the Retrieval+Context condition.[2]

6.3.2.2 PRODUCTIVE WORD KNOWLEDGE. There was a significant main effect of Practice Condition on productive word knowledge, $F(2, 86) = 5.28$, $p = .007$, $\eta_p^2 = .11$. Pairwise comparisons revealed that, as for the receptive test, performance was lower in the Context condition ($M_{est} = 0.37$, $SE = 0.03$) than in the Retrieval condition ($M_{est} = 0.42$, $SE = 0.03$, $p = .001$, $d = 0.24$) and in the Retrieval+Context condition ($M_{est} = 0.42$, $SE = 0.03$, $p = .008$, $d = 0.22$). Performance did not differ significantly between the two retrieval conditions, $p = .824$, $d = 0.02$.

2 Mixed logit models were fitted using the glmer function in the lme4 package (Bates, Mächler, Bolker, & Walker, 2015; Bates, Maechler, Bolker, & Walker, 2014) in R (version 3.1.2, R Core Team, 2014), with accuracy on the delayed receptive recall test as binary outcome variable (correct or incorrect). The model with the best fit as determined with the maximum likelihood criterion was a model with random intercepts for participants and words, and a fixed effect of Practice Condition. The dummy-coded regression coefficients showed that the odds that receptive recall was successful were significantly higher in the Retrieval condition than in the Context-inference condition (Odd's Ratio $OR = 1.58$), and were higher in the Retrieval+Context condition than in the Context-inference condition ($OR = 1.26$). These contrasts replicate the results from the RM ANOVA. However, the mixed model also showed a significant difference between the two retrieval conditions: the odds for correct recall were significantly higher in the Retrieval condition than in the Retrieval+Context condition ($OR = 1.25$). Confidence intervals obtained with bootstrapping and p-values obtained with Satterthwaite's approximation in lmerTest indicated that this effect was significant with $.01 < p < .05$. Note that this was the only contrast in the mixed model that led to a different conclusion than the ANOVA on the aggregated data.

6.3.3 DISCUSSION

In Experiment 2, both retrieval conditions led to significantly higher productive and receptive word knowledge seven days after practice than the Context-inference condition. With feedback, uninformative sentences that required memory retrieval thus led to better retention than informative sentences from which word meaning could be inferred. These results provide further evidence that reducing the amount of contextual information during practice can enhance foreign word retention by triggering retrieval. In Experiment 1, this testing effect was only found for those items that were successfully translated. In Experiment 2, after addition of feedback, a testing effect was found for all items. This result strengthens the tentative conclusion from Experiment 1 that retrieval practice leads to better word retention than context-inference practice if learners have access to the meaning of words. By adding corrective feedback in Experiment 2, access to word meaning was guaranteed and retrieval practice led to better performance than context-inferences. This effect was found on tests of both productive and receptive word knowledge, as in Experiment 1.

An unexpected finding in Experiment 2 was the trend towards a negative effect of providing contextual information after retrieval. The difference between the Retrieval and the Retrieval+Context condition was small and reached statistical significance only in the mixed model analysis but not in the analysis of the aggregated data (where p = .10). Nevertheless, this result is remarkable because it directly contradicted our prediction that participants in the Retrieval+Context condition would benefit from the best of both other conditions if they first tried to translate words in the uninformative retrieval sentences from memory, and then processed additional contextual information to infer word meaning and evaluate their answer. In the Retrieval condition, participants could not infer word meaning from context and instead waited for four seconds after responding. Still, the Retrieval condition led to better learning outcomes than the Retrieval+Context condition. This finding provides further evidence that extra contextual information does not always enhance word learning and can sometimes even have negative effects. Possibly, participants paid more attention to the target words in the period after the response submission in the Retrieval condition than in the Retrieval+Context condition, in which they may have instead focused at the sentence context. The contextual information could also have changed how participants approached the retrieval task in the second and third practice round, if participants translated the words by recognizing or recalling the context rather than by activating the form-meaning association. In both cases the focus on comprehension might have reduced encoding of the word form (Barcroft, 2002; Deconinck et al., 2015; Hu & Nassaji, 2012). Irrespective of the specific mechanism underlying the effect, additional contextual information had a negative effect on word retention in this experiment as it resulted both in lower performance

in the Context-inference condition compared to the Retrieval condition, and in lower performance in the Retrieval+Context condition compared to the Retrieval condition.

The unexpected results from the combined Retrieval+Context condition raised the question whether inferring a word's meaning from context has benefits at all for word retention over time, once a learner can understand the word. Comparing again translation accuracy in the first round of retrieval practice to recall on the immediate test, we found significantly better performance after practice than before practice for all three conditions. This difference was large and significant for both retrieval conditions (d_{Retr} = 1.03, $d_{Retr+Ctx}$ = 0.79), but only small for the Context-inference condition, $t(43) = 1.90$, $p = .03$ (one-sided), $d = 0.26$. These results suggest that once learners understand a word, repetitions in an uninformative context that triggers retrieval are more beneficial for retention than repetitions in a rich context from which word meaning can easily be guessed. Repetitions in such a rich context had only a weak effect on word retention in Experiment 2.

6.4 EXPERIMENT 3

Experiment 1 and 2 showed better recall of word form and meaning after learners had practiced words repeatedly in an uninformative context that required memory retrieval than after learners had practiced words in an informative context from which word meaning could be inferred. We attribute this result to the beneficial effects of memory retrieval on retention (e.g., Roediger & Butler, 2011). However, an alternative explanation is that retrieval practice and the final translation tests both required that learners activated word form-meaning associations from memory, whereas the Context-inference condition may have focused attention more on word and context meaning. This difference in overlap between processing during practice and test may have biased results in favor of the Retrieval condition due to *transfer-appropriate processing*, that is, the phenomenon that practice tends to have larger benefits when it involves similar cognitive processes as the final performance test (Morris, Bransford, & Franks, 1977; Veltre, Cho, & Neely, 2015; Winstanley, 1996). Indeed, L1 studies suggest that benefits of practicing words in context can become visible when tests are sensitive to semantic associations or require the use of words in context even when recall tests show no such benefits (see Frishkoff, Perfetti, & Collins-Thompson, 2011).

Experiment 3 was therefore conducted to see if the benefits of retrieval practice compared to context-inferences could be replicated with a final test that was more sensitive to semantic associations and more similar to context-inference practice. We constructed a test on which participants had to judge whether the practiced words were appropriately used in different sentences. Some test items presented

the target words in a sentence that included words and semantic concepts from the Context-inference condition; other test items presented the target words in a new, unrelated context. Based on the idea that the overlap between practice and final test enhances performance, we expected to find smaller or no benefits of retrieval practice over context-inference practice on this test compared to the previous experiments. Furthermore, differences in performance on familiar and unfamiliar test items were investigated to see to what extent benefits of context-inference practice were restricted to the specific context from practice or also transferred to a new context. The answer scale measured both accuracy and confidence of responses to have an objective and a subjective measure of word learning.

6.4.1 METHODS

Experiment 3 included only the Retrieval condition and the Context-inference condition. The pre-training was identical to that in Experiment 2; practice trials were similar to Experiment 2 but shortened; the test format was changed. See Figure 6.1 for an overview of the differences between experiments.

6.4.1.1 PARTICIPANTS. Forty-one university students (40 female, M_{age} = 20.0, SD_{age} = 2.3) took part in Experiment 3. Again, all participants spoke Dutch fluently (92.7 % native speakers), and none of them had prior knowledge of Swahili nor participated in Experiment 1 or 2.

6.4.1.2 STIMULI. We used 100 of the 102 words from Experiment 2.

Retrieval and context-inference practice. Immediately after participants submitted a response, the same feedback (correct translation) as in Experiment 2 was displayed for 4.5 seconds.

6.4.1.3 SENTENCE JUDGMENT TEST. For each Swahili word, four test sentences were constructed (see Table 6.2 for examples). These were two sentences in which the Swahili words fit into the context (*Fit*), and two sentences in which the Swahili words did not fit (*NoFit*). For each word, one of the two Fit test sentences and one of the two NoFit test sentences were semantically related to the practice sentences from the Context-inference condition (short: *familiar*). The other Fit and NoFit test sentence were different from practice (short: *unfamiliar*). The familiar Fit sentences were constructed using words or concepts from the context-inference practice sentences. For the familiar NoFit sentences, Swahili words were inserted into the familiar Fit sentence of a different Swahili word, thereby creating a test sentence in which the word did not fit. The unfamiliar sentences were constructed using words and topics that did not occur in the practice context. The presentation order of the test items was random, but it was ensured that for each participant, for half of the words from each condition, the familiar Fit sentences were presented first, and for the other half of the words, the unfamiliar Fit sentences were presented first.

Accuracy and confidence measures. Participants rated each test item on a 6-point scale that indicated whether they thought that the word fit into the context (left half of the scale) or not (right half of the scale) and how confident they were in their answer (from "*guess*" (1) to "*I am sure*" (3), see the full scale in Table 6.2). Accuracy (of answering on the left or right half of the scale given the fit of the word into the sentence), confidence, and confidence for accurate responses only, were then aggregated per participant.

6.4.1.4 STATISTICAL ANALYSES. Repeated measures ANOVAs were conducted with Practice Condition (Retrieval, Context-inference) and Familiarity of test context (Familiar, Unfamiliar) as within-subject factors, and as dependent variables participant means for accuracy and confidence ratings. We also conducted separate mixed model analyses of accuracy and confidence measures, which replicated the reported significant effects.

6.4.2 RESULTS

Descriptive statistics of accuracy and confidence of the test responses can be found in Table 6.1.

6.4.2.1 ACCURACY. The accuracy of judgments of words in context was higher in the Retrieval condition than in the Context-inference condition, $F(1, 40) = 19.32$, $p < .001$, $d = 0.24$. Participants also showed higher accuracy on the familiar test items than on the unfamiliar test items, $F(1,40) = 89.93$, $p < .001$, $d = 0.50$. There was no interaction between Practice Condition and Familiarity, $F(1,40) = 2.56$, $p = .12$, $\eta_p^2 = .06$.

6.4.2.2 CONFIDENCE. Similar to the results on accuracy, confidence was higher in the Retrieval condition than in the Context-inference condition, $F(1, 40) = 12.94$, $p < .001$, $d = 0.14$, and higher for the familiar than the unfamiliar test items, $F(1,40) = 206.14$, $p < .001$, $d = 0.43$. Additionally, there was a significant interaction between Practice Condition and Familiarity of Test Context, $F(1, 40) = 8.31$, $p = .006$, $\eta_p^2 = .17$, reflecting significantly higher confidence when participants rated words from the Retrieval condition than when they rated words from the Context-inference condition for the unfamiliar test items, $p < .001$, $d = 0.19$, but not for the familiar test items, $p = .075$, $d = 0.09$[3]. To ensure that these results were not driven by differences in accuracy, confidence ratings were also aggregated for accurate responses only. This analysis led to the same pattern of results as the analysis of all responses.

3 In contrast, the mixed model analysis indicated that this effect, too, was significant with $.01 < p < .05$. The odds that confidence was high were significantly higher in the Retrieval condition than in the Context-inference condition for familiar (*OR* = 1.32) as well as for unfamiliar test items (*OR* = 1.43).

Table 6.2 Example Items from the Sentence Judgment Test in Experiment 3

	Test sentence type			
	Familiar Fit	**Familiar NoFit**	**Unfamiliar Fit**	**Unfamiliar NoFit**
Example Item 1 *mkate* = bread	I'll quickly go to the bakery to get some *mkate*.	During an asthma attack, *mkate* cannot enter the lungs freely.	Fresh *mkate* tastes best.	He walks with crutches because he hurt his *mkate*.
Example Item 2 *hewa* = air	During an asthma attack, *hewa* cannot enter the lungs freely.	She knits a scarf of fine *hewa*.	Many factories in this area pollute the *hewa*.	The *hewa* was sharing my friend's bike.
Answer scale	(1) *I am sure that the word does not fit in this context* (2) *I think the word does not fit in this context* (3) *I don't know but my guess is that the word does not fit in this context* (4) *I don't know but my guess is that the word fits in this context* (5) *I think the word fits in this context* (6) *I am sure that the word fits in this context.*			
Correct Response	The word fits in this context.	The word does not fit in this context	The word fits in this context.	The word does not fit in this context

Note. During the test, all words were presented in each of the four types of test sentences (familiar Fit, unfamiliar Fit, familiar NoFit, and unfamiliar NoFit test items). The familiar Fit sentences were created based on the practice sentences from the Context-inference condition (see also Figure 1). For the familiar NoFit items, a Swahili word was inserted into a sentence that pertained to a different Swahili word such that the word did not fit into the context. The familiar NoFit sentence for Item 1 is an example of this; it was constructed based on the familiar Fit test sentence for Item 2 (marked with an arrow).

6.4.3 DISCUSSION

Learners more accurately and more confidently recognized words in context seven days after retrieval practice than after context-inference practice. Experiment 3 thus replicated the benefits of retrieval practice found in Experiment 1 and 2, now with a test that involved the presentation of words in a sentence context. Judgments on this test were more accurate after prior retrieval practice than after prior context-inference practice, both when test items were familiar because they resembled the sentences used during context-inference practice and when test items were unfamiliar. In addition, participants were more confident when they judged words from retrieval practice than when they judged words from context-inference practice. This effect was more pronounced when words were presented in an unfamiliar context than when words were presented in a familiar context (where $p = .075$ in the ANOVA, but $p < .05$ in the mixed model).

As we had predicted, familiar sentences were rated more accurately and more confidently than unfamiliar sentences, possibly due to transfer-appropriate processing (Veltre et al., 2015; Winstanley, 1996). It is important to note that the familiar test items were constructed based on the sentences from context-inference practice, and were thus actually only familiar to participants for the words practiced in that condition. This greater overlap between context-inference practice and the familiar test context than between retrieval practice and the familiar test context might explain why the data showed stronger evidence for benefits of the Retrieval condition for the unfamiliar test items than for the familiar test items. For the latter, benefits of retrieval may have been counteracted by the greater overlap between test items and context-inference practice. Independent of familiarity, accuracy was better for the words practiced in the Retrieval condition than in the Context-inference condition on both familiar and unfamiliar test items. Overall, this suggests that although transfer-appropriate processing might influence how well participants recognize words in context, benefits of retrieval over context-inference practice are robust and exist even on a test that presents words in context.

6.5 GENERAL DISCUSSION

Given that language learners often practice words in context, it is important to understand the effect of textual characteristics on word retention. This study focused on contextual richness as a trigger of context-inferences and memory retrieval during intentional vocabulary practice. In three different experiments, words were remembered better after practice with an uninformative context that required memory retrieval to access word meaning than after practice with an informative context from which word meaning could be inferred. In Experiment 1, this testing effect was found only for words that participants had translated successfully during practice: Performance was higher after successful retrieval practice than after successful context-inference practice both immediately after learning as well as after seven days, and on tests of both productive and receptive word knowledge. In Experiment 2, feedback was added to the practice phase and a testing effect was again found, now for all items. This confirmed that the testing effect in Experiment 1 was not an artifact of item selection but indeed due to benefits of successful retrieval on retention. Moreover, Experiment 2 showed that a combined Retrieval+Context condition did not enhance performance compared to a pure Retrieval condition, suggesting that benefits of context-inferences are limited once learners can retrieve the meaning of words from memory. Finally, in Experiment 3, the testing effect was replicated with a final test that presented words in a sentence context. Both accuracy and confidence on this test

were higher after retrieval practice than context-inference practice, showing that the testing effect was not restricted to recall tests. Overall, the three experiments showed that memory retrieval enhances the long-term retention of vocabulary words more than context-inferences, as we had predicted based on the extensive testing effect literature (e.g., Roediger & Butler, 2011; Rowland, 2014) and the comparably limited empirical support for benefits of context-inferences on word retention (Mondria, 2003; Mondria & Wit-de Boer, 1991). The fact that reducing the amount of contextual information to trigger memory retrieval had a consistent positive influence on word retention confirms that testing effects cannot only be evoked through explicit recall exercises but also more indirectly by creating a need to retrieve information from memory when that information is not accessible from context.

The present results appear at odds with the widely-held view that an informative context is conducive to word learning because contextual clues facilitate the inference of word meaning (e.g., Seibert, 1945), and understanding a word's meaning is necessary to establish a form-meaning connection (Li, 1988). However, the comprehension of words in context (e.g., during reading) and the retention of words over time are distinct processes (Lawson & Hogben, 1996; Verspoor & Lowie, 2003). As a case in point, the present study showed that the amount of contextual information affected comprehension and retention in different ways: Contextual information increased the chance that learners found the correct word meaning during practice, but it reduced the retention of these words over time. This somewhat counterintuitive finding parallels other learning conditions that facilitate practice but lead to worse long-term outcomes, such as massed repetition (as opposed to spaced repetition) and continuous practice with the same task (as opposed to practice with varying tasks) (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Nakata, 2016a; Nakata & Webb, 2016). Such conditions lead to high performance during practice but allow learners to bypass the effort and engagement necessary for durable learning, resulting in worse long-term outcomes (Bjork, 1994; Yan, Clark, & Bjork, 2016). In contrast, conditions like retrieval constitute so-called *desirable difficulties* (term coined by Bjork, 1994) that require more effortful, often slower and more error-prone processing of the to-be learned information but also lead to better long-term outcomes. Following this framework, reducing contextual information in our experiments likely created a desirable difficulty because learners had to engage in effortful retrieval, whereas rich contextual information gave learners easy access to word meaning and involved only superficial processing of the form-meaning association. These results demonstrate that it is crucial to consider the effects that a manipulation of the context has not only on comprehension of words, but also on the way in which learners process and subsequently remember the words.

Although the present study showed that reducing contextual information during practice can enhance word learning, it should not be seen as an argument that words should always be presented in an uninformative context rather than an informative context. Whether difficulties during learning are desirable, depends on learner capabilities and prior knowledge (McNamara, Kintsch, Songer, & Kintsch, 1996). Here, we focused at the effect of contextual richness during *later* repetitions of words, which are supposed to be "one of the most important phases in vocabulary learning which has not been researched sufficiently" (Peters, Hulstijn, Sercu, & Lutjeharms, 2009, p. 118). During later repetitions of words, learners have already acquired some word knowledge that must be consolidated through further repetitions. In this situation, a reduction of contextual information was found to be beneficial if learners succeeded at retrieving word meaning from memory. In contrast, during initial repetitions, learners are less likely to successfully retrieve word meaning from memory so that the trade-off between better comprehension with more context and better retention over time with less context poses a larger challenge. The previous finding that exposure to an uninformative context might become more beneficial after a number of prior exposures (Webb, 2008) support this idea. One solution to ensure that learners can benefit from retrieval opportunities already earlier during practice could be to provide feedback, as in Experiment 2 and 3; another solution might be to reduce contextual richness gradually over the course of several repetitions (see also Finley, Benjamin, Hays, Bjork, & Kornell, 2011). More research is needed to establish if under such conditions, a reduction of contextual clues is beneficial already during earlier stages of learning.

Given the importance of feedback in the present experiments, it is noteworthy that the literature on testing effects is not limited to retrieval practice for consolidation of previously learned materials. It also describes so-called test-potentiated encoding. This entails that information that is provided after learners have (unsuccessfully) attempted to retrieve the information from memory, is remembered better than information that is directly presented to learners (e.g., Richland, Kornell, & Kao, 2009). Possibly, retrieval attempts enhance learners' involvement (similar concepts have been discussed in the language literature as the *need* to process a word (Laufer & Hulstijn, 2001)) or lead to a more thorough inspection of available cues, such as the word form. In any case, studies on test-potentiated learning suggest that learners might benefit from retrieval attempts even when the retrieval fails, as long as corrective feedback is available (Rowland & DeLosh, 2015; Chapter 2: van den Broek et al., 2014). In the present study, we did not distinguish between the indirect effects of retrieval on

feedback processing and the direct effects of the retrieval itself[4], but the mechanisms of feedback processing could be interesting for follow-up research. For example, one practically relevant question is whether feedback after a retrieval attempt has to be explicit, or whether a context sentence from which word meaning can be derived, is similarly effective.

A number of directions for future research can be derived from characteristics and limitations of the present study. First, in the present study, learners practiced with single sentences, typed in the translation of target words, and saw the words' translations as feedback to their responses. These are characteristics of intentional vocabulary practice. An interesting venue for future studies would be to test whether reducing contextual information can also trigger retrieval and enhance word retention in more incidental learning situations, such as during the study of text passages or free reading. It is not clear whether retrieval can be triggered in the same way in these situations. For instance, learners pay less attention to novel words in longer passages than in single sentences (Wochna & Juhasz, 2013). Moreover, learners regularly ignore novel words during free reading (e.g., Hulstijn et al., 1996). On the other hand, a response may not be necessary to obtain benefits of retrieval. Covert retrieval – thinking of an answer but not providing an overt response – produces similar benefits for retention as overt retrieval (Smith, Roediger, & Karpicke, 2013). Therefore, it would be interesting to see if reading materials for language learners such as short texts in handbooks or guided readers could be adapted to trigger the retrieval of target word meaning. Feedback could be realized through glossaries, or by providing contextual information a few sentences after the retrieval cue, similar to the combined Retrieval+Context condition in Experiment 2.

Second, this study presented learners with foreign words in a first-language context. The L1-context allowed us to manipulate contextual richness while ensuring that all words in the context were known to the learners and all target words were unknown. This manipulation, too, may have had a benefit especially for the Context-inference condition because it made it more likely that learners understood the contextual clues. It is thus not likely that retrieval benefits were due to the choice of language. Nevertheless, in practice, a text in the foreign language may be useful for learners to strengthen their knowledge of the words in the context in addition to learning specific experimental target words. It is therefore a relevant question if the

4 For further information about the cognitive mechanisms that might underlie benefits of retrieval practice on retention, we refer readers to discussions in the recent literature (e.g., Carpenter & Yeung, 2017; Whiffen & Karpicke, 2017; for overview publications, see Butler & Roediger, 2010; Rowland, 2014). See also Rowland, Littrell-Baez, Sensenig, and DeLosh (2014) on retrieval-induced suppression in mixed list designs.

effect of contextual richness found in the present study is the same when learners read texts in a foreign language. Moreover, it remains to be tested if contextual richness has the same effect when learners try to acquire conceptually complex words. Studies on word learning in the first language suggest that in this case, more presentations in an informative context might be beneficial – at least until comprehension has been achieved (see Frishkoff, Perfetti, & Collins-Thompson, 2010; Frishkoff et al., 2011).

Finally, we focused at words presented in one either very informative or neutral, uninformative context, to isolate the effect of memory retrieval from context-inferences. In reality, the context surrounding a word falls on a continuum from defining to uninformative to misleading (see also Webb, 2008). This raises additional questions, for example, whether retrieval is also beneficial if it is triggered by a distracting or irrelevant context, and whether retrieval is beneficial compared to more effortful context-inferences. Some authors have argued that the effort involved in inferences may increase deeper processing and lead to higher retention (e.g., Hu & Nassaji, 2012; Haastrup, 1989 in Nation, 2001). A related point is that deeper or more beneficial processing may have occurred if words had been inferred from different context sentences instead of repeatedly the same context. Encoding variability is thought to enhance memory by creating additional associations that enrich memory representations and make them easier to be re-activated (Benjamin & Tullis, 2010). Context variability specifically has been shown to be beneficial for learning word meaning (e.g., Bolger, Balass, Landen, & Perfetti, 2008). On the other hand, varying the context may further draw participants' attention to comprehension instead of the novel word form, and could therefore lead to weaker form-meaning associations.

In conclusion, this study focused at the influence of contextual richness on word learning. Three experiments showed that practice of newly learned words in an uninformative context that required memory retrieval improved word retention compared to context-inference practice in an informative context. These testing effects were found on different outcome measures both immediately and seven days after learning, and including recall of word form, meaning, and the recognition of words in context. Reducing contextual information creates desirable difficulties during vocabulary practice; when done after some prior encoding, it is a means to trigger testing effects and to enhance the long-term retention of words.

6.6 REFERENCES

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baleghizadeh, S., & Shahry, M. N. N. (2011). The effect of three consecutive context sentences on EFL vocabulary-learning. *TESL Canada Journal*, *28*(2), 74–89. https://doi.org/10.18806/tesl.v28i2.1073

Barcroft, J. (2002). Semantic and structural elaboration in l2 lexical acquisition. *Language Learning*, *52*(2), 323–363. https://doi.org/10.1111/0023-8333.00186

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, *57*(1), 35–56. https://doi.org/10.1111/j.1467-9922.2007.00398.x

Barcroft, J. (2015). Can retrieval opportunities increase vocabulary learning during reading? *Foreign Language Annals*, *48*(2), 236–249. https://doi.org/10.1111/flan.12139

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beck, I. L., McKeown, M. G., & McCaslin, E. S. (1983). Vocabulary development: All contexts are not created equal. *The Elementary School Journal*, *83*(3), 177–181. https://doi.org/10.1086/461307

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology*, *61*(3), 228–247. https://doi.org/10.1016/j.cogpsych.2010.05.004

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

Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. A. (2008). Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes*, *45*(2), 122–159. https://doi.org/10.1080/01638530701792826

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, *13*(5), 826–830. https://doi.org/10.3758/BF03194004

Carpenter, S. K., Sachs, R. E., Martin, B., Schmidt, K., & Looft, R. (2012). Learning new vocabulary in German: The effects of inferring word meanings, type of feedback, and time of test. *Psychonomic Bulletin & Review*, *19*(1), 81–86. https://doi.org/10.3758/s13423-011-0185-7

Carpenter, S. K., & Yeung, K. L. (2017). The role of mediator strength in learning from retrieval. *Journal of Memory and Language*, *92*, 128–141. https://doi.org/10.1016/j.jml.2016.06.008

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Choi, S., Kim, J., & Ryu, K. (2014). Effects of context on implicit and explicit lexical knowledge: An event-related potential study. *Neuropsychologia*, *63*, 226–234. https://doi.org/10.1016/j.neuropsychologia.2014.09.003

Deconinck, J., Boers, F., & Eyckmans, J. (2015). "Does the form of this word fit its meaning?" The effect of learner-generated mapping elaborations on L2 word recall. *Language Teaching Research*, *21*(1), 31–53. https://doi.org/10.1177/1362168815614048

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*(2), 170–177. https://

doi.org/10.1037/1082-989X.1.2.170

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Finley, J. R., Benjamin, A. S., Hays, M. J., Bjork, R. A., & Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language*, *64*(4), 289–298. https://doi.org/10.1016/j.jml.2011.01.006

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*(2), 273–293. https://doi.org/10.2307/40264523

Frishkoff, G. A., Collins-Thompson, K., Hodges, L., & Crossley, S. (2016). Accuracy feedback improves word learning from context: Evidence from a meaning-generation task. *Reading and Writing*, *29*(4), 609–632. https://doi.org/10.1007/s11145-015-9615-7

Frishkoff, G. A., Perfetti, C. A., & Collins-Thompson, K. (2010). Lexical quality in the brain: ERP evidence for robust word learning from context. *Developmental Neuropsychology*, *35*(4), 376–403. https://doi.org/10.1080/87565641.2010.480915

Frishkoff, G. A., Perfetti, C. A., & Collins-Thompson, K. (2011). Predicting robust vocabulary growth from measures of incremental learning. *Scientific Studies of Reading*, *15*(1), 71–91. https://doi.org/10.1080/10888438.2011.539076

Fukkink, R. G., & de Glopper, K. (1998). Effects of instruction in deriving word meaning from context: A meta-analysis. *Review of Educational Research*, *68*(4), 450–469. https://doi.org/10.2307/1170735

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., & Tabbers, H. K. (2014). The effect of retrieval practice in primary school vocabulary learning. *Applied Cognitive Psychology*, *28*(1), 135–142. https://doi.org/10.1002/acp.2956

Grace, C. A. (1998). Retention of word meanings inferred from context and sentence-level translations: Implications for the design of beginning-level call software. *The Modern Language Journal*, *82*(4), 533–544. https://doi.org/10.2307/330223

Haastrup, K. (1991). *Lexical inferencing procedures, or, talking about words: Receptive procedures in foreign language learning with special reference to english*. Tuebingen, Germany: Narr.

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*(4), 801–812. https://doi.org/10.1037/a0023219

Hu, H. M., & Nassaji, H. (2012). Ease of inferencing, learner inferential strategies, and their relationship with the retention of word meanings inferred from context. *Canadian Modern Language Review*, *68*(1), 54–77. https://doi.org/10.1353/cml.2011.0036

Hulstijn, J. H. (1992). Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 113–125). London: Maxmillan.

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal*, *80*(3), 327–339. https://doi.org/10.1111/j.1540-4781.1996.tb01614.x

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*(3), 539–558. https://doi.org/10.1111/0023-8333.00164

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775. https://doi.org/10.1126/science.1199327

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning. *Psychology of Learning and Motivation*, *61*, 237–284. https://doi.org/10.1016/B978-0-12-800283-4.00007-1

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

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

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. https://doi.org/10.1016/j.jml.2011.04.002

Kuhn, M., & Stahl, S. (1998). Teaching children to learn word meanings from context: A synthesis and some questions. *Journal of Literacy Research*, *30*(1), 119–138. https://doi.org/10.1080/10862969809547983

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*(1), 1–26. https://doi.org/10.1093/applin/22.1.1

Lawson, M. J., & Hogben, D. (1996). The vocabulary-learning strategies of foreign-language students. *Language Learning*, *46*(1), 101–135. https://doi.org/10.1111/j.1467-1770.1996.tb00642.x

Li, X. (1988). Effects of contextual cues on inferring and remembering meanings of new words. *Applied Linguistics*, *9*(4), 402–413. https://doi.org/10.1093/applin/9.4.402

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological Science*, *25*(3), 639–647. https://doi.org/10.1177/0956797613504302

McNamara, D., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*(1), 1–43. https://doi.org/10.1207/s1532690xci1401_1

Mondria, J.-A. (2003). The effects of inferring, verifying, and memorizing on the retention of L2 word meanings. *Studies in Second Language Acquisition*, *25*(04), 473–499. https://doi.org/https://doi.org/10.1017/S0272263103000202

Mondria, J.-A., & Wit-de Boer, M. (1991). The effects of contextual richness on the guessability and the retention of words in a foreign language. *Applied Linguistics*, *12*(3), 249–267. https://doi.org/10.1093/applin/12.3.249

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Nakata, T. (2016a). Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. *Studies in Second Language Acquisition*, Advance online publication. https://doi.org/10.1017/S0272263116000280

Nakata, T. (2016b). Effects of retrieval formats on second language vocabulary learning. *International Review of Applied Linguistics in Language Teaching*, *54*(3), 257–289. https://doi.org/10.1515/iral-2015-0022

Nakata, T., & Webb, S. (2016). Does studying vocabulary in smaller sets increase learning? The effects of part and whole learning on second language vocabulary acquisition. *Studies in Second Language Acquisition*, *38*(3), 523–552. https://doi.org/10.1017/S0272263115000236

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press. Retrieved from https://www.cambridge.org/core/books/learning-vocabulary-in-another-language/491314AA1B451AD04F3536000F1C9F0D

Nation, I. S. P. (2015). Principles guiding vocabulary learning through extensive reading. *Reading in a Foreign Language*, *27*(1), 136–145.

Peters, E., Hulstijn, J. H., Sercu, L., & Lutjeharms, M. (2009). Learning L2 German vocabulary through reading: The effect of three enhancement techniques compared. *Language Learning*, *59*(1), 113–151. https://doi.org/10.1111/j.1467-9922.2009.00502.x

Pressley, M., Levin, J. R., & McDaniel, M. A. (1987). Remembering versus inferring what a word means: Mnemonic and contextual approaches. In M. G. McKeown & M. E. Curtis (Eds.), *The Nature of Vocabulary Acquisition* (pp. 107–127). Hillsdale, NJ: Lawrence Erlbaum.

Prince, P. (1996). Second language vocabulary learning: the role of context versus translations as a function of proficiency. *The Modern Language Journal*, *80*(4), 478–493. https://doi.org/10.1111/j.1540-4781.1996.tb05468.x

R Core Team. (2014). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. Retrieved from http://www.R-project.org/

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243–257. https://doi.org/10.1037/a0016496

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

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

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

Rowland, C. A., & DeLosh, E. L. (2015). Mnemonic benefits of retrieval practice at short retention intervals. *Memory*, *23*(3), 403–419. https://doi.org/10.1080/09658211.2014.889710

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

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363. https://doi.org/10.1177/1362168808089921

Schneider, V. I., Healy, A. F., & Bourne, L. E. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, *46*(2), 419–440. https://doi.org/10.1006/jmla.2001.2813

Schouten-van Parreren, C. A. (1989). Vocabulary learning through reading: Which conditions should be met when presenting words in texts. *AILA Review*, *6*(1), 75–85.

Seibert, L. C. (1945). A study on the practice of guessing word meanings from a context. *The Modern Language Journal*, *29*(4), 296–322. https://doi.org/10.2307/318219

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An individual's rate of forgetting is stable over time but differs across materials. *Topics in Cognitive Science*, *8*(1), 305–321. https://doi.org/10.1111/tops.12183

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

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, *56*(4), 252–257. https://doi.org/10.1027/1618-3169.56.4.252

van den Broek, G. S. E., Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? Insights from immediate and delayed retrieval speed. *Memory*, *22*(7), 803–812. https://doi.org/10.1080/09658211.2013.831455

Veltre, M. T., Cho, K. W., & Neely, J. H. (2015). Transfer-appropriate processing in the testing effect. *Memory*, *23*(8), 1229–1237. https://doi.org/10.1080/09658211.2014.970196

Verspoor, M., & Lowie, W. (2003). Making sense of polysemous words. *Language Learning*, *53*(3), 547–586. https://doi.org/10.1111/1467-9922.00234

Webb, S. (2007a). Learning word pairs and glossed sentences: the effects of a single context on vocabulary knowledge. *Language Teaching Research*, *11*(1), 63–81. https://doi.org/10.1177/1362168806072463

Webb, S. (2007b). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65. https://doi.org/10.1093/applin/aml048

Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language*, *20*(2), 232–245.

Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*, *43*(7), 1036–1046. https://doi.org/10.1037/xlm0000379

Winstanley, P. A. D. (1996). Generation effects and the lack thereof: The role of transfer-appropriate processing. *Memory*, *4*(1), 31–48. https://doi.org/10.1080/741940667

Wochna, K. L., & Juhasz, B. J. (2013). Context length and reading novel words: An eye-movement investigation. *British Journal of Psychology*, *104*(3), 347–363. https://doi.org/10.1111/j.2044-8295.2012.02127.x

Yan, V. X., Clark, C. M., & Bjork, R. A. (2016). Memory and metamemory considerations in the instruction of human beings revisited: Implications for optimizing online learning. In J. C. Horvath, J. M. Lodge, & J. Hattie (Eds.), *From the laboratory to the classroom: Translating science of learning for teachers*. London, UK: Routledge.

Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *Canadian Modern Language Review*, *57*(4), 541–572.

6.7 APPENDIX

6.7.1 PILOT PROCEDURE TO CONSTRUCT PRACTICE SENTENCES

The practice items were piloted in two online experiments with 49 participants in the first pilot round and 68 participants in the second pilot round. Each participant read half of the sentences in the Retrieval condition and half in the Context-inference condition, whereby Swahili words were replaced by a blank (Pilot 1) or introduced as pseudowords (Pilot 2). The participants were asked to fill in up to three words that could complete the sentence (Pilot 1) or up to three possible meanings of the "pseudoword" (Pilot 2) and rate on a 5-point scale how confident they were that their first guess was correct. After Pilot 1, problematic sentences were removed or changed before presentation in Pilot 2. The sentences eventually chosen for the experiment fulfilled the following criteria: (1) For the version of the sentence with rich information (context-inference condition), more than 75% of the participants filled in the correct word or a closely related synonym such as "cap" instead of "hat" as first choice. (2) For the version without information about word meaning (retrieval condition), (a) more than 75% of the participants responded with low certainty (< 3 on a 5-point scale), (b) less than 25% of the participants filled in the target word as first or second choice, and (c) no other word was filled in by more than 50% of the participants as first or second choice. In short, the stimuli were designed in such a way that a participant without prior exposure to the words, like the pilot participants, would still be able to derive the meaning of the target words from the context-inference sentences but would not know what the words in the retrieval sentences mean, and would associate neither the target word nor a different word consistently with the retrieval sentences.

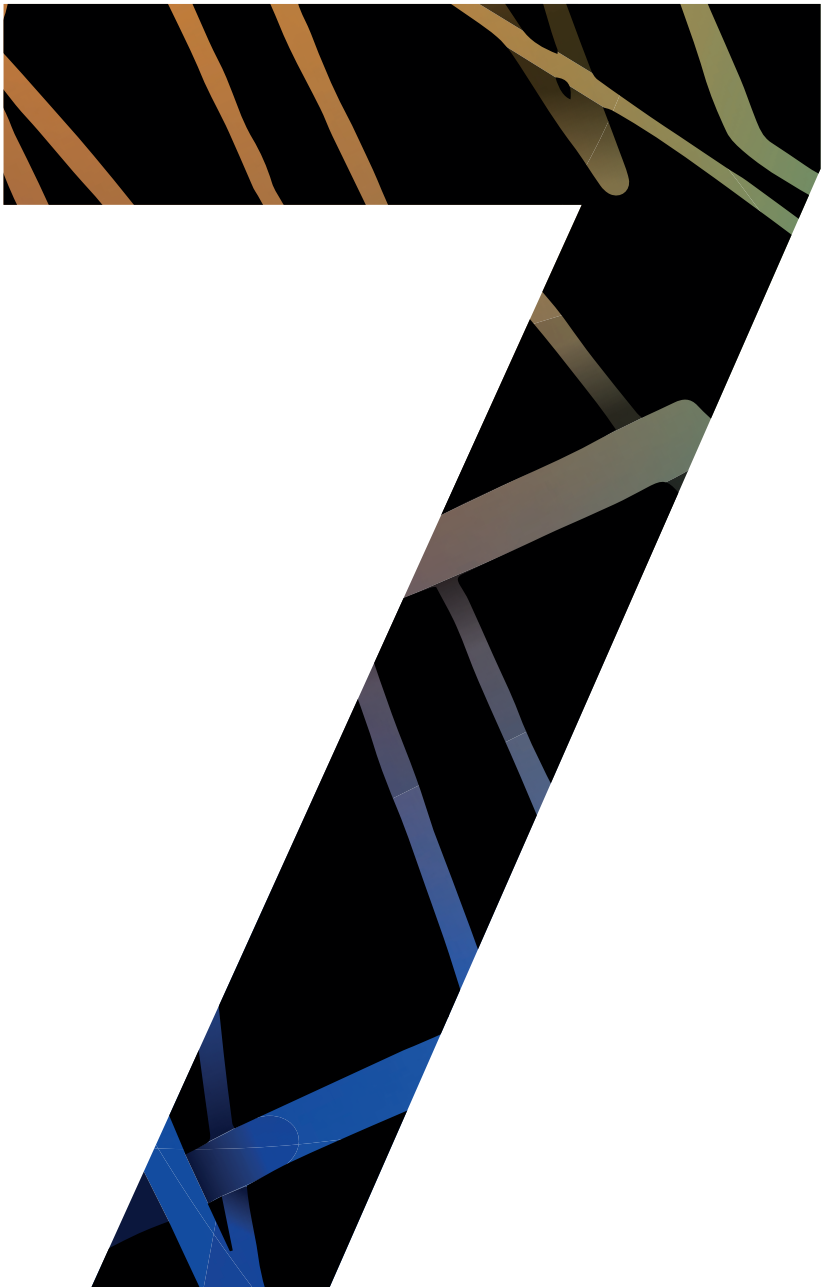
6.7.2 ADDITIONAL INFORMATION ON DATA REPORTED IN FIGURE 6.2

Table 6.3 Mean Accuracy of Receptive Recall, by Practice Condition and Number of Correct Translations during Practice

	Number of correct translations during practice			
Condition	*0 correct*	*1 correct*	*2 correct*	*3 correct*
Context-Inference	0.07	0	0.38	0.61
Retrieval	0.04	0.26	0.53	0.75

Table 6.4 Number of Observations, by Practice Condition and Number of Correct Translations during Practice

	Number of correct translations during practice				
Condition	*0 correct*	*1 correct*	*2 correct*	*3 correct*	*Total*
Context-Inference	114	44	132	4390	4680
Retrieval	852	188	270	3370	4680



GENERAL DISCUSSION

Abstract. This dissertation explores the cognitive and neural underpinnings of retrieval practice and the effects of retrieval practice during vocabulary exercises. The final chapter provides a summary of the main results of the previous chapters and discusses overarching theoretical and practical implications. The retrieval of information from memory is not a simple read-out process; each retrieval act changes the future accessibility of memories. Retrieval *practice* is an effective technique for learners who want to remember large amounts of information, for example, when practicing vocabulary in a foreign language. Results from behavioral and neuroimaging studies suggest that retrieval is a controlled, effortful process that becomes facilitated with practice. In case of word learning, this facilitation might involve the selective strengthening of form-meaning associations through semantic elaboration or the inhibition of competing information. Retrieval processes can become dependent on the support that is available during practice, such as prompts and contextual information. Therefore, retrieval practice with extensive support has limited benefits for later recall without support. An additional boundary condition is that benefits of retrieval practice depend on retrieval success or the availability of feedback. The findings reported in this thesis lead to a number of practical recommendations for the design of learning situations, which are presented at the end of this chapter.

The majority of laymen assume that human memory works like a video camera, which records, stores, and later replays fixed memories (Simons & Chabris, 2011). However, the benefits of retrieval practice show that this view is incorrect and memory is a more complex, dynamic system. The retrieval of information from memory is not a simple replay process; each retrieval act changes the future accessibility of memories (Dudai, 2012). This characteristic of memory is of great interest for learners who need to remember large amounts of information, such as language learners who need to learn thousands of words to master a new language (Schmitt, Cobb, Horst, & Schmitt, 2017). For them, retrieval practice is a promising technique to enhance the retention of words over time - more promising, for example, than studying words with translations (e.g., Carrier & Pashler, 1992). The first part of this thesis focuses on the cognitive and neural underpinnings of this *testing effect.* In spite of a wealth of research on testing effects, the underlying mechanisms are poorly understood (Roediger & Butler, 2011). Using reaction time measures and neuroimaging data, an attempt was made to go beyond measures of recall accuracy to better understand why retrieval practice is beneficial for the long-term retention of words. The second part of the thesis focuses on the integration of memory retrieval in vocabulary exercises. As described in the introduction, retrieval is not a frequent topic in vocabulary learning studies (cf. Barcroft, 2007, 2015). Therefore, two studies were conducted to test the effect of stimulating retrieval during vocabulary exercises, during adaptive computerized practice with feedback and during contextualized vocabulary practice. This final chapter provides a summary of the main results of each part of the thesis, before discussing theoretical and practical implications.

7.1 SUMMARY OF MAIN FINDINGS

7.1.1 COGNITIVE MECHANISMS AND NEURAL CORRELATES OF RETRIEVAL PRACTICE.

The first aim of this thesis was to provide insight into the cognitive mechanisms that underlie the benefits of memory retrieval for vocabulary learning. Reaction time data and neuroimaging data were used to test predictions derived from accounts of the testing effect. First, it was shown in Chapter 2 that elaboration (Carpenter, 2009, 2011) and selection accounts (Lehman, Smith, & Karpicke, 2014; Thomas & McDaniel, 2013) can explain changes in testing effects over time. Then, these accounts were evaluated against neural correlates of testing effects in Chapter 3 and 4.

7.1.1.1 EXPLAINING THE TIMING OF TESTING EFFECTS. Chapter 2 provides evidence for the bifurcation model (Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011), which explains why testing effects are more robust on delayed tests than on

immediate tests (Rowland & DeLosh, 2015; Toppino & Cohen, 2009). According to the model, low difficulty of the final test and limited retrieval success during practice can create the false impression that testing effects only occur on delayed tests although successful retrieval actually immediately leads to stronger memories than restudying. The reaction times reported in Chapter 2 support the assumption of the model that items recalled after retrieval practice are more accessible than items recalled after restudying. Both immediately and seven days after practice, students translated words more quickly after prior retrieval than after restudy practice. I interpreted this as higher memory strength after retrieval practice because stronger, more accessible memories are thought to lead to faster response times (J. R. Anderson, 1981; MacLeod & Nelson, 1984). Moreover, learning outcomes on a later test were related to retrieval success during practice. As predicted by the bifurcation model, recall was significantly higher for those items that were successfully retrieved during practice than for restudied items (and unsuccessfully retrieved items). Together, these results support the proposition of the bifurcation model that changes in testing effects over time can be an artifact of limited retrieval success during practice and low difficulty of the final test. Mechanistic accounts which predict that retrieval induces immediate beneficial changes in semantic associations, like the elaboration (Carpenter, 2009, 2011) or selection accounts (Lehman et al., 2014; Thomas & McDaniel, 2013), are thus compatible with reports of delayed testing effects. Moreover, faster response times after retrieval practice than after restudying suggest that retrieval indeed becomes facilitated with repetition. This is an important addition to the literature because the facilitation of retrieval processes has so far mostly been measured as recall accuracy.

7.1.1.2 NEURAL CORRELATES OF TESTING EFFECTS. Two main results of the empirical study reported in Chapter 3 were replicated by other studies reviewed in Chapter 4: Higher activation in ventrolateral prefrontal cortex (VLPFC) during retrieval practice than during restudy practice, and differences in the relation between later recall and inferior parietal and middle temporal activation during retrieval and restudying. These results were respectively related to the retrieval effort hypothesis (Pyc & Rawson, 2009), and to elaboration (Carpenter, 2009, 2011) and selection accounts (Lehman et al., 2014; Thomas & McDaniel, 2013).

Prefrontal activations during memory retrieval are typically interpreted as demands on top-down cognitive control, for example, to enable the activation of target information among competing memories (e.g., Badre & Wagner, 2007; Blumenfeld & Ranganath, 2007). Higher VLPFC activation during retrieval than during restudying thus suggests stronger demands on cognitive control. Adding to this finding from Chapter 3, other fMRI studies reviewed in Chapter 4 showed that over the course of repeated retrieval practice (Karlsson-Wirebring et al., 2015; Wimber, Alink, Charest, Kriegeskorte, & Anderson, 2015), and during a final recall test after

prior retrieval practice compared to prior restudying (Eriksson, Kalpouzos, & Nyberg, 2011), activations in VLPFC were reduced. Together, these findings are consistent with the *retrieval effort hypothesis* that retrieval leads to more effortful, controlled processing than restudying (Pyc & Rawson, 2009) and the general idea that retrieval becomes facilitated with practice (e.g., Carpenter, 2009; Lehman et al., 2014).

The facilitation of retrieval processes with practice is thought to be due to changes in associations within semantic memory networks, for example, through elaboration or selective strengthening of mental associations between two words (Carpenter, 2009, 2011; Karpicke & Zaromb, 2010; Thomas & McDaniel, 2013). Therefore, I focused on the engagement of temporo-parietal areas during retrieval and restudying. In the experiment reported in Chapter 3, the angular gyrus and supramarginal gyrus in the inferior parietal lobe (IPL) and part of the middle temporal gyrus (MTG) were more active during restudying compared to retrieval practice, but activation in these areas was only predictive of later memory when measured during retrieval practice but not when measured during restudying. Other studies that compared brain activation during retrieval and restudying reported similar results (Vannest et al., 2012; Wing, Marsh, & Cabeza, 2013), as reviewed in Chapter 4. The IPL and MTG have been related to semantic processing (meta-analysis in Binder, Desai, Graves, & Conant, 2009), possibly as higher order association areas that bind individual concepts into coherent semantic combinations (Binder & Desai, 2011; Price, Bonner, Peelle, & Grossman, 2015) or as part of a cognitive control network involved in semantic processing (Noonan, Jefferies, Visser, & Lambon Ralph, 2013; Ralph, Jefferies, Patterson, & Rogers, 2017). Tentatively, we related the lower activation in these areas during retrieval than during restudying to reduced semantic processing. Furthermore, the fact that activations during retrieval but not during restudying were predictive of subsequent memory, suggests that the reduced activation during retrieval could reflect a beneficial process, like a focus of attention on relevant information. These finding are further discussed in Section 7.2.1 *Elaboration, Selection, or Selective Elaboration?*.

7.1.2 RETRIEVAL PRACTICE DURING VOCABULARY EXERCISES.

The second aim of this thesis was to investigate the effect of retrieval opportunities during vocabulary exercises. The classroom experiments reported in Chapter 5 revealed two boundary conditions. First, there was only limited transfer of retrieval practice with hints to later recall situations without hints. Second, time constraints meant that longer feedback processing reduced the time available for further repetitions. In Chapter 6, triggering retrieval through a manipulation of the context in which the target words appeared produced robust benefits both on later recall and on the later recognition of words in context. This effect depended on retrieval success during practice.

7.1.2.1 RETRIEVAL AND FEEDBACK. Chapter 5 contains a study on the effect of retrieval during feedback processing. As demonstrated in Chapter 2, if a learner cannot retrieve the translation of a word from memory, the retrieval attempt has few benefits for the form-meaning association unless followed by feedback that creates a new encoding opportunity (see also Jang, Wixted, Pecher, Zeelenberg, & Huber, 2012; Kornell et al., 2011). Feedback therefore significantly increases the benefits of retrieval practice, as it allows learners to correct their errors and strengthen unconfident responses (Butler & Roediger, 2008; Thomas & McDaniel, 2013). Most studies on retrieval practice contain simple show-answer feedback, but elaborate feedback might be more beneficial (van der Kleij, Feskens, & Eggen, 2015). We therefore conducted three classroom experiments in which we tested whether feedback can be made more effective if it includes an extra retrieval opportunity. High school students practiced vocabulary words from a foreign language with an adaptive computer programme (described in Sense, Behrens, Meijer, & van Rijn, 2016), which either contained simple show-answer feedback or hints feedback with which students could try again to retrieve the answer from memory.

None of the experiments in Chapter 5 produced evidence that hints feedback was preferable over show-answer feedback. Across three experiments with orthographic, mnemonic and cross-language hints, hints feedback led to a shift in the distribution of the available practice time from further repetitions to longer feedback processing after errors. Nevertheless, hints did not reduce (repeated) errors during practice. There was also no positive effect of hints feedback on learning outcomes on a recall test several days after learning. Unexpectedly, the only effects of hints feedback were found when the hints from practice were available again on the test (in Exp. 1 and 2). Compared to the show-answer condition, students used more orthographic recall prompts on the final test after practice with orthographic hints and showed better recall on a later test with mnemonic hints after practice with mnemonic hints feedback. This suggests that students do not automatically transfer what they retrieve *with* hints during practice to a recall situation *without* hints, which could reflect how retrieval became dependent on the cues available during practice (Smith & Handy, 2016). Overall, the findings reported in Chapter 5 extended the small number of prior studies on hints feedback, demonstrating that hints feedback is not generally more beneficial than standard feedback (Hall, Adams, & Tardibuono, 1968; Kornell & Vaughn, 2016; Kornell, Klein, & Rawson, 2015; but see Finn & Metcalfe, 2010). The common preconception that hints feedback is beneficial for learning may not hold in realistic learning situations and under time constraints; the elaborate feedback processing had few benefits for learning but reduced the time for further repetitions.

7.1.2.2 RETRIEVAL AND CONTEXT. In the experiments reported in Chapter 6, we found that learners benefited more from practice with an uninformative sentence context that stimulated the retrieval of word meaning from memory than from practice with an informative sentence context that allowed learners to infer word meaning. This effect was reliably found when retrieval was successful or combined with feedback, both immediately and several days after learning, and on productive and receptive recall as well as accuracy and confidence of the recognition of words in context. Thus, the manipulation of the context in which target words appeared evoked a testing effect.

The finding that reducing contextual information enhanced word learning is perhaps counter-intuitive, because contextual information is an important source of information to establish the meaning of hitherto unknown words (Beck, McKeown, & McCaslin, 1983; Seibert, 1945). For this reason rich contextual information is usually considered to be beneficial for word learning (Schouten-van Parreren, 1989). However, understanding a word in context does not ensure that the word is also remembered over time (e.g., Hulstijn, 1992; Mondria & Wit-de Boer, 1991), and the value of contextual inferences for the *retention* of words has been debated (Pressley, Levin, & McDaniel, 1987). In line with this, the experiments in Chapter 6 showed that a weak context that required retrieval to access word meaning led to better retention than a rich context that made it easy for learners to infer word meaning from context. This demonstrates that, at least under certain conditions, retrieval practice can be a more effective vocabulary practice strategy than context inferences. Moreover, the study shows that retrieval cannot only be evoked through explicit recall exercises but also more indirectly by manipulating the context in which a target word appears.

7.2 THEORETICAL IMPLICATIONS

7.2.1 ELABORATION, SELECTION, OR SELECTIVE ELABORATION?

There is an ongoing debate in the literature about the mechanisms that underlie testing effects. As introduced in Chapter 1, some authors propose semantic elaboration of cue-target associations through the creation of additional retrieval routes (Carpenter, 2009, 2011; Carpenter & Yeung, 2017; Rawson, Vaughn, & Carpenter, 2015) whereas others emphasize selective processing that concentrates activation on the target response (Karpicke, Lehman, & Aue, 2014; Karpicke & Zaromb, 2010; Lehman et al., 2014; Thomas & McDaniel, 2013). The reaction time and fMRI results reported in Chapters 2, 3 and 4 were collected to evaluate these different accounts.

In the studies reported in Chapter 2 and 3, we showed that learners not only recalled *more* words after retrieval than after restudy practice, but also recalled the

retrieved words more *quickly*. These faster response times after retrieval practice suggest that words become more accessible in memory, possibly due to reduced processing steps needed for recall. This supports theoretical accounts that testing effects are due to increased efficiency of retrieval processes (e.g., Roediger & Butler, 2011). The reduced response times also seem to be more in line with selection accounts than with elaboration accounts. Based on models of memory that postulate that recall depends on the degree to which available cues activate cue-target associations *to the exclusion of competing associations* (e.g., Nairne, 2002), Karpicke et al. (2014) argued that elaborate semantic associations should make recall *less* efficient rather than *more* efficient. This is because elaborate semantic associations increase the amount of information that is activated in addition to the target information. Such a larger number of associations with a cue is thought to reduce the amount of activation that is passed on to associations with the target, causing lower accuracy and speed of target recall (Danker, Fincham, & Anderson, 2011). Following these arguments, increasing semantic elaboration over the course of repeated retrieval should lead to an increasing spread of activation and increasingly slow down recall as learners have to filter out more and more competing responses. Therefore, faster responses after retrieval practice than after restudying are difficult to explain under the assumptions of the elaboration account. Faster reaction times seem more in line with *selective* strengthening of cue-response associations which reduces the search-set of candidate items that become activated (Karpicke et al., 2014; Karpicke & Zaromb, 2010; Lehman et al., 2014; Thomas & McDaniel, 2013).

The patterns of brain activation during retrieval and restudy described in Chapters 3 and 4 are in line with such selective processing but nevertheless point at a role of semantic processing. Subsequent memory was predicted by activity increases in core semantic association areas in IPL and MTG when measured during retrieval but not when measured during restudy. This contradicts the idea that semantic processing during retrieval and restudying has similar beneficial effects for subsequent memory but is enhanced during retrieval, as was put forward in early semantic elaboration accounts (Carpenter & Delosh, 2006). Instead, results suggest that semantic processing during retrieval is more beneficial for memory than semantic processing during restudying, possibly because it reflects the activation of more relevant associations. I therefore proposed in Chapter 4 that an alternative cognitive model might be needed that can accommodate both semantic processing and the selective nature of retrieval. For example, elaboration during retrieval could selectively focus on associations that strengthen the word-meaning link, whereas associations that compete with the correct meaning may be suppressed. In line with this, Carpenter and Yeung recently suggested that the benefits of semantic elaboration during retrieval may not be due to "the quantity of information that can be activated during retrieval, but the strength

of the mediating information and how it facilitates cue-target connections" (Carpenter & Yeung, 2017, p. 139). Thus, elaborations may be beneficial if they selectively strengthen the association between a word and its translation, rather than increase the quantity of alternative associations in a broad semantic network. Different ideas have been put forward how cue-target associations could be selectively strengthened during retrieval in this way.

7.2.2 SELECTIVE STRENGTHENING OF CUE-TARGET ASSOCIATIONS DURING RETRIEVAL

7.2.2.1 INHIBITION OF COMPETING RESPONSES. Some researchers have argued that inhibitory control mechanisms suppress competing information that becomes activated during retrieval in order to limit (future) distraction by competitors (Kuhl, Dudukovic, Kahn, & Wagner, 2007; Storm & Levy, 2012; Wimber et al., 2015). These inhibition processes have been related to activity in prefrontal areas of the brain (e.g., Kuhl et al., 2007; Wimber et al., 2008; Wimber, Rutschmann, Greenlee, & Bäuml, 2009), which might contribute to a top-down control signal that resolves competition during retrieval (Badre & Wagner, 2007; Danker et al., 2011; Wimber et al., 2015). Following this interpretation, the higher activation in VLPFC observed during retrieval than restudy reported in Chapter 3 and 4 could reflect the controlled inhibition of incorrect translations to enable the retrieval of the correct translation. Moreover, reduced VLPFC activations as a consequence of retrieval practice could reflect reduced demands on inhibition as less and less competing information is activated. It should be noted, however, that although competitor-suppression is thought to enhance recall of target information, only limited empirical evidence exists for a relation between competitor suppression and target enhancement (Storm & Levy, 2012)[1]. This parallels the findings in Chapter 3 and 4, in which no relation was found between prefrontal activations during practice and later memory. Thus, the pattern of activations in VLPFC during retrieval fit the idea that a suppression of incorrect responses takes place, but it is not clear if this suppression influences the strengthening of cue-target associations and later recall.

1 Inhibition has mostly been investigated in the context of so-called retrieval-induced forgetting (i.e., retrieval-induced suppression of competing information). This is typically studied in experiments in which the selective retrieval of specific associations suppresses competing associations that are not retrieved. For example, practicing the response "pineapple" when cued with "fruit – p.. ?" inhibits the response "pear" (M. C. Anderson, Bjork, & Bjork, 1994). In such paradigms, the amount of prefrontal activations correlates with competitor suppression (e.g., Kuhl, Dudukovic, Kahn, & Wagner, 2007; Wimber, Alink, Charest, Kriegeskorte, & Anderson, 2015).

7.2.2.2 ELABORATION OF CUE-TARGET ASSOCIATIONS. Karpicke et al. (Karpicke et al., 2014; Karpicke & Zaromb, 2010; Lehman et al., 2014) proposed that the mechanism underlying the selective strengthening of cue-target associations during retrieval is context reinstatement. Supposedly, the context of earlier encounters with a word is re-activated during retrieval, and then becomes integrated with the contextual information that is available during retrieval. Over the course of repeated retrieval, this is thought to lead to a refinement or extension of the context representation associated with target information because effective cues are more likely reactivated. This context reinstatement account has so far mostly been used to explain benefits of retrieval on free recall (e.g., Whiffen & Karpicke, 2017). A different mechanism was recently proposed to explain testing effects on cue-target associations like word form-meaning associations (Carpenter & Yeung, 2017). In a revised (or reworded) version of Carpenter's original elaboration account (Carpenter, 2009, 2011), Carpenter and Yeung (2017) suggested that semantic information that is processed during retrieval might become *integrated* in cue-target associations. Thus, elaborations such as keyword mediators (Pyc & Rawson, 2010) or incorrect guesses (Potts & Shanks, 2014; Yan, Yu, Garcia, & Bjork, 2014) might become merged into the association, rather than form additional, possibly competing, associations. Tentatively, this could be an explanation why activation in IPL and MTG, which might reflect semantic processing, predicted later performance when measured during retrieval practice (in Chapter 3 and 4).

It should be noted that if mental operations like semantic elaborations are considered part of the encoding context of a word, the context reinstatement and the selective elaboration ideas share important properties: Both accounts predict that repeated retrieval leads to the refinement or extension of associations that connect cue and target information, thereby increasing the chance that later-on, available retrieval cues overlap with parts of the semantic network that lead to the activation of target information. This, in turn, could lead to a reduction of the search-set of candidate items that are activated and narrow the scope of the memory search to hone in on target information (Thomas & McDaniel, 2013). A question that remains for future research is what distinguishes unbeneficial, competing associations from beneficial, merged associations. At the time of writing, neither the elaboration nor the context reinstatement account has clearly specified this distinction. Another open question is whether associations change with repeated retrieval, as for example, keyword mediators that link vocabulary words to their translation are known to become less activated after extensive retrieval practice (Crutcher & Ericsson, 2000; Kole & Healy, 2013). There is a substantial literature on the increasing integration of word knowledge into semantic memory and the loss of episodic details through consolidation processes (Takashima & Bakker, in press). It could be an interesting

venue for future research to document the effect of repeated retrieval on such semantization processes.

To summarize, both the reaction time measures and neuroimaging results of VLPFC activations as a consequence of retrieval suggest that retrieval is a controlled, effortful process that becomes facilitated with practice. Reaction times and the differential involvement of temporo-parietal activations during restudying and retrieval suggest that testing effects are not due to generally enhanced semantic elaboration. Instead, the data are more compatible with the selective strengthening of cue-target associations. This might involve decreased activation of competing associations through active inhibition and/or selective strengthening of cue-target associations. Further research is needed to clarify the mechanisms underlying the selective strengthening of form-meaning associations through retrieval practice, but a refinement of the mental representation of the learning context (Karpicke et al., 2014; Whiffen & Karpicke, 2017) or strengthening of semantic associations (Carpenter & Yeung, 2017) might play a role.

7.3 PRACTICAL IMPLICATIONS FOR EFFECTIVE VOCABULARY LEARNING

Retrieval practice has a strong potential to enhance word retention. However, as the results in this thesis show, boundary conditions apply. These lead to a number of practical recommendations for instructional design.

7.3.1 THE IMPORTANCE OF RETRIEVAL SUCCESS

The importance of retrieval success for testing effects is a topic in all chapters of this thesis. Chapter 2 demonstrated that in the absence of feedback only successfully retrieved words benefitted from retrieval practice, thereby confirming a basic assumption of the bifurcation model. Similarly, testing effects were more pronounced in Chapter 3 and 6 when analyses were restricted to the successfully retrieved items than when all items were included. This importance of retrieval success creates a dilemma for designers of instructional situations, because there is a trade-off between the potency of retrieval and the probability that the retrieval succeeds (e.g., Finley, Benjamin, Hays, Bjork, & Kornell, 2011). As discussed in Chapter 5, potential benefits increase when retrieval is made more difficult (Carpenter & Delosh, 2006; Pyc & Rawson, 2009), but the chance that retrieval fails also increases. The context manipulation in Chapter 6 is a good example of this: the uninformative context that induced retrieval led to better retention only when combined with feedback. Without feedback, the beneficial effects of the retrieval opportunity were cancelled out by the

fact that learners more often accessed the correct word meaning in the informative context condition.

Chapter 5 introduced two possible approaches to deal with the trade-off between the potency of effortful retrieval and retrieval success. One approach is to adjust the retrieval difficulty to learner performance. There are promising technical solutions that allow the adjustment of the timing of repetitions during vocabulary practice, one of which was used in Chapter 5 (Sense et al., 2016). Another approach is to combine retrieval practice with feedback. Feedback allows learners to correct their errors and encode the correct answer. Although elaborate feedback may cost time compared to show-answer feedback (Chapter 5), receiving feedback about the correct answer clearly leads to better performance than not having access to the correct answer (van der Kleij et al., 2015). Likewise, show-answer feedback significantly enhances the benefits of retrieval practice compared to retrieval without feedback (e.g., Finn & Metcalfe, 2010; Kang et al., 2011; Kang, McDermott, & Roediger, 2007). Another argument for the use of feedback is that in addition to direct benefits of retrieval, retrieval enhances subsequent encoding (Grimaldi & Karpicke, 2012; Izawa, 1971; Richland, Kornell, & Kao, 2009). Such *test-potentiated encoding* was further discussed in Chapter 4, and predicts that learners' encoding of feedback may be enhanced by a prior failed retrieval attempt. From these findings, the following practical recommendations can be derived:

- **Recommendation (1). Combine retrieval with feedback** that allows learners to encode words that they cannot yet retrieve.
- **Recommendation (2). Provide repeated retrieval opportunities** so that learners benefit from both test-potentiated encoding of the feedback, and from the direct effects of successful retrieval.

7.3.2 THE SPECIFICITY OF RETRIEVAL PRACTICE

In the previous sections, I summarized that the available evidence suggests that retrieval practice involves the selective strengthening of cue-target associations and the suppression of competing information. The selectiveness of these associations influences learning outcomes when learners cannot use the specific associations that they strengthened during practice in a later recall situation. As a case in point, retrieval practice with support in the form of hints (Chapter 5) or rich contextual information (Chapter 6) did not enhance later recall when that support was not available anymore (i.e., on a recall test without hints or without context). These findings show that retrieval can become dependent on cues that are available during practice, hampering later recall without those cues. This is in line with several principles introduced in Chapter 1, including the transfer-appropriate processing account that the overlap

between practice and later recall predicts performance (see Veltre, Cho, & Neely, 2015 for a recent overview), and the idea that memories can become dependent on contextual information available during practice (Smith & Handy, 2016).

It is useful to consider the specificity of retrieval practice effects in the context of the new theory of disuse (Bjork & Bjork, 1992), which distinguishes between the current accessibility of words (*retrieval strength*) and their long-term retention (*storage strength*). Learners tend to mistake the momentary retrieval strength of memories for their storage strength, and tend to predict better later learning outcomes when they experience higher fluency during practice (Koriat & Bjork, 2006; Yan, Bjork, & Bjork, 2016). This is often incorrect because fluency during practice largely depends on cues available in the environment that determine the momentary retrieval strength, rather than on the long-term storage strength. More importantly, practice conditions that increase momentary retrieval strength lead to less increase in storage strength. Supposedly, these conditions "act like crutches that artificially support performance during practice. When those crutches are absent in the posttraining environment, performance collapses" (Bjork, 1994, p. 196). For example, massed practice of words leads to high fluency and retrieval success during practice, but worse later recall than interleaved practice (Yan, Bjork, et al., 2016). Similarly, contextual information made it easier for learners to translate the words during practice but led to worse later recall in Chapter 6. Therefore, an important practical recommendation is to ensure that practice conditions do not artificially increase performance during practice at the cost of learning. Practice conditions should correspond to the desired learning outcomes, so that transfer from practice to later recall situations is likely.

- **Recommendation (3): Make retrieval practice challenging and be conservative with measures that facilitate retrieval during practice, such as hints.** Otherwise later retrieval can fail if the support from practice is no longer available.
- **Recommendation (4): Distinguish between fluency of retrieval during practice and benefits for later performance.** Conditions that increase the ease of retrieval may lead to high performance and confidence during practice, but to worse retention over time.

7.3.3 USING LIMITED PRACTICE TIME

The main conclusion from the experiment in Chapter 5 is that a manipulation that enhances learning when investigated in isolation, namely retrieving the meaning of a word from memory instead of restudying the word together with its meaning, is not automatically a beneficial addition to each practice situation (see also Kornell et al., 2015). When added to adaptive retrieval practice, hints feedback took time away from further repetitions but did not enhance later recall. The fifth recommendation is therefore:

- **Recommendation (5): Consider that the effects of a manipulation can be complex and depend on the baseline condition**. Even a seemingly beneficial manipulation like replacing simple feedback with more elaborate feedback incurs costs because it takes study time away from other activities. With an efficient baseline condition, *more is not always more.*

7.3.4 RETRIEVAL IN DIFFERENT LEARNING SITUATIONS

The study reported in Chapter 6 shows that retrieval can be triggered not only in obvious recall tasks, but also in other situations in which a word or word meaning needs to be retrieved from memory. Language learners rely on memory retrieval in various situations: when they hear or read a word and the context does not immediately specify its meaning, when they are in conversation and feel a need to use a particular word, or when an exercise creates a need for them to retrieve a word form or meaning from memory. As Chapter 6 showed, such incidentally triggered word retrieval is likely to be beneficial for retention. Conversely, when vocabulary words are presented in a highly restrictive context, or together with a translation, retrieval may not occur.

- **Recommendation (6): Integrate opportunities for retrieval in different learning situations.** Small manipulations of exercises can suffice to trigger retrieval, for example, by presenting a word first in an uninformative context and only later in an informative context (see Chapter 6). The previous recommendations apply: Retrieval should be followed by feedback and if possible, the difficulty of retrieval should be adjusted to learners' performance to create challenging but successful retrieval practice.

7.4 LIMITATIONS OF THE PRESENT WORK AND RECOMMENDATIONS FOR FUTURE RESEARCH

The studies in this thesis draw from different research disciplines, using methodologies from cognitive neuroscience (Chapter 3, 4), paradigms and models from applied cognitive psychology (Chapters 1, 2, 5, and 6), and materials relevant to language learning (all chapters, but especially Chapter 5 and 6). These multiple perspectives were combined in order to gain deeper insight into the benefits of retrieval practice for word learning and its potential use for vocabulary exercises. Interdisciplinary work is often called for to allow more evidence-based education (Beauchamp & Beauchamp, 2013) but a consequence of interdisciplinary work can be that from the point of view of the different sub-disciplines, studies have limitations. This is because compromises

are involved when different conceptual frames and terms are linked, experimental paradigms are combined, and materials and outcome measures are selected. Here, I offer suggestions for future research based on these three overarching issues. Specific methodological limitations of the studies presented in the thesis have been discussed in the previous chapters.

One challenge for the work discussed in this thesis was to link neuroimaging results to the behavioral literature. This is not trivial for several reasons, including differences in terminology and paradigms (see Chapter 4). Here, I focus on the interpretation of fMRI data in this thesis. It is important to be aware that the reported studies rely substantially on the mapping of cognitive functions on activations in specific brain areas, and vice versa. Such mappings are known as forward inference (predicting activations in brain area B given the involvement of process X) and reverse inference (predicting the involvement of process X given activation in brain area B) (Poldrack, 2006). Reverse inference is common in the literature, and was important to relate differences in brain activation between retrieval and restudying to the cognitive literature, in Chapters 3 and 4. However, most available neuroimaging studies, including meta-analyses, were designed to identify neural correlates of experimental manipulations (forward inference) and not to interpret observed activity in terms of cognitive processes (reverse inference) (e.g., Moran & Zaki, 2013; Poldrack, 2006, 2012). Using these data for reverse inferences is problematic because many brain regions are activated by different tasks (see, for example, Humphreys & Lambon Ralph, 2015 on IPL activation), making it impossible to definitely relate observed activation to one specific cognitive process (Poldrack, 2012; Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011). Fortunately, more information is increasingly available on the specificity of mappings between neural and cognitive functions, for example, from databases that are made using text-mining and machine-learning techniques to summarize published imaging studies (e.g., neurosynth, Yarkoni et al., 2011). Such initiatives might allow more rigorous inferences in the future, taking into account how selectively patterns of activation are associated with different cognitive processes (e.g., Poldrack, 2012). At present, it is important to keep in mind that the reported neuroimaging data are not definite evidence for the involvement of specific cognitive mechanisms. However, they are an addition to the available behavioral data and a possible source of new hypotheses. Whereas available behavioral experiments test hypotheses on semantic processing during retrieval by measuring *later* behaviour (e.g., Carpenter, 2011; Lehman & Karpicke, 2016) neuroimaging results provide an online measure of neural processes involved *during* practice. This makes the imaging data an interesting additional source of information about the cognitive processes involved in retrieval practice.

A second limitation of the studies reported here is that some of the experimental paradigms that we derived from prior testing effect studies differ from realistic learning situations. Based on the findings in Chapter 5 and 6, I recommended not using hints to support retrieval because learners might fail to transfer what they practice with the support of hints to later recall situations without such support. This interpretation fits the distinction made in the literature between current performance during practice and later (Bjork, 1994; Yan, Clark, & Bjork, 2016, as discussed in Section 7.3.2). Yet, this recommendation also draws attention to a fundamental issue in the use of retrieval practice: retrieval practice is only possible when learners have already encoded the materials sufficiently to retrieve them from memory. It is not clear how (initial) encoding should best be combined with retrieval, and how learners can best be supported to make the step from initial encoding to successful retrieval practice. In the studies in Chapter 2, 3, and 6, I circumvented this issue by using extensive pre-training to ensure high success during subsequent retrieval practice without feedback, and focused analyses on the retrieval phase. This paradigm isolated the direct benefits of retrieval practice from indirect benefits that occur when retrieval enhances subsequent feedback processing (e.g., Karpicke et al., 2014), which was useful for the study of underlying cognitive mechanisms. However, from a more applied perspective, this design limits the generalizability of the findings. In reality, learners do not first go through an extensive encoding procedure before engaging in retrieval practice. It is more likely, and probably more effective, if learners include retrieval already earlier during learning.

In Chapter 5, such a learning situation without pre-training was investigated. We used a scheduling algorithm (Sense et al., 2016) which presented each word once for encoding and then soon afterwards for retrieval practice, thus incorporating retrieval already early during learning. Then, over the course of practice, the time until the next repetition of a word was gradually increased to adjust the difficulty of retrieval to the learners' increasing word knowledge. Such adaptive technology that ensures high retrieval success is effective (Lindsey, Shroyer, Pashler, & Mozer, 2014; Sense et al., 2016) but not available for every task, and not every retrieval exercise can take place at a computer. Therefore, the question how to make the transition from initial encoding to retrieval practice remains relevant: For instance, is it better to use difficult retrieval with feedback, as Chapter 5 suggested, or are there situations in which it is more beneficial to make the first retrieval attempts easier, for example, with hints (see also Finley et al., 2011)? The experiments in Chapter 6, for example, showed that triggering successful retrieval with weak contextual information is beneficial compared to practicing words with rich contextual information. However, the context was manipulated after a pre-training and it is unclear if it is also more beneficial to present words in an uninformative context at earlier stages of learning. Alternatively,

learners might remember more if words are first presented in an informative context that facilitates retrieval (or inferences), and later-on retrieve word meaning in an uninformative context. Finally, if retrieval is initially facilitated, an open question is how this should occur. For example, gradually increasing spacing may be preferable over the use of prompts because spacing changes the difficulty but not the nature of the retrieval task, whereas prompts may influence which specific associations are strengthened.

A last limitation is the small range of outcome measures used in this thesis. Building up a form-meaning association is a crucial part of vocabulary learning, but it is by far not the only aspect of it (Nation, 2001). Moreover, we focused on receptive word knowledge in most chapters, using tests on which learners translated foreign vocabularies into their native language. It is not clear if the investigated manipulations would affect recall of the foreign word form in the same way. Recall of a new word form is more difficult to acquire, as it involves the "formation of complete orthographic representations of the new L2 words, whereas L2-to-L1 learning requires only discriminable, but not necessarily complete, representations of the new L2 words" (Schneider, Healy, & Bourne, 2002, p. 420). The results in Chapter 6 are promising, however, in that they showed that retrieval of the word meaning during practice enhanced later recall of the word meaning but also of the word form as well as the recognition of words in context. Thus, some transfer occurs from practicing the retrieval of word meaning to other aspects of word knowledge (see also Carpenter, Pashler, & Vul, 2006). It would be relevant to investigate whether retrieval practice enhances also more complex aspects of word learning, such as the ability to spontaneously use a word in conversation. Possibly, a controlled adjustment of retrieval practice - for example, the variety of retrieval cues and the semantic associations activated - might enhance benefits of retrieval practice for these aspects of word learning.

7.5 CONCLUSION

The retrieval of information from memory is not a simple readout process; each retrieval act increases the accessibility of the retrieved information and suppresses competing information. This makes retrieval practice a promising technique for learners who want to remember large amounts of information, for example, when practicing vocabulary in a foreign language. This thesis presents converging evidence from behavioral and neuroimaging studies that an important mechanism underlying the benefits of retrieval practice is the selective strengthening of form-meaning associations. A practical consequence of the selectiveness of this process is that retrieval practice that includes prompts does not necessarily enhance later recall without prompts. Benefits

of retrieval have mostly been studied with explicit recall tasks, but they can also be triggered more indirectly. For example, retention is enhanced when the context in which words appear stimulates retrieval of word meaning from memory. These findings lead to a number of practical recommendations for the design of learning situations, which show how research into the basic architecture of human memory can be used to realize effective practice.

7.6 REFERENCES

Anderson, J. R. (1981). Interference: The relationship between response latency and response accuracy. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 326–343. https://doi.org/10.1037/0278-7393.7.5.326

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063–1087. https://doi.org/10.1037/0278-7393.20.5.1063

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

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, *57*(1), 35–56. https://doi.org/10.1111/j.1467-9922.2007.00398.x

Barcroft, J. (2015). Can retrieval opportunities increase vocabulary learning during reading? *Foreign Language Annals*, *48*(2), 236–249. https://doi.org/10.1111/flan.12139

Beauchamp, C., & Beauchamp, M. H. (2013). Boundary as bridge: An analysis of the educational neuroscience literature from a boundary perspective. *Educational Psychology Review*, *25*(1), 47–67. https://doi.org/10.1007/s10648-012-9207-x

Beck, I. L., McKeown, M. G., & McCaslin, E. S. (1983). Vocabulary development: All contexts are not created equal. *The Elementary School Journal*, *83*(3), 177–181. https://doi.org/10.1086/461307

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–536. https://doi.org/10.1016/j.tics.2011.10.001

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796. https://doi.org/10.1093/cercor/bhp055

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

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

Blumenfeld, R. S., & Ranganath, C. (2007). Prefrontal cortex and long-term memory encoding: An integrative review of findings from neuropsychology and neuroimaging. *The Neuroscientist*, *13*(3), 280–291. https://doi.org/10.1177/1073858407299290

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604–616. https://doi.org/10.3758/MC.36.3.604

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*(6), 1563–1569. https://doi.org/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*(6), 1547–1552. https://doi.org/10.1037/a0024140

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*(2), 268–276. https://doi.org/10.3758/BF03193405

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, *13*(5), 826–830. https://doi.org/10.3758/BF03194004

Carpenter, S. K., & Yeung, K. L. (2017). The role of mediator strength in learning from retrieval. *Journal of Memory and Language*, *92*, 128–141. https://doi.org/10.1016/j.jml.2016.06.008

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–642. https://doi.org/10.3758/BF03202713

Crutcher, R. J., & Ericsson, K. A. (2000). The role of mediators in memory retrieval as a function of practice: Controlled mediation to direct access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1297–1317. https://doi.org/10.1037//0278-7393.26.5.1297

Danker, J. F., Fincham, J. M., & Anderson, J. R. (2011). The neural correlates of competition during memory retrieval are modulated by attention to the cues. *Neuropsychologia*, *49*(9), 2427–2438. https://doi.org/10.1016/j.neuropsychologia.2011.04.020

Dudai, Y. (2012). The restless engram: Consolidations never end. *Annual Review of Neuroscience*, *35*(1), 227–247. https://doi.org/10.1146/annurev-neuro-062111-150500

Eriksson, J., Kalpouzos, G., & Nyberg, L. (2011). Rewiring the brain with repeated retrieval: A parametric fMRI study of the testing effect. *Neuroscience Letters*, *505*(1), 36–40. https://doi.org/10.1016/j.neulet.2011.08.061

Finley, J. R., Benjamin, A. S., Hays, M. J., Bjork, R. A., & Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language*, *64*(4), 289–298. https://doi.org/10.1016/j.jml.2011.01.006

Finn, B., & Metcalfe, J. (2010). Scaffolding feedback to maximize long-term error correction. *Memory & Cognition*, *38*(7), 951–961. https://doi.org/10.3758/MC.38.7.951

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*(4), 801–812. https://doi.org/10.1037/a0023219

Hall, K. A., Adams, M., & Tardibuono, J. (1968). Gradient- and full-response feedback in computer assisted instruction. *Journal of Educational Research*, *61*(5), 195–199.

Hulstijn, J. H. (1992). Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 113–125). London: Maxmillan.

Humphreys, G. F., & Lambon Ralph, M. A. (2015). Fusion and fission of cognitive functions in the human parietal cortex. *Cerebral Cortex*, *25*(10), 3547–3560. https://doi.org/10.1093/cercor/bhu198

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability. *The Quaterly Journal of Experimental Psychology*, *65*(5), 962–975. https://doi.org/10.1080/17470218.2011.638079

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4–5), 528–558. https://doi.org/10.1080/09541440601056620

Kang, S. H. K., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K., & Mozer, M. C. (2011). Does incorrect guessing impair fact learning? *Journal of Educational Psychology*, *103*(1), 48–59. https://doi.org/10.1037/a0021977

Karlsson-Wirebring, L., Wiklund-Hörnqvist, C., Eriksson, J., Andersson, M., Jonsson, B., & Nyberg, L. (2015). Lesser Neural Pattern Similarity across Repeated Tests Is Associated with Better Long-Term Memory Retention. *Journal of Neuroscience*, *35*(26), 9595–9602. https://doi.org/10.1523/JNEUROSCI.3550-14.2015

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning. *Psychology of Learning and Motivation*, *61*, 237–284. https://doi.org/10.1016/B978-0-12-800283-4.00007-1

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

Kole, J. A., & Healy, A. F. (2013). Is retrieval mediated after repeated testing? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(2), 462–472. https://doi.org/10.1037/a0028880

Koriat, A., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition*, *34*(5), 959–972. https://doi.org/10.3758/BF03193244

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. https://doi.org/10.1016/j.jml.2011.04.002

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 283–294. https://doi.org/10.1037/a0037850

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. *Psychology of Learning and Motivation*, *65*, 183–215. https://doi.org/10.1016/bs.plm.2016.03.003

Kuhl, B. A., Dudukovic, N. M., Kahn, I., & Wagner, A. D. (2007). Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nature Neuroscience*, *10*(7), 908–914. https://doi.org/10.1038/nn1918

Lehman, M., & Karpicke, J. D. (2016). Elaborative retrieval: Do semantic mediators improve memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(10), 1573–1591. https://doi.org/10.1037/xlm0000267

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1787–1794. https://doi.org/10.1037/xlm0000012

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological Science*, *25*(3), 639–647. https://doi.org/10.1177/0956797613504302

MacLeod, C. M., & Nelson, T. O. (1984). Response latency and response accuracy as measures of memory. *Acta Psychologica*, *57*(3), 215–235. https://doi.org/doi: 10.1016/0001-6918(84)90032-5

Mondria, J.-A., & Wit-de Boer, M. (1991). The effects of contextual richness on the guessability and the retention of words in a foreign language. *Applied Linguistics*, *12*(3), 249–267. https://doi.org/10.1093/applin/12.3.249

Moran, J. M., & Zaki, J. (2013). Functional neuroimaging and psychology: What have you done for me lately? *Journal of Cognitive Neuroscience*, *25*(6), 834–842. https://doi.org/10.1162/jocn_a_00380

Nairne, J. S. (2002). The myth of the encoding-retrieval match. *Memory*, *10*(5–6), 389–395. https://doi.org/10.1080/09658210244000216

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press. Retrieved from /core/books/learning-vocabulary-in-another-language/491314AA1B451AD04F3536000F1C9F0D

Noonan, K. A., Jefferies, E., Visser, M., & Lambon Ralph, M. A. (2013). Going beyond inferior prefrontal involvement in semantic control: evidence for the additional contribution of dorsal angular gyrus and posterior middle temporal cortex. *Journal of Cognitive Neuroscience*, *25*(11), 1824–1850. https://doi.org/10.1162/jocn_a_00442

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59–63. https://doi.org/10.1016/j.tics.2005.12.004

Poldrack, R. A. (2012). The future of fMRI in cognitive neuroscience. *NeuroImage*, *62*(2), 1216–1220. https://doi.org/10.1016/j.neuroimage.2011.08.007

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*(2), 644–667. https://doi.org/10.1037/a0033194

Pressley, M., Levin, J. R., & McDaniel, M. A. (1987). Remembering versus inferring what a word means: Mnemonic and contextual approaches. In M. G. McKeown & M. E. Curtis (Eds.), *The Nature of Vocabulary Acquisition* (pp. 107–127). Hillsdale, NJ: Lawrence Erlbaum.

Price, A. R., Bonner, M. F., Peelle, J. E., & Grossman, M. (2015). Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *Journal of Neuroscience*, *35*(7), 3276–3284. https://doi.org/10.1523/JNEUROSCI.3446-14.2015

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*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*(6002), 335. https://doi.org/10.1126/science.1191465

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55. https://doi.org/10.1038/nrn.2016.150

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory & Cognition*, *43*(4), 619–633. https://doi.org/10.3758/s13421-014-0477-z

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Rowland, C. A., & DeLosh, E. L. (2015). Mnemonic benefits of retrieval practice at short retention intervals. *Memory*, *23*(3), 403–419. https://doi.org/10.1080/09658211.2014.889710

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland and Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching*, *50*(2), 212–226. https://doi.org/10.1017/S0261444815000075

Schneider, V. I., Healy, A. F., & Bourne, L. E. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, *46*(2), 419–440. https://doi.org/10.1006/jmla.2001.2813

Schouten-van Parreren, C. A. (1989). Vocabulary learning through reading: Which conditions should be met when presenting words in texts. *AILA Review*, *6*(1), 75–85.

Seibert, L. C. (1945). A study on the practice of guessing word meanings from a context. *The Modern Language Journal*, *29*(4), 296–322. https://doi.org/10.2307/318219

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An individual's rate of forgetting is stable over time but differs across materials. *Topics in Cognitive Science*, *8*(1), 305–321. https://doi.org/10.1111/tops.12183

Simons, D. J., & Chabris, C. F. (2011). What people believe about how memory works: A representative survey of the U.S. population. *PLoS ONE*, *6*(8), e22757. https://doi.org/10.1371/journal.pone.0022757

Smith, S. M., & Handy, J. D. (2016). The crutch of context-dependency: Effects of contextual support and constancy on acquisition and retention. *Memory*, *24*(8), 1134–1141. https://doi.org/10.1080/09658211.2015.1071852

Storm, B. C., & Levy, B. J. (2012). A progress report on the inhibitory account of retrieval-induced forgetting. *Memory & Cognition*, *40*(6), 827–843. https://doi.org/10.3758/s13421-012-0211-7

Takashima, A., & Bakker, I. (in press). Memory consolidation. In H.-J. Schmid (Ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge* (pp. 177–200). Boston, MA, US: De Gruyter Mouton. https://doi.org/10.1037/15969-009

Thomas, R. C., & McDaniel, M. A. (2013). Testing and feedback effects on front-end control over later retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 437–450. https://doi.org/10.1037/a0028886

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, *56*(4), 252–257. https://doi.org/10.1027/1618-3169.56.4.252

van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, *85*(4), 475–511. https://doi.org/10.3102/0034654314564881

Vannest, J., Eaton, K. P., Henkel, D., Siegel, M., Tsevat, R. K., Allendorfer, J. B., .. Szaflarski, J. P. (2012). Cortical correlates of self-generation in verbal paired associate learning. *Brain Research*, *1437*, 104–114. https://doi.org/10.1016/j.brainres.2011.12.020

Veltre, M. T., Cho, K. W., & Neely, J. H. (2015). Transfer-appropriate processing in the testing effect. *Memory*, *23*(8), 1229–1237. https://doi.org/10.1080/09658211.2014.970196

Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*, *43*(7), 1036–1046. https://doi.org/10.1037/xlm0000379

Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., & Anderson, M. C. (2015). Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nature Neuroscience*, *18*(4), 582–589. https://doi.org/10.1038/nn.3973

Wimber, M., Bäuml, K.-H., Bergström, Z., Markopoulos, G., Heinze, H.-J., & Richardson-Klavehn, A. (2008). Neural markers of inhibition in human memory retrieval. *Journal of Neuroscience*, *28*(50), 13419–13427. https://doi.org/10.1523/jneurosci.1916-08.2008

Wimber, M., Rutschmann, R. M., Greenlee, M. W., & Bäuml, K.-H. (2009). Retrieval from episodic memory: Neural mechanisms of interference resolution. *Journal of Cognitive Neuroscience*, *21*(3), 538–549. https://doi.org/10.1162/jocn.2009.21043

Wing, E. A., Marsh, E. J., & Cabeza, R. (2013). Neural correlates of retrieval-based memory enhancement: an fMRI study of the testing effect. *Neuropsychologia*, *51*(12), 2360–2370.

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, *145*(7), 918–933. https://doi.org/10.1037/xge0000177

Yan, V. X., Clark, C. M., & Bjork, R. A. (2016). Memory and metamemory considerations in the instruction of human beings revisited: Implications for optimizing online learning. In J. C. Horvath, J. M. Lodge, & J. Hattie (Eds.), *From the laboratory to the classroom: Translating science of learning for teachers*. London, UK: Routledge.

Yan, V. X., Yu, Y., Garcia, M. A., & Bjork, R. A. (2014). Why does guessing incorrectly enhance, rather than impair, retention? *Memory & Cognition*, *42*(8), 1373–1383. https://doi.org/10.3758/s13421-014-0454-6

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. https://doi.org/10.1038/nmeth.1635

APPENDIX

SUMMARY[1]

Many people incorrectly assume that human memory works like a video camera, which records, stores, and later replays fixed memories. In fact, memory is a more complex, dynamic system. Memory retrieval, in particular, is not a simple replay process. Instead, each retrieval can change the content and accessibility of memories.

For learners, it is particularly interesting that **memory retrieval can be practiced:** When learners retrieve information from memory, it becomes easier to later retrieve that information again. This is of great interest for anyone who needs to remember large amounts of information, such as language learners who memorize hundreds of words when mastering a new language. For them, retrieval practice is a promising technique to enhance their ability to later recall the practiced words. Retrieval practice is also more effective than other learning techniques during which the complete information is presented, for example, repeatedly reading words with translations. The benefits of *retrieval practice* compared to other forms of *restudying* are known in the literature as the **testing effect.**

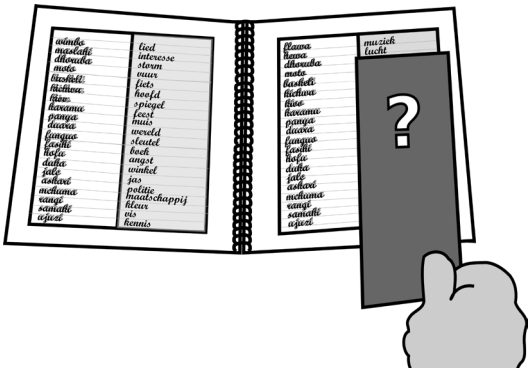


Figure A.1 A learner practices the retrieval of words from memory, thereby making it easier to later recall the words again. Such self-testing or *retrieval practice* leads to better long-term outcomes than other forms of *restudying*, like repeated reading. This phenomenon is called the *testing effect.*

1 This is a summary for a broad audience. A more comprehensive summary is given in the general discussion in Chapter 7.

The first part of **this dissertation** focuses on the cognitive and neural underpinnings of the testing effect**.** In spite of a wealth of research showing testing effects, the underlying mechanisms are poorly understood. I measured reaction times and brain activation during retrieval and restudying, to explain why retrieval practice is beneficial for the long-term retention of words. The second part of the dissertation focuses on the integration of memory retrieval in vocabulary exercises. This part first reports several classroom experiments in which I tested the effect of retrieval prompts with high school students who practiced vocabulary words with an adaptive computer program. Then, I discuss the effect of contextual information during retrieval on word learning. Finally, the dissertation – and this summary - conclude with practical recommendations on how to create opportunities for retrieval during vocabulary learning.

PART I. WHY RETRIEVAL PRACTICE ENHANCES WORD RETENTION

REACTION TIMES AFTER RETRIEVAL AND RESTUDY

In Chapters 2 and 3, I showed that learners not only recall *more* words after retrieval practice (compared to restudy practice), but also recall the words more *quickly*. Thus, there is a testing effect not only on the accuracy of later recall but also on recall speed. This **faster recall speed suggests that words become more accessible in memory with retrieval practice,** which supports theoretical accounts from the literature which hold that retrieval becomes more efficient with practice. Possibly, the mental search for vocabulary items becomes easier every time a word is retrieved from memory. A requirement for such effects is that the retrieval is successful: In several experiments reported in the dissertation, **later recall only benefited from retrieval practice, when participants managed to retrieve words from memory during practice.**

DIFFERENCES IN BRAIN ACTIVITY DURING RETRIEVAL AND RESTUDY

Chapters 3 and 4 present neuroimaging data that was collected while participants practiced vocabulary words in an MR scanner (see Figure A2). Differences in brain activation during retrieval practice and restudy practice were measured with functional magnetic resonance imaging (fMRI), a technique that picks up changes in regional blood flow due to the magnetic properties of oxygen in the blood. Comparing brain activations during retrieval and restudy practice, there were two main findings:

1. Retrieval likely involves more mental effort than restudying, which is reflected in higher activations in frontal brain areas. The effort is reduced with practice, and likewise frontal activations decrease with practice.

Areas in the front of the brain, specifically in the ventrolateral prefrontal cortex (VLPFC), showed higher activation during memory retrieval than during restudying. These brain areas are often involved in tasks that require cognitive control, that is, concentration and an intentional focus of attention. Cognitive control enables, for example, the selection of relevant information among distracting, irrelevant input. **Higher VLPFC activation during retrieval than during restudying likely reflects stronger demands on cognitive control** during retrieval. Controlled, effortful practice often leads to better long-term learning outcomes than less effortful practice, and more effort during retrieval than restudying could therefore be an explanation for testing effects. In addition, over the course of repeated retrieval practice and after prior retrieval practice compared to prior restudying, activations in VLPFC decreased. Together, these findings suggest that **retrieval leads to more effortful, controlled processing than restudying but it becomes easier with practice.**

2. Activations in brain areas important for semantic (meaningful) processing suggest that retrieval practice might lead to a focus of attention on relevant information and strengthen associations that make later recall easier.

During retrieval practice, activations in inferior parietal and middle temporal brain areas were higher when participants practiced words that they later *remembered* than when participants practiced words that they later *forgot*. During restudying, there was no such difference. I tentatively explain these findings with differences in the quality of semantic processing during retrieval and restudying, because the areas involved are thought to play an important role in semantic processing. Retrieval practice might direct attention more to relevant semantic associations than restudying does, leading to a stronger relation between brain activation and later performance. **A focus on relevant associations during retrieval might lead to a stronger memory representation that can be better recalled later-on.** In contrast, although restudying seemed to involve more semantic processing overall than retrieval practice, this processing did not predict better later recall. This may be because more irrelevant information was activated during restudying, possibly due to mind-wandering and lapses of attention.

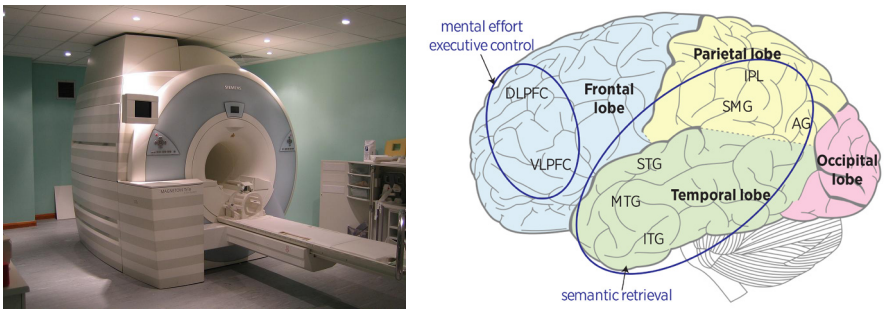


Figure A.2 Left: Photo of an MR scanner. In Chapters 3 and 4, results are presented from neuroimaging studies in which adults practiced vocabulary words through retrieval practice or restudying in an MR scanner. (Photo: Image Editor, 2006.). The activation of different parts of the brain was measured using functional magnetic resonance imaging, a technique that is based on the coupling between neural activation and regional blood flow. Right: The second illustration shows a view on the left side of the brain (Carter, 1918). The most important brain areas that are discussed in Chapters 3 and 4 are highlighted. See Figure 4.2 for further information.

PART II. HOW TO USE RETRIEVAL PRACTICE IN VOCABULARY EXERCISES

FEEDBACK AND HINTS DURING RETRIEVAL PRACTICE

In the second part of the dissertation, I investigate retrieval during vocabulary exercises. In Chapter 5, I describe three experiments with high school students who practiced vocabulary words at the computer, using retrieval practice. Different feedback formats are compared to test if it is beneficial to provide hints during practice.

Feedback increases the benefits of retrieval practice, because it allows learners to correct their errors and makes them more confident in their responses. In contrast, if a learner cannot retrieve a word from memory and there is no feedback, the retrieval attempt has few benefits. Most studies on retrieval practice contain simple feedback in which the correct answer is shown. In Chapter 5, I test whether feedback becomes more effective if learners first receive hints that allow them to correct their answer (see Figure A3).

The experiments reported in Chapter 5 produced no evidence that hints during retrieval practice are beneficial for learning. The hints did not reduce (repeated) errors during practice and there was also no positive effect on a later vocabulary test. On the contrary, because it took students time to process the hints, less time remained for further repetitions and students practiced fewer words overall. Moreover, benefits of hints were only found when the hints from practice were also available on the

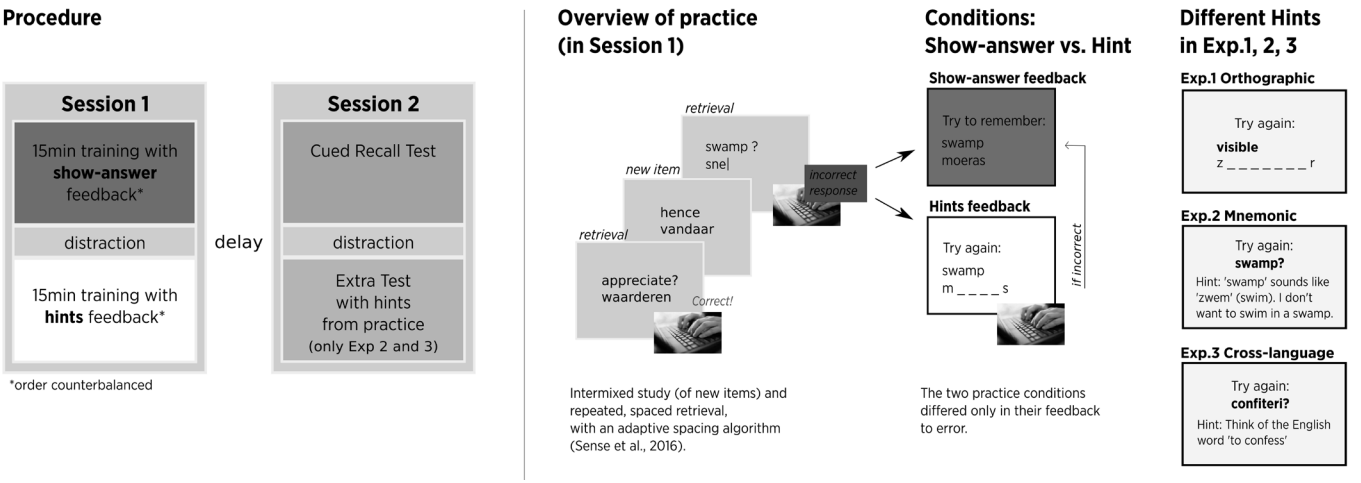


Figure A.3 In the experiments described in Chapter 5, high school students practiced English or Latin vocabulary items with different versions of an adaptive computer program. The program timed the repetitions of vocabulary items based on typical forgetting rates and increased the delays between repetitions of the same word over the course of practice. If students made an error during practice, they received either standard *show-answer feedback* or *hints feedback* that consisted of different types of hints in the three experiments. The hints did not enhance performance during practice or on a later vocabulary test, unless that test contained the same hints from practice. In addition, some hints reduced the time available for further practice trials.

subsequent vocabulary test. This suggests that, during practice with hints, word knowledge became partly dependent on the availability of hints. Thus, **feedback with hints had no benefits for recall on a later test without hints.**

CONTEXTUAL INFORMATION AND RETRIEVAL PRACTICE

In a different study on the effect of retrieval during vocabulary exercises, reported in Chapter 6, I found that new words were better remembered when participants practiced retrieving the words from memory, than when they inferred word meaning from a relevant, informative context. Participants studied new words (e.g., the word *funguo*) and then practiced the words either in an uninformative sentence ("I need the *funguo*.") or in an informative sentence ("I want to unlock the door. I need the *funguo*."). Participants had to retrieve the word meaning from memory when practicing with the uninformative sentences, but could infer word meaning when practicing with

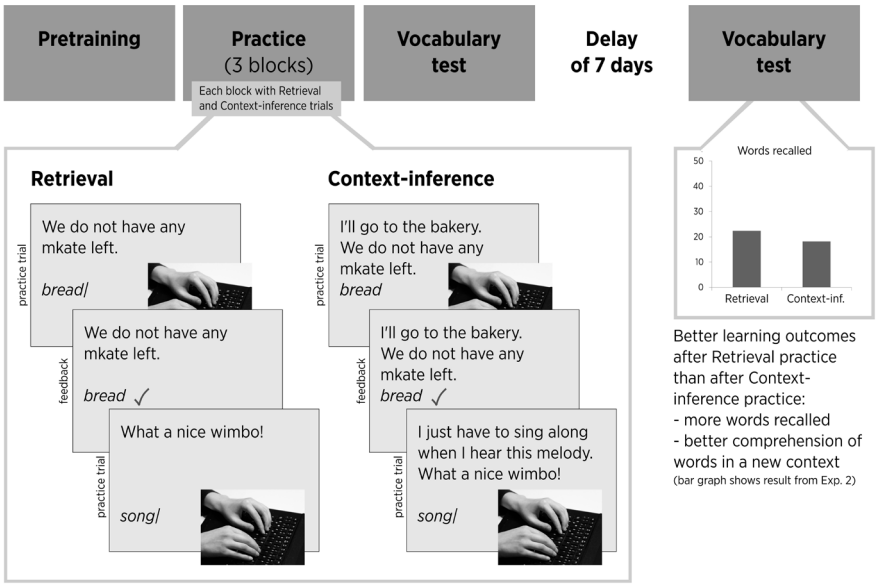


Figure A4 In the experiments reported in Chapter 6, participants studied new words (e.g., the word wimbo = song) and then further practiced the words either in an uninformative Retrieval sentence ("What a nice *wimbo*!") or in an informative context-inference sentence ("I just have to sing along when I hear this melody. What a nice *wimbo*!"). Participants had to retrieve the word meaning from memory when practicing with the uninformative sentences but could infer word meaning from the informative sentences. Practice with Retrieval sentences led to significantly higher performance on vocabulary tests, both immediately and several days after learning. The graph shows one of several significant differences found in Experiment 2.

the informative sentences (here: *funguo* = key). The informative sentences thus made it easier to understand words during practice. However, a vocabulary test several days later showed that participants remembered words better after retrieval practice with uninformative sentences.

The finding that the uninformative context led to more word learning than the informative context is counter-intuitive, because contextual information is typically a beneficial source of information that learners can use to understand new words. However, **when a learner understands a word in context, that does not automatically mean that the learner also remembers the word over time**. In line with this, Chapter 6 showed that a manipulation that made it easier to understand the target words (namely: adding contextual information during practice), made it *less* likely that the words were remembered over time. The uninformative context, which required effortful retrieval of the word from memory, led to better long-term retention than the informative context, which allowed learners to infer the word meaning from context. These experiments demonstrate that the **benefits of retrieval can be evoked through the context in which a word appears. Reducing contextual information to trigger memory retrieval can enhance learning.**

CONCLUSIONS AND PRACTICAL RECOMMENDATIONS

The retrieval of information from memory is not a simple readout process; each retrieval act increases the accessibility of the retrieved information. This makes retrieval practice a powerful technique to remember information over time.

This dissertation presents converging evidence from behavioral and neuroimaging studies regarding the cognitive mechanisms that underlie the benefits of retrieval practice for vocabulary learning. It suggests that retrieval is an effortful process that becomes facilitated with practice. This facilitation may be due to a selective focus on information necessary to achieve the retrieval, such as the association between novel word form and meaning in vocabulary learning. This may strengthen the memory representation in such a way that it can be reactivated more easily later on.

The studies reported in the thesis also highlight a number of requirements for the successful integration of retrieval practice during vocabulary exercises, from which practical recommendations can be derived for the design of learning situations.

Practical recommendations

- **The retrieval of information from memory, for example during self-testing, is a powerful learning technique.** Therefore, do not only present key information to the learners but let them produce target information from memory during practice.
- **Combine retrieval with feedback.** This allows learners to encode information that they cannot yet retrieve from memory. Unsuccessful retrieval has few, if any, benefits without feedback.
- **Provide repeated retrieval opportunities**. This way, later retrieval becomes increasingly facilitated. Repetition also ensures that there is a new chance for a successful retrieval after a prior error.
- **Aim for retrieval practice that fits the learners' abilities. Ideally, practice should be successful but challenging.** Conditions that increase the ease of retrieval too much may lead to high performance and confidence during practice, but to worse retention over time.
- **Ensure that the conditions of retrieval practice fit the learning goal.** For example, practice with hints may not enhance later recall without hints. Similarly, a multiple-choice test may not enhance later free recall. Therefore, be conservative with support during practice and consider whether the support is also available later-on.
- **Distinguish between fluency of retrieval during practice and benefits of practice for later performance.** Fluency during practice is not a good predictor of later performance, if the practice situation facilitates the retrieval in an undesirable way (e.g., through massed repetition).
- **Integrate opportunities for retrieval in different learning situations.** Small manipulations of exercises can suffice to trigger retrieval, for example, by first presenting a word in an uninformative context and only later in an informative context that reveals word meaning.

IMAGE CREDITS

Figure A2:

Image Editor. (2006). *01 Siemens MAGNETOM Trio* [Digital image]. Retrieved from Flickr: https://www.flickr.com/photos/11304375@N07/3081315619/sizes/o/. Creative Commons Attribution License.

Carter, H.V. (1918). Principal fissures and lobes of the cerebrum viewed laterally. Figure 728 from H. Gray (Ed.), *Anatomy of the human body* (20th edition, revised by Warren H. Lewis). Philadelphia: Lea & Febiger. Image edited by O. Räisänen for wikimedia [Public domain]. Retrieved from https://commons.wikimedia.org/wiki/File:Lobes_of_the_brain_NL.svg

SAMENVATTING[1]

Veel mensen denken dat het menselijk geheugen werkt als een camera die herinneringen opneemt, opslaat en later weer afspeelt. Deze vergelijking is echter niet correct: het menselijke geheugen is in werkelijkheid een systeem dat beduidend complexer en dynamischer is. Een belangrijke eigenschap van het geheugen is dat het oproepen van een herinnering de inhoud en de latere toegankelijkheid van de herinnering beïnvloedt.

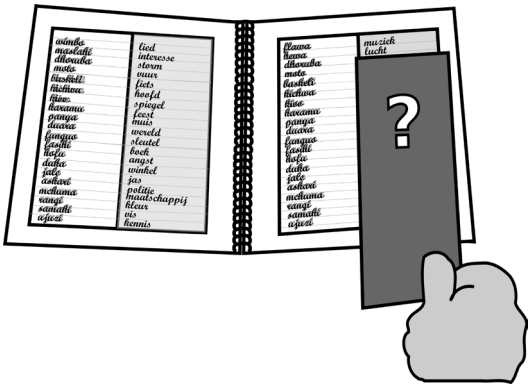
Voor leerlingen is het relevant dat **het oproepen van informatie uit het geheugen geoefend kan worden:** elke oproep van een herinnering uit het geheugen (in het vervolg *retrieval* genoemd) maakt het makkelijker om dezelfde herinnering later opnieuw op te roepen. Dit is bijzonder interessant voor iedereen die een grote hoeveelheid informatie wil onthouden, bijvoorbeeld om een woordenschat in een vreemde taal op te bouwen (zie figuur A.5). Het oefenen van retrieval (*retrieval practice*) is een effectieve techniek om woorden op lange termijn te onthouden. Retrieval practice is daarnaast duidelijk effectiever dan oefenmethoden waarbij volledige informatie bestudeerd wordt, bijvoorbeeld het herhaaldelijk doorlezen van nieuwe woorden samen met de vertaling. Deze voordelen van retrieval practice vergeleken met herbestuderen[2] worden in de literatuur **testeffect** genoemd.

Het eerste gedeelte van **dit proefschrift** behandelt de cognitieve en neurale mechanismen die ten grondslag liggen aan testeffecten. Hoewel veel wetenschappelijke studies testeffecten hebben aangetoond, is tot nu toe weinig bekend over de onderliggende denkprocessen. Ik beschrijf reactietijden en hersenactiviteit gemeten tijdens retrieval en restudy om te verklaren waarom retrieval practice ervoor zorgt dat woorden op lange termijn beter onthouden worden.

Het tweede gedeelte van dit proefschrift is gericht op de integratie van retrieval in oefeningen voor woordenschatverwerving. Hoofdstuk 5 bevat een rapportage van drie experimenten waarin ik heb onderzocht welk effect verschillende soorten feedback en tips tijdens retrieval practice hebben op het woordleren van middelbare scholieren. Vervolgens bespreek ik in Hoofdstuk 6 het effect van contextuele

1 Deze samenvatting geeft een compact overzicht van de belangrijkste bevindingen van het proefschrift. De samenvatting is bedoeld voor een breed publiek. Een uitgebreidere samenvatting in het Engels is opgenomen in hoofdstuk 7, sectie 7.1.

2 Het oefenen van het reproduceren van informatie uit het geheugen (*retrieval*) wordt in deze samenvatting *retrieval practice* genoemd (van het Engelse woord *memory retrieval)*. De leerresultaten na retrieval practice worden vaak vergeleken met een controleconditie van herhaaldelijk doorlezen; deze wordt *herbestuderen* of *restudy* genoemd.



Figuur A.5 Een leerling vertaalt woorden uit het geheugen. Dit maakt het later herinneren van de woorden makkelijker. Het herhaaldelijk oproepen van informatie uit het geheugen (*retrieval practice*), bijvoorbeeld tijdens het overhoren, leidt tot betere langetermijnresultaten dan andere oefenmethoden zoals het herhaaldelijk doorlezen van informatie (*restudy*). Dit fenomeen wordt in de literatuur *testeffect* genoemd.

informatie tijdens retrieval op het woordleren aan de hand van drie experimenten met volwassen proefpersonen. Het proefschrift en deze samenvatting sluiten af met praktische aanbevelingen voor het gebruik van retrieval tijdens het woordleren.

DEEL I. *WAAROM* ZORGT RETRIEVAL PRACTICE ERVOOR DAT WOORDEN BETER WORDEN ONTHOUDEN?

REACTIETIJDMETINGEN

In Hoofdstuk 2 en 3 laat ik zien dat mensen die retrieval practice toepassen later niet alleen *meer* woorden uit het geheugen kunnen reproduceren (vergeleken met herhaling door restudy), maar deze woorden ook *sneller* reproduceren. Dit betekent dat een testeffect niet alleen de kans verhoogt op een correcte reproductie maar ook de snelheid van de reproductie. Ditbevestigt hypothesen uit de literatuur die stellen dat **retrieval door oefenen efficiënter** wordt. Mogelijk gaat elke keer wanneer een woord succesvol uit het geheugen gereproduceerd wordt, het vinden van dat woord in het geheugen makkelijker en sneller. De studies in dit proefschrift laten verder zien dat een voorwaarde voor het testeffect is, dat de retrieval tijdens het oefenen succesvol is. **Alleen wanneer de proefpersonen de woorden tijdens het oefenen uit het geheugen konden reproduceren, was retrieval practice voordelig voor het onthouden van woorden.**

METINGEN VAN HERSENACTIVITEIT

In hoofdstuk 3 en 4 worden neuro-imaging-data gepresenteerd die verzameld zijn terwijl proefpersonen in een scanner nieuwe woorden oefenden. Verschillen in hersenactiviteit tijdens retrieval en restudy practice werden met behulp van *fMRI* (functional magnetic resonance imaging) gemeten (zie figuur A.6). fMRI is gebaseerd op het verband tussen lokale neurale activiteit en veranderingen in de doorbloeding van de hersenen. De doorbloeding wordt hierbij gemeten op basis van de verschillende magnetische eigenschappen van zuurstofarm en zuurstofrijk bloed. De vergelijking van hersenactiviteit tijdens retrieval en restudy leidde tot twee hoofdresultaten:

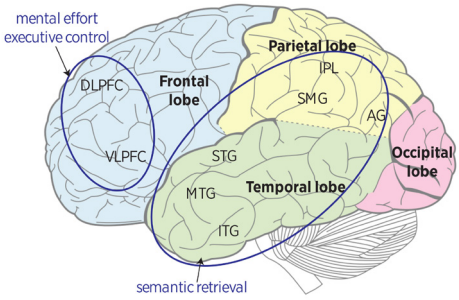
1. Retrieval vereist waarschijnlijk meer mentale inspanning dan restudy practice, wat gepaard gaat met hogere activatie in frontale hersengebieden. De inspanning neemt met herhaalde retrieval af evenals de frontale activatie.

Gebieden in het voorste gedeelte van het brein, in de ventrolaterale prefrontale cortex (kort: VLPFC, zie figuur A.6), vertoonden in de studies beschreven in hoofdstuk 3 en 4 een hogere activatie tijdens retrieval practice dan tijdens restudying. Deze gebieden zijn vaak actief wanneer taken mentale controle vereisen, dat wil zeggen concentratie en het bewust richten van aandacht. Dergelijke controle maakt het bijvoorbeeld mogelijk om relevante informatie te selecteren en afleidende, irrelevante informatie te negeren. **Hogere activatie in de VLPFC tijdens retrieval practice – vergeleken met restudying – duidt op een sterkere rol van mentale controle tijdens retrieval practice.** Over het algemeen leiden inspanning en gecontroleerde verwerking van informatie tot betere langetermijnresultaten dan verwerking zonder inspanning. Daarom zou de verhoogde inspanning tijdens retrieval een verklaring kunnen zijn voor testeffecten. Tegelijkertijd laten studies in hoofdstuk 4 zien dat tijdens herhaalde retrieval practice en na retrieval practice (vergeleken met restudy practice) de activatie van de VLPFC afneemt. Alles bij elkaar genomen suggereren deze resultaten dat **retrieval inspanning en gecontroleerde verwerking van informatie vereist maar door herhaaldelijk oefenen steeds minder moeite kost.**

2. Activatie in hersengebieden die belangrijk zijn voor semantische (betekenisvolle) verwerking duiden erop dat retrieval practice leidt tot een focus van aandacht op relevante informatie en hierdoor associaties versterkt die later het herinneren makkelijker maken.

Tijdens retrieval practice waren gebieden in de parietale en temporale cortex (inferior parietal lobe met SMG en AG, en middle temporal gyrus, zie figuur A.6) actiever wanneer de proefpersonen woorden oefenden die ze op een test na het oefenen nog bleken te kennen dan wanneer de proefpersonen woorden oefenden

die ze op de test vergeten bleken te zijn. De activiteit in deze hersengebieden was dus voorspellend voor het leerresultaat. Tijdens restudy was er daarentegen geen verschil tussen de onthouden en vergeten woorden; de hersenactiviteit tijdens het herhaaldelijk doorlezen was dus niet voorspellend voor het leerresultaat. Mogelijk komt dit door verschillen in de kwaliteit van de semantische (op betekenis gerichte) verwerking van de woorden tijdens retrieval en restudy, aangezien in de literatuur wordt aangenomen dat de betrokken hersengebieden een belangrijke rol spelen bij semantische verwerking. **Retrieval practice zou de aandacht tijdens het oefenen kunnen richten op relevante informatie die de retrieval mogelijk maakt,** zoals bijvoorbeeld ezelsbruggen die een link leggen tussen de spelling van nieuwe woorden en de woordbetekenis. Daarentegen was semantische verwerking tijdens restudy niet voorspellend voor het leerresultaat. Een mogelijke verklaring hiervoor is dat meer irrelevante informatie geactiveerd wordt tijdens restudy omdat er sprake is van afleiding en minder doelgerichte verwerking.



Figuur A.6 Links: Foto van een MRI scanner. In hoofdstuk 3 en 4 worden resultaten gepresenteerd van neuro-imaging-studies waarin volwassenen nieuwe woorden oefenden door middel van retrieval of restudy practice in een MRI scanner (foto: Image Editor, 2006). De activatie van verschillende delen van de hersenen werd gemeten door middel van functional magnetic resonance imaging (fMRI), een techniek die gebaseerd is op de koppeling tussen neurale activiteit en veranderingen in de doorbloeding. Rechts: De illustratie laat de menselijke hersenen zien vanaf de linkerkant (gebaseerd op Carter, 1918). De meest belangrijke hersengebieden die in hoofdstuk 3 en 4 behandeld worden zijn hier gemarkeerd. Zie figuur 4.2 voor meer informatie.

DEEL II. HOE KAN RETRIEVAL PRACTICE GEBRUIKT WORDEN TIJDENS HET LEREN VAN NIEUWE WOORDEN?

FEEDBACK EN TIPS TIJDENS RETRIEVAL PRACTICE

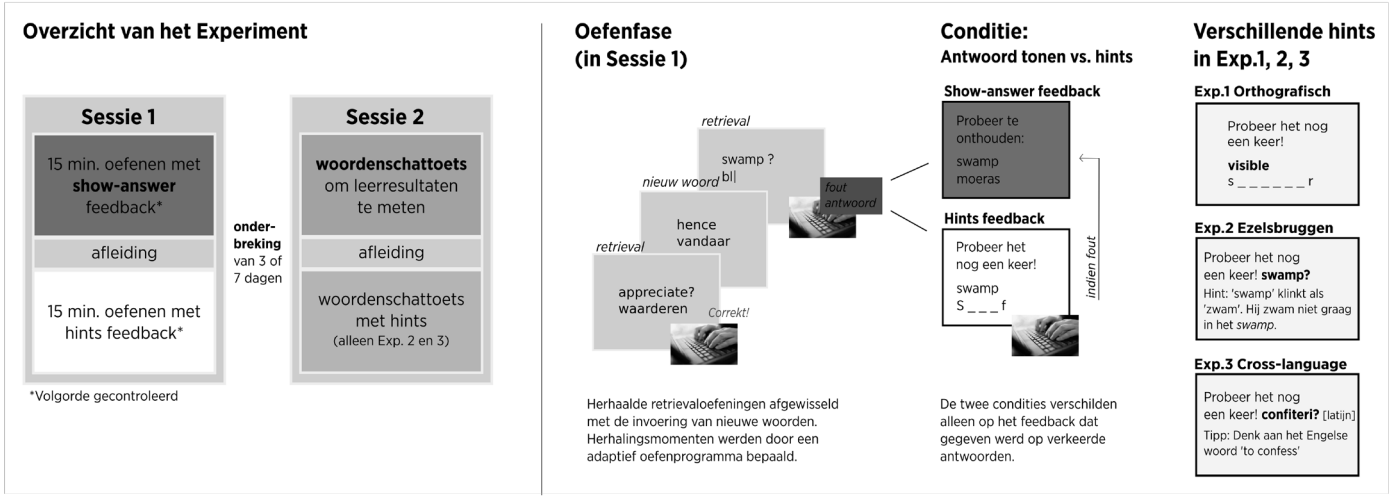
In het tweede gedeelte van het proefschrift heb ik het effect van retrieval tijdens woordenschatoefeningen onderzocht. In hoofdstuk 5 beschrijf ik hiervoor drie experimenten met middelbare scholieren die nieuwe woorden oefenden met behulp van een computerprogramma met retrieval practice. Verschillende soorten feedback werden vergeleken om te bepalen of tussentijdse tips het leerresultaat verbeteren.

Feedback verhoogt het positieve effect van retrieval practice omdat leerlingen hun fouten kunnen corrigeren en zekerder worden van hun antwoorden. Wanneer daarentegen een leerling een woord niet uit het geheugen kan oproepen en er geen feedback gegeven wordt, hebben retrieval-pogingen weinig nut. De meeste studies van retrieval practice bevatten daarom feedback waarin het correcte antwoord wordt vertoond. In hoofdstuk 5 onderzoek ik of feedback effectiever wordt wanneer de leerlingen eerst een tip krijgen waarmee ze zelf hun antwoord kunnen verbeteren (zie figuur A.7). Dit was in onze experimenten niet het geval.

De experimenten in hoofdstuk 5 lieten geen bewijs zien dat tips tijdens het oefenen nuttig zijn. Tips verminderden (herhaalde) fouten tijdens het oefenen niet en hadden ook geen positief effect op het leerresultaat op een latere toets. Integendeel, omdat het de leerlingen extra tijd kostte om de tips te verwerken, bleef minder tijd over om woorden te herhalen zodat de leerlingen in totaal minder woorden konden oefenen. Opvallend was dat leerlingen wel beter presteerden na het oefenen met tips (vergeleken met oefenen zonder tips) op een toets waarop dezelfde tips opnieuw beschikbaar waren. Dit duidt erop dat de leerlingen er wel beter in werden om de woorden met tips op te roepen uit het geheugen, maar hierdoor niet beter werden op een toets zonder tips. Mogelijk wordt de **woordkennis door het oefenen met tips deels afhankelijk van de latere beschikbaarheid van dezelfde tips.** Verder bleek dat oefenen met tips de leerresultaten niet verbeterde.

CONTEXTUELE INFORMATIE EN RETRIEVAL

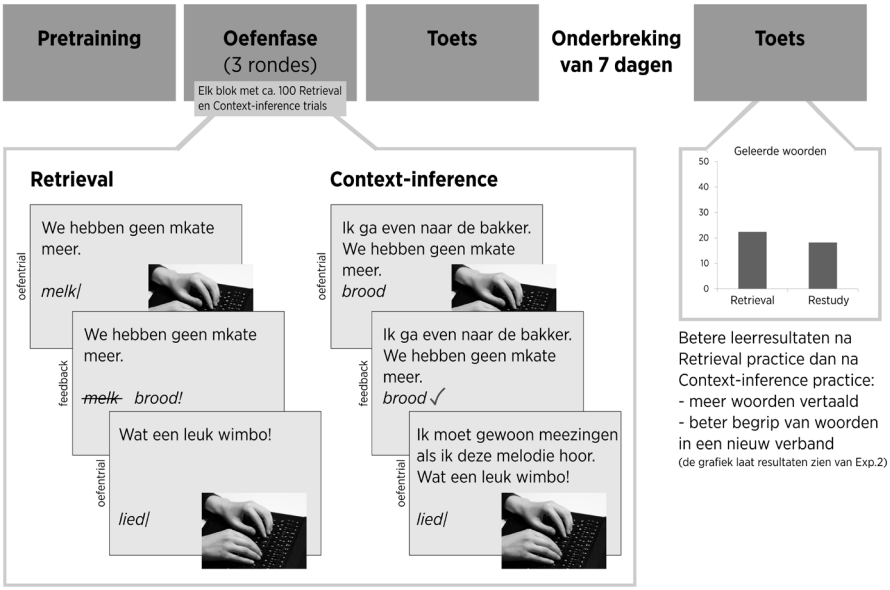
In hoofdstuk 6 beargumenteer ik dat **nieuwe woorden beter onthouden worden wanneer proefpersonen oefenen om de woorden uit het geheugen op te roepen (retrieval) dan wanneer ze de woorden uit een relevante, informatieve context afleiden.** De proefpersonen leerden nieuwe woorden (bijvoorbeeld het woord *funguo*) en oefenden deze woorden vervolgens óf in een niet-informatieve zin („Waar is de *funguo*?“) óf in een informatieve zin („Ik wil graag de deur op slot doen. Waar is de *funguo*?“). De proefpersonen moesten de woordbetekenis uit het geheugen oproepen wanneer woorden in de niet-informatieve zinnen aangeboden werden maar konden



Figuur A.7 In de experimenten die in hoofdstuk 5 beschreven worden, oefenden middelbare scholieren Engelse of Latijnse woorden met verschillende versies van een adaptief computerprogramma. Dit programma gebruikte een algoritme om de herhalingen van de woorden te baseren op gemiddelde vergeetcurves en vergrootte de tijd tussen herhalingen van hetzelfde woord geleidelijk. Wanneer leerlingen een fout maakten tijdens het oefenen, ontvingen zij óf feedback waarbij het antwoord getoond werd, óf feedback met tips. Deze tips verschilden in de drie experimenten. Geen van de tips verbeterde prestaties tijdens het oefenen of op een latere toets, behalve wanneer de test dezelfde tips bevatte die tijdens het oefenen aangeboden werden. Bovendien verminderden de tips de tijd die beschikbaar was voor verdere herhalingen.

de woordbetekenis uit de informatieve zinnen afleiden (hier: *funguo* = sleutel). De informatieve zinnen vergrootten dus de kans dat de proefpersonen de woorden tijdens het oefenen begrepen. Een test enkele dagen later liet echter zien dat de proefpersonen meer woorden onthielden na het oefenen met de niet-informatieve zinnen die retrieval uitlokten (zie ook figuur A.8).

De bevinding dat een niet-informatieve context leidt tot meer woordleren dan een informatieve context is tegen-intuïtief omdat contextuele informatie over het algemeen een nuttige bron van informatie is waarmee leerlingen nieuwe woorden kunnen begrijpen. Het is echter belangrijk om onderscheid te maken tussen begrip en leren: **wanneer een leerling een woord in context begrijpt, betekent dit niet automatisch dat de leerling het woord ook over de tijd heen onthoudt.** Hierop aansluitend liet hoofdstuk 6 zien dat een manipulatie die het makkelijker maakte



Figuur A.8 In de experimenten die in Hoofdstuk 6 besproken worden, bestudeerden de proefpersonen nieuwe woorden (bijvoorbeeld het woord *wimbo* = lied) en oefenden deze woorden vervolgens óf in een niet-informatieve retrieval zin („Wat een leuk wimbo!“) óf in een informatieve context-inference zin („Ik moet gewoon meezingen als ik deze melodie hoor. Wat een leuk wimbo!“). De proefpersonen moesten de woordbetekenis uit het geheugen oproepen wanneer ze met niet-informatieve zinnen oefenden, maar konden de woordbetekenis uit de informatieve zinnen afleiden. Het oefenen met de niet-informatieve zinnen leidde tot betere resultaten op toetsen onmiddellijk en meerdere dagen na het leren. De grafiek rechts vertoont exemplarisch de verschillen die in Experiment 2 gevonden werden.

om sleutelwoorden te begrijpen (namelijk de toevoeging van contextuele informatie tijdens het oefenen), het *minder* waarschijnlijk maakte dat de woorden over de tijd heen onthouden werden. De niet-informatieve context die retrieval vereiste, leidde tot betere langetermijnresultaten dan de informatieve context waaruit de betekenis afgeleid kon worden. Deze experimenten laten zien dat **testeffecten uitgelokt kunnen worden door de context waarin een woord verschijnt. Het verminderen van contextuele informatie kan retrieval nodig maken en hierdoor het leerresultaat verbeteren.**

CONCLUSIE EN PRAKTISCHE AANBEVELINGEN

Het ophalen van informatie uit het geheugen is geen simpel afspeelproces; elke retrieval verhoogt de toegankelijkheid van de opgehaalde informatie. Dit maakt retrieval een effectieve techniek om informatie op lange termijn te onthouden.

Dit proefschrift rapporteert data van gedragsstudies en neuro-imaging-studies naar de cognitieve mechanismen die ten grondslag liggen aan de positieve effecten van retrieval practice op het leren en onthouden van woorden. De resultaten wijzen erop dat retrieval inspanning vereist, maar toenemend makkelijker en sneller gaat met oefening. De reden hiervoor zou kunnen zijn dat retrieval de aandacht richt op informatie die relevant is om woorden uit het geheugen op te roepen, zoals het verband tussen de schrijfwijze en betekenis van nieuwe woorden.

De studies in dit proefschrift laten ook een aantal voorwaarden zien voor het succesvolle gebruik van retrieval practice in oefeningen voor vocabulaireverwerving, waaruit praktische aanbevelingen afgeleid kunnen worden voor het ontwerp van leersituaties.

Praktische aanbevelingen

- **Het oproepen van informatie uit het geheugen, bijvoorbeeld tijdens het overhoren, is een effectieve leerstrategie.** Bied belangrijke informatie niet alleen aan tijdens het oefenen maar laat leerlingen de informatie ook uit het geheugen reproduceren!
- **Combineer retrieval met feedback.** Feedback maakt het mogelijk voor leerlingen om informatie te bestuderen die ze nog niet uit het geheugen kunnen reproduceren. Zonder feedback hebben mislukte retrievalpogingen nauwelijks positieve effecten en kunnen fouten blijven bestaan.
- **Bied herhaaldelijk gelegenheid voor retrieval.** Zo kunnen oproepprocessen steeds gemakkelijker worden. Herhaling creëert ook nieuwe kansen tot succesvolle retrieval na een eerdere fout.
- **Pas de retrievaloefeningen aan op de vaardigheden van leerlingen! Idealiter is retrieval practice haalbaar maar inspannend.** Condities die de retrieval makkelijk maken leiden tijdens het oefenen tot goede prestaties en een (te) hoge inschatting van het eigen leerresultaat, maar op lange termijn tot slechtere resultaten. Zorg er bijvoorbeeld voor dat er voldoende tijd tussen herhalingen ligt zodat elke retrieval uitdagend blijft.
- **Pas de retrieval-situatie aan op het gewenste leerresultaat.** Zo verbetert het oefenen zonder tips de latere reproductie zonder tips. Daarentegen is de waarde van meerkeuzevragen voor het verbeteren van vrije reproductie beperkt. Zorg er daarom voor dat de manier van oefenen aansluit op de manier waarop leerlingen hun kennis later moeten kunnen gebruiken.
- **Maak onderscheid tussen de snelheid en het gemak van antwoorden tijdens het oefenen en de effecten van het oefenen op het leerresultaat.** Snelle antwoorden tijdens het oefenen zijn geen goede voorspellers van langetermijnresultaten wanneer de oefensituatie de retrieval makkelijk maakt, bijvoorbeeld door snelle herhaling van dezelfde woorden. Juist retrieval die inspanning vereist tijdens het oefenen zorgt voor een beter leerresultaat.
- **Creëer gelegenheid voor retrieval in verschillende leersituaties.** Een kleine aanpassing van een oefening kan al retrievalprocessen uitlokken, bijvoorbeeld wanneer een sleutelwoord eerst in een onduidelijke samenhang wordt aangeboden en de woordbetekenis pas later wordt uitgelegd.

BEELDRECHTEN

Figuur A.5:

Image Editor. (2006). *01 Siemens MAGNETOM Trio* [Digital image]. Retrieved from Flickr: https://www.flickr.com/photos/11304375@N07/3081315619/sizes/o/. Creative Commons Attribution License.

Carter, H.V. (1918). Principal fissures and lobes of the cerebrum viewed laterally. Figure 728 from H. Gray (Ed.), *Anatomy of the human body* (20th edition, revised by Warren H. Lewis). Philadelphia: Lea & Febiger. Image edited by O. Räisänen for wikimedia [Public domain]. Retrieved from https://commons.wikimedia.org/wiki/File:Lobes_of_the_brain_NL.svg

ZUSAMMENFASSUNG[1]

Viele Menschen gehen davon aus, dass das menschliche Gedächtnis wie eine Kamera funktioniert, die Erinnerungen aufzeichnet, speichert, und später wieder abspielt. Dieser Vergleich ist nicht korrekt: Das menschliche Gedächtnis ist in Wirklichkeit sehr viel komplexer und dynamischer. Eine wichtige Eigenschaft des Gedächtnisses ist, dass das Abrufen einer Erinnerung aus dem Gedächtnis den Inhalt und die spätere Zugänglichkeit der Erinnerung beeinflussen kann.

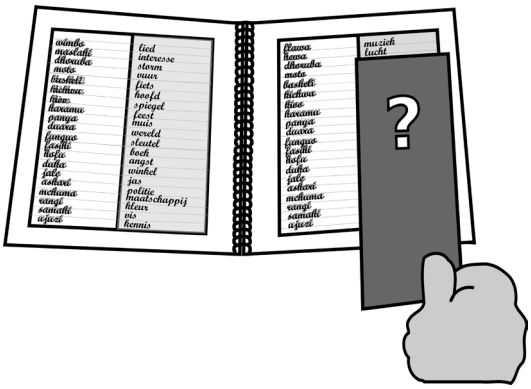
Für Lernende ist besonders relevant, dass **der Gedächtnisabruf[2] (engl. memory retrieval) trainiert werden kann.** Jeder Abruf einer Erinnerung aus dem Gedächtnis macht es leichter, die gleiche Information später erneut abzurufen. Dies ist von besonderem Interesse für alle, die große Mengen Informationen behalten möchten, beispielsweise um viele hundert Wörter einer Fremdsprache zu lernen (siehe Grafik A.9). *Retrieval Practice* (auf Deutsch etwa „Abrufübung") ist eine wirkungsvolle Methode, um Wörter über einen längeren Zeitraum zu behalten. Retrieval Practice ist außerdem deutlich effektiver als andere Wiederholungsmethoden, bei denen die zu behaltene Information dem Lernenden vollständig präsentiert wird wie etwa beim wiederholten Durchlesen von Wörtern mit Übersetzung (*Restudy*). Diese positiven Auswirkungen von Retrieval Practice verglichen mit anderen Übungsmethoden werden in der Literatur als **Testeffekt** bezeichnet.

Der erste Teil **dieser Doktorarbeit** beschäftigt sich mit den kognitiven und neuronalen Grundlagen des Testeffektes. Obwohl viele Studien Testeffekte dokumentieren, ist bislang wenig über die zugrundeliegenden Denkmechanismen bekannt. Durch Messungen von Reaktionszeiten und neuronaler Aktivität während *Retrieval* und *Restudy*, werden Erklärungen abgeleitet, warum Retrieval Practice einen positiven Effekt auf das Behalten von Wörtern hat.

Der zweite Teil der Doktorarbeit befasst sich mit der Anwendung von Retrieval Practice während Vokabelübungen. Hierfür wurden in mehreren Experimenten im Fremdsprachenunterricht an weiterführenden Schulen die Effekte von verschiedenen Formen von Feedback und von Hinweisen auf den Vokabelerwerb untersucht. Darüber hinaus beschreibt der zweite Teil der Arbeit mehrere Studien zum Effekt von relevanten kontextuellen Informationen während des Retrieval-Prozesses. Zum Abschluss werden praktische Empfehlungen für den Gebrauch von Retrieval in Lernsituationen gegeben.

1 Diese Zusammenfassung gibt einen kompakten Überblick über die wichtigsten Ergebnisse der Arbeit. Sie richtet sich an eine breite Leserschaft. Eine ausführlichere Zusammenfassung ist in Abschnitt 7.1 aufgenommen.

2 Das Üben des Reproduzierens oder Abrufens von Informationen aus dem Gedächtnis wird in dieser Zusammenfassung als *Retrieval Practice* bezeichnet (von engl. *memory retrieval*); die gängige Vergleichskondition des wiederholten Durchlesens zum Zweck des Behaltens wird als *Restudy* bezeichnet.



Grafik A.9 Ein/e Lernende/r übersetzt Vokabeln aus dem Gedächtnis. Dies erleichtert die spätere Erinnerung an die Wörter. Ein solches Abhören (*Retrieval Practice*) führt zu besseren Langzeitergebnissen als andere Übungsmethoden, wie etwa das wiederholte Durchlesen von Informationen (*Restudy*). Dieses Phänomen wird in der Literatur als *Testeffekt* bezeichnet.

TEIL I. *WARUM* RETRIEVAL PRACTICE DAZU FÜHRT, DASS WÖRTER BESSER BEHALTEN WERDEN

REAKTIONSZEITMESSUNGEN

In Kapitel 2 und 3 zeige ich, dass Lernende durch Retrieval Practice nicht nur *mehr* Wörter aus dem Gedächtnis reproduzieren können (verglichen mit Wiederholung durch Restudy), sondern diese Wörter auch *schneller* reproduzieren. Das bedeutet, dass nicht nur ein Testeffekt auf den Erfolg der späteren Reproduktion existiert, sondern auch auf die Geschwindigkeit der Reproduktion. Dies unterstützt Thesen aus der Literatur, dass **Retrieval-Prozesse durch Übung effizienter** werden. Möglicherweise werden künftige Suchen im Gedächtnis jedes Mal einfacher, wenn ein Wort erfolgreich aus dem Gedächtnis reproduziert wird. Die Studien in dieser Dissertation zeigen, dass eine Voraussetzung für diese Testeffekte ist, dass der Retrieval erfolgreich ist**. Nur wenn es den Teilnehmern tatsächlich gelang, die Wörter beim Üben aus dem Gedächtnis zu reproduzieren, waren die Retrieval-Übungen von Vorteil für das Behalten der Wörter.**

NEUROIMAGING-DATEN

Kapitel 3 und 4 präsentieren Neuroimaging-Daten die aufgezeichnet wurden, während Teilnehmer Vokabeln in einem Kernspintomographen übten. Unterschiede in der Hirnaktivität während Retrieval Practice und Restudy Practice wurden mit Hilfe funktioneller Magnetresonanztomographie (fMRT) beschrieben (siehe Grafik

A.10). Diese Technik nutzt den Zusammenhang zwischen neuronaler Aktivität und regionaler Durchblutung verschiedener Hirnareale, wobei letztere gemessen werden kann auf Grund unterschiedlicher magnetischer Eigenschaften von sauerstoffarmem und sauerstoffreichem Blut. Der Vergleich der Hirnaktivität während Retrieval- und Restudy-Übungen führte zu zwei Hauptergebnissen:

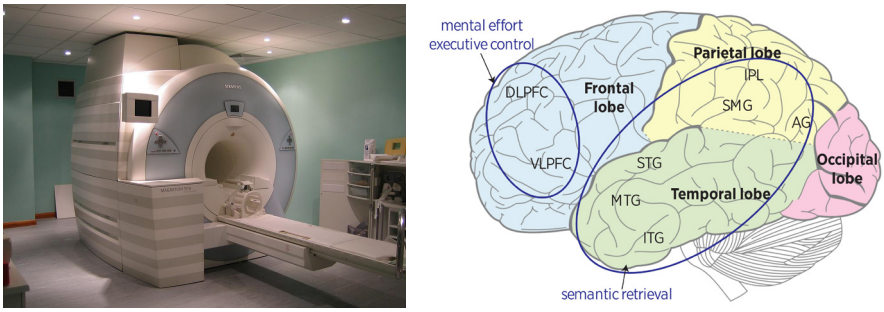
1. Retrieval erfordert vermutlich mehr mentale Anstrengung als Restudy Practice, was sich in höherer Aktivität in frontalen Hirnregionen widerspiegelt. Diese Anstrengung nimmt mit wiederholter Retrieval- Übung ebenso wie die frontale Aktivität ab.

Gebiete im vorderen Teil des Gehirns, genauer im ventrolateralen Präfrontalcortex (VLPFC), zeigten in verschiedenen Studien beschrieben in Kapitel 3 und 4 erhöhte Aktivierung während Retrieval Practice verglichen mit Restudying. Diese Hirnareale sind häufig aktiv wenn Aufgaben mentale Kontrolle erfordern, d.h. erhöhte Konzentration und die bewusste Steuerung von Aufmerksamkeit. Solche mentale Kontrolle ermöglicht es beispielsweise, relevante Informationen auszuwählen und ablenkende, irrelevante Informationen zu ignorieren. **Höhere Aktivität im VLPFC während Retrieval – verglichen mit Restudying – deutet auf eine stärkere Rolle von mentaler Kontrolle während Retrieval Practice hin**. Kontrolliertes, angestrengtes Verarbeiten von Informationen führt häufig zu besseren Langzeit-Lernergebnissen als unangestrengtes Verarbeiten und könnte daher eine Erklärung für Testeffekte sein. Darüber hinaus zeigten die Studien, die in Kapitel 4 beschrieben werden, dass im Laufe von wiederholter Retrieval Practice und nach Retrieval Practice (verglichen mit Restudy Practice) die Aktivität des VLPFC abnahm. Zusammen weisen diese Ergebnisse darauf hin, dass **Retrieval angestrengtes, kontrolliertes Verarbeiten von Informationen beinhaltet, aber durch wiederholtes Üben mit stets weniger Anstrengung erfolgt.**

2. Aktivität in Hirnarealen für semantische (bedeutungsvolle) Verarbeitung deutet darauf hin, dass Retrieval Practice zu einem Fokus von Aufmerksamkeit auf relevante Informationen führt und damit Assoziationen verstärkt, die die spätere Erinnerung erleichtern.

Während Retrieval Practice waren Gebiete im unteren parietalen und mittleren temporalen Cortex aktiver wenn die Teilnehmer Wörter übten, die sie bei einem Vokabeltest nach dem Üben noch kannten, als wenn sie Wörter übten, die sie beim Vokabeltest vergessen hatten. Die Aktivität in diesen Hirnarealen war also höher, wenn Vokabeln erfolgreich geübt wurden. Während des Restudying bestand kein solcher Unterschied zwischen behaltenen und vergessenen Wörtern; die Hirnaktivität während des wiederholten Durchlesens war also nicht vorhersagend für die Ergebnisse beim

späteren Vokabeltest. Möglicherweise ist dies auf Unterschiede in der Qualität der semantischen (d.h., auf die Bedeutung gerichteten) Verarbeitung der Wörter während Retrieval und Restudying zurückzuführen, da in der Literatur davon ausgegangen wird, dass die betroffenen Areale eine Rolle bei semantischer Verarbeitung spielen. **Retrieval Practice könnte die Aufmerksamkeit während des Übens auf relevante Informationen lenken,** etwa auf auffällige Schreibweisen der Vokabeln oder Eselsbrücken, die einen Zusammenhang herstellen zwischen Wortschreibweise und Bedeutung. Dies kann **zu einer Gedächtnisrepräsentation führen, die leichter reproduziert werden kann**. Restudy dagegen könnte aufgrund von Ablenkung und weniger zielgerichtetem Verarbeiten der angebotenen Wörter eine vermehrte Verarbeitung von irrelevanten Informationen beinhalten. Daher ist die Aktivität in Gebieten für semantische Verarbeitung während Restudy nicht vorhersagend für das Lernergebnis.



Grafik A.10 Links: Foto eines Kernspintomographen. In Kapitel 3 und 4 werden Ergebnisse von Neuroimaging Studien präsentiert, in denen erwachsene Probanden Vokabeln durch Retrieval Practice (Abhören) oder Restudying (wiederholtes Durchlesen) übten. Die Veränderung der Aktivität in verschiedenen Hirnarealen wurde mit Hilfe funktioneller Magnetresonanztomographie (fMRT) beschrieben (Foto: Image Editor, 2006.). Die Illustration rechts zeigt ein Gehirn von der linken Seite (Illustration basiert auf Carter, 1918). Die wichtigsten Areale, die in den Studien in Kapitel 3 und 4 besprochen werden, sind hier hervorgehoben. Für weitere Informationen siehe Grafik 4.2 in Kapitel 4.

TEIL II. WIE RETRIEVAL PRACTICE EINGESETZT WERDEN KANN BEIM VOKABELLERNEN

FEEDBACK UND HINWEISE WÄHREND RETRIEVAL PRACTICE

Der zweite Teil der Dissertation beschreibt den Effekt von Retrieval während verschiedener Vokabelübungen. In Kapitel 5 werden drei Experimente beschrieben, in denen Schüler Vokabeln am Computer mit Hilfe von Retrieval Practice übten.

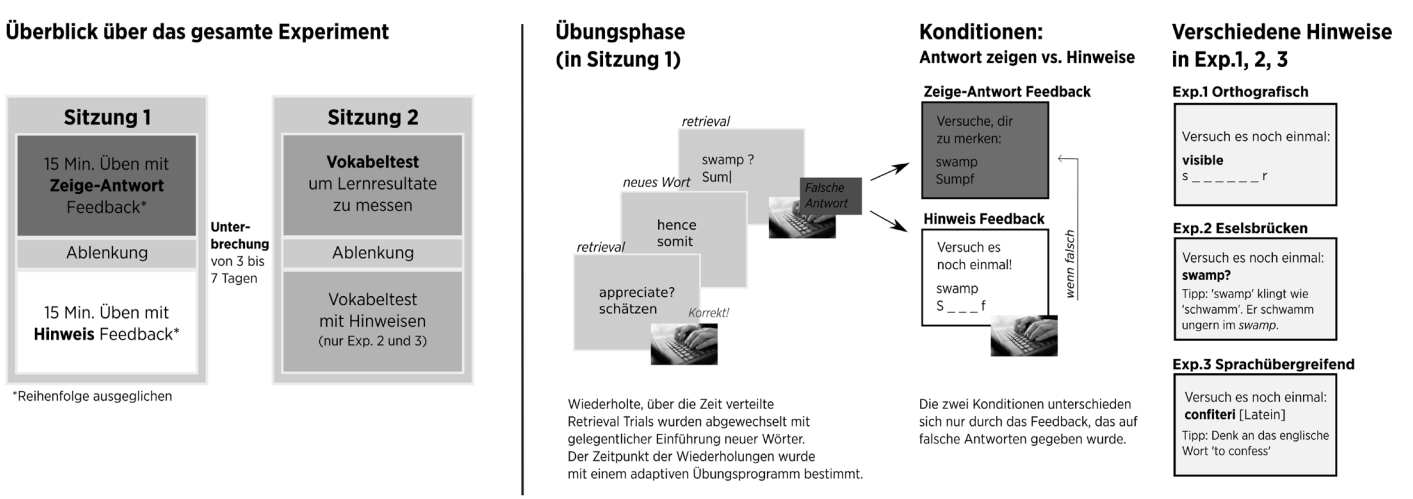
Verschiedene Rückmeldungen während des Übens (*Feedback*) wurden dabei verglichen um festzustellen, ob es hilfreich ist, Hinweise anzubieten.

Rückmeldungen (Feedback) verstärken den positiven Effekt von Retrieval Practice, da sie den Lernenden erlauben, Fehler zu verbessern und das Vertrauen in unsichere Antworten erhöhen. Dagegen hat ein fehlgeschlagener Retrieval-Versuch – etwa wenn ein Lernender ein Wort nicht aus dem Gedächtnis reproduzieren kann – ohne Feedback kaum positive Effekte. Die meisten Studien in der Literatur über Testeffekte verwenden einfache Rückmeldungen, bei denen die korrekte Antwort gezeigt wird. In den Experimenten in Kapitel 5 wurde untersucht, ob Rückmeldungen effektiver sind, wenn die Lernenden zunächst Hinweise bekommen, die es ihnen ermöglichen, ihre Antwort selbst zu korrigieren (siehe Grafik A.11). Dies war nicht der Fall.

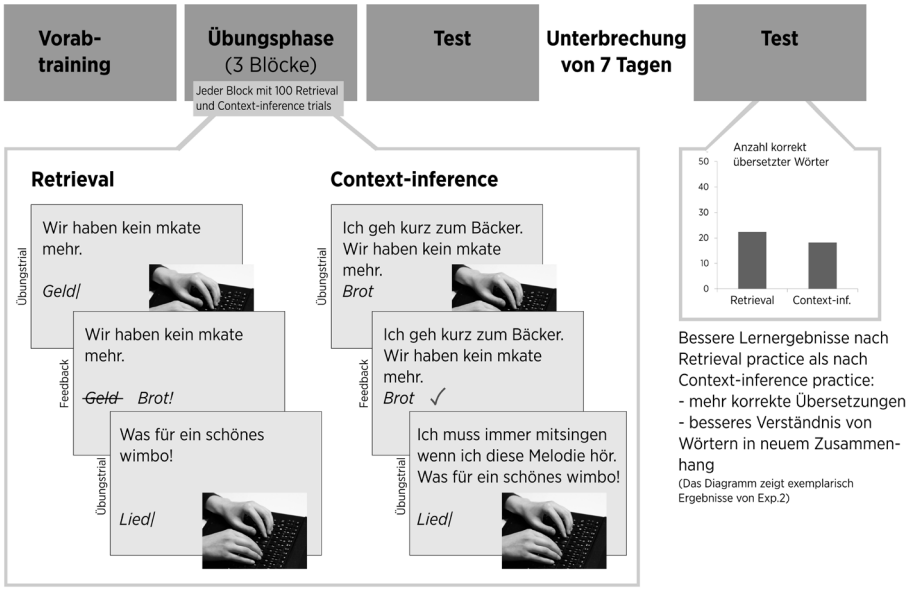
Die Experimente in Kapitel 5 lieferten keinerlei Beweis dafür, dass Hinweise vorteilhaft sind. Die Hinweise reduzierten die Anzahl der (wiederholten) Fehler während des Übens nicht und verbesserten auch nicht die Ergebnisse bei einem späteren Vokabeltest. Da es die Schüler Zeit kostete, die Hinweise zu verarbeiten, blieb weniger Zeit für Wiederholungen und die Schüler konnten insgesamt weniger Wörter üben. Der einzige positive Effekt der Hinweise wurde gefunden, wenn die gleichen Hinweise wie beim Üben auch während des Vokabeltests erneut angeboten wurden. Dies deutet darauf hin, dass die **Schüler lediglich besser darin wurden, Wörter mit Hinweisen zu übersetzen, aber dieses Wissen nicht auf einem Vokabeltest ohne Hinweise anwenden konnten**. Insgesamt zeigten die drei Experimente, dass **Retrieval Practice mit Hinweisen nicht zu effektiverem Vokabellernen führte als Retrieval Practice mit gewöhnlichem Feedback.**

KONTEXTINFORMATIONEN UND RETRIEVAL

Die Experimente in Kapitel 6 zeigten, dass die Teilnehmer **Wörter besser behielten, wenn sie die Wörter wiederholt aus dem Gedächtnis abriefen (Retrieval) als wenn sie die Bedeutung der Wörter aus einem relevanten, informativen Zusammenhang (Kontext) ableiteten.** Die Teilnehmer lernten zunächst einige Wörter einer für sie unbekannten Sprache (zum Beispiel „*funguo*“) und übten diese Wörter dann entweder in einem nicht-informativen Satz („Ich brauche den *funguo.*“) oder in einem informativen Satz („Ich möchte diese Tür aufschließen. Ich brauche den *funguo.*“, vgl. Grafik A.12). Die Teilnehmer mussten die Bedeutung der Schlüsselwörter aus dem Gedächtnis abrufen, wenn sie mit nicht-informativen Sätzen übten, aber konnten die Bedeutung der Wörter aus dem Kontext ableiten, wenn sie mit einem informativen Satz übten (hier: *funguo* = Schlüssel). Die informativen Sätze machten es also leichter, die Wörter während des Übens zu verstehen. Jedoch zeigte ein Vokabeltest mehrere Tage später, dass die Teilnehmer die Wörter besser nach Retrieval Practice mit nicht-informativen Sätzen behalten hatten.



Grafik A.11 In den Experimenten in Kapitel 5 übten Schüler englische oder lateinische Vokabeln mit verschiedenen Versionen eines speziellen Computerprogramms: Dieses Programm bestimmt die Wiederholungsrate der Vokabeln ausgehend von typischen Vergessenskurven und passt den Abstand zwischen Wiederholungen einzelner Wörter während des Übens an die Leistung des Benutzers an. Wenn Schüler einen Fehler machten, wurde entweder Standard Zeige-Antwort-Feedback gezeigt (die korrekte Antwort) oder es wurden verschiedene Hinweise gezeigt, mit denen die Schüler noch einmal versuchten, die richtige Antwort zu geben. In den Experimenten wurden hierfür drei verschiedene Sorten von Hinweisen verwendet. Diese verbesserten weder die Ergebnisse während des Übens noch die Ergebnisse bei einem späteren Vokabeltest. Darüber hinaus hatten einige Hinweise einen negativen Effekt, weil sie die Zeit verringerten, die für weitere Wiederholungen verfügbar war.



Grafik A.12 In den Experimenten in Kapitel 6 lernten de Teilnehmer zunächst unbekannte Wörter (z.B., *wimbo* = Lied) und übten diese Wörter dann entweder mit nicht-informativen Retrieval-Sätzen („Was für ein nettes *wimbo*!“) oder informativen Sätzen, aus denen die Bedeutung abgeleitet werden konnte („Ich muss einfach mitsingen wenn ich diese Melodie höre. Was für ein nettes *wimbo*!“). Die Teilnehmer mussten ihr Wissen über die Wörter aus dem Gedächtnis abrufen, um die Wörter in den nicht-informativen Sätzen zu verstehen, konnten aber die Bedeutung aus den informativen Sätzen ableiten. Das Üben mit den nicht-informativen Retrieval-Sätzen führte zu besseren Leistungen bei späteren Vokabeltests. Der Graph rechts zeigt exemplarisch die durchschnittlichen Ergebnisse für eine der Messungen in Experiment 2.

Das Ergebnis, dass nicht-informative Sätze zu besserem Vokabellernen führten, ist überraschend, da Kontextinformationen im Allgemeinen eine nützliche Quelle von Informationen sind, mit der Lernende neue Wörter im Zusammenhang verstehen können. Das *Verstehen* eines Wortes im Kontext führt jedoch nicht automatisch auch zum *Behalten* des Wortes. So zeigte Kapitel 6, dass eine Manipulation, die es *leichter* machte, die Schlüsselvokabeln zu verstehen (nämlich das Präsentieren eines relevanten Zusammenhangs), es *weniger* wahrscheinlich machte, dass die Wörter behalten wurden. Die nicht-informativen Sätze, die einen Abruf der Wörter aus dem Gedächtnis erforderten, führten zu besserer Langzeitretention als die informativen Sätze. Diese Experimente zeigen, dass Testeffekte auch durch eine Manipulation des Textes hervorgerufen werden können, in dem ein Wort erscheint. Da eine Reduzierung

von Kontextinformationen einen Retrieval Prozess erforderlich macht, kann sie zu besseren Lernergebnissen führen – wenn der Retrieval gelingt und die Lernenden, die Informationen aus dem Gedächtnis reproduzieren können.

SCHLUSSFOLGERUNG UND PRAKTISCHE EMPFEHLUNGEN

Der Abruf von Informationen aus dem Gedächtnis (*Retrieval*) ist kein simpler Ausleseprozess: Jeder Abruf einer Erinnerung aus dem Gedächtnis macht es leichter, die gleiche Information später erneut abzurufen. Dies macht Retrieval Practice zu einer wirkungsvollen Lerntechnik, um Informationen langfristig zu behalten.

Diese Dissertation beschreibt anhand von Verhaltensstudien und Studien mit bildgebenden Verfahren die kognitiven Mechanismen, die die positive Wirkung von Retrieval erklären. Die Ergebnisse weisen darauf hin, dass Retrieval ein anstrengender Prozess ist, der durch Übung leichter wird. Diese Erleichterung könnte durch selektive Verarbeitung von semantischen Informationen entstehen, die den späteren Abrufprozess erleichtern – beispielsweise Assoziationen zwischen der Schreibweise und Bedeutung neuer Vokabeln. Die Verstärkung solcher Assoziationen könnte Wortrepräsentationen im Gedächtnis so verändern, dass diese später leichter und schneller abgerufen werden können.

Die Ergebnisse dieser Dissertation zeigen eine Reihe von Voraussetzungen für die erfolgreiche Nutzung von Retrieval Practice in Vokabelübungen. Aus diesen Voraussetzungen können praktische Empfehlungen für die Gestaltung von Lernsituationen abgeleitet werden. Diese sind ein Beispiel dafür, wie grundlegende Untersuchungen der Architektur des menschlichen Gedächtnisses zur Verbesserung von Lernsituationen genutzt werden können.

Praktische Empfehlungen

- **Das Abrufen von Informationen aus dem Gedächtnis, beispielsweise während des Abhörens von Vokabeln, ist eine wirkungsvolle Lerntechnik.** Präsentieren Sie Schlüsselinformationen also nicht nur beim Üben, sondern lassen Sie die Lernenden die Schlüsselinformationen auch aus dem Gedächtnis reproduzieren!
- **Kombinieren Sie Retrieval mit Feedback.** Dies erlaubt es Lernenden, Informationen zu studieren die sie noch nicht aus dem Gedächtnis reproduzieren können. Gescheiterte Retrieval-Versuche haben kaum positive Effekte ohne eine solche Rückmeldung.
- **Bieten Sie wiederholt Gelegenheit zum Retrieval.** Sie ermöglichen so, dass Abrufprozesse zunehmend erleichtert werden. Wiederholung stellt auch sicher, dass nach einem gescheiterten Versuch eine neue Chance zum erfolgreichen Retrieval geboten wird.
- **Passen Sie Retrieval Practice an die Fertigkeiten der Lernenden an.** Idealerweise ist Retrieval Practice erfolgreich aber herausfordernd. Bedingungen, die den Retrieval erleichtern, können zu guter Leistung und (unrealistisch) hohen Einschätzungen des eigenen Lernstands während des Übens führen, gleichzeitig aber zu schlechteren Langzeitergebnissen. Stellen Sie zum Beispiel sicher, dass Zeit zwischen Wiederholungen liegt, sodass der Retrieval herausfordernd bleibt.
- **Passen Sie die Bedingungen des Retrieval Practice an das gewünschte Lernergebnis an.** Das Üben ohne Hinweise erleichtert zum Beispiel die spätere Reproduktion ohne Hinweise. Dagegen sind die Vorteile von Multiple-Choice-Übungen für spätere freie Reproduktion begrenzt. Wählen Sie ein Format, das zum gewünschten Lernergebnis passt.
- **Unterscheiden Sie zwischen dem Verlauf von Retrieval Prozessen während des Übens und den Auswirkungen der Übung auf spätere Leistungen.** Wenn die Übungssituation den Retrieval vereinfacht (etwa durch schnelle, gehäufte Wiederholung) kann aus der Leichtigkeit mit der Antworten während des Übens gegeben werden, keine Vorhersage über den Lernerfolg getroffen werden. Eine Übung, die zu angestrengterem Retrieval führt, kann im Vergleich zu einem höheren Lernerfolg führen.
- **Schaffen Sie Gelegenheiten für Retrieval in verschiedenen Lernsituationen.** Schon kleine Anpassungen von Übungen können Retrieval hervorrufen, zum Beispiel indem man ein Schlüsselwort zunächst in einem undeutlichen Zusammenhang präsentiert und die Wortbedeutung erst später erklärt.

BILDNACHWEISE

Grafik A.10:

Image Editor. (2006). *01 Siemens MAGNETOM Trio* [Digital image]. Retrieved from Flickr: https://www.flickr.com/photos/11304375@N07/3081315619/sizes/o/. Creative Commons Attribution License.

Carter, H.V. (1918). Principal fissures and lobes of the cerebrum viewed laterally. Figure 728 from H. Gray (Ed.), *Anatomy of the human body* (20th edition, revised by Warren H. Lewis). Philadelphia: Lea & Febiger. Image edited by O. Räisänen for wikimedia [Public domain]. Retrieved from https://commons.wikimedia.org/wiki/File:Lobes_of_the_brain_NL.svg

ACKNOWLEDGEMENTS | DANKWOORD | DANKWORT

There are many people who contributed to the completion of this book. I want to use these last pages to thank you all for your help, guidance, and company.

My supervisors were a source of positivity and encouragement throughout this project. I always left our meetings in high spirit and with the idea that everything could be done in no time. That impression was usually wrong - but very motivating while it lasted.

Ludo, your bird's eye view on the structure of our papers and the humorous reminders to keep things simple(r) helped me become a better writer. Thank you for your trust and for giving me the freedom to carve my own path even when it took me from imaging research to classroom interventions. I still remember my excitement when you walked into my office and invited me to the SSSR symposium in Montréal in the first months of my project. Thank you for this and many other opportunities!

Atsuko, thank you for your detailed, helpful responses to all of my questions. You are very thorough in your work and supervision, and your kindness and patience make you a wonderful mentor. Thank you for the great conversations and advice, and for being involved in this project at every step of the way.

Eliane, I hope that one day I will be as good as you are at giving your students confidence and motivating them to get things done. Your critical questions and constructive feedback helped me with many decisions in this project. Thank you also for the practical, no-nonsense perspective that you brought to the supervision team.

Guillén, our contacts were infrequent but I appreciated your reliable, quick responses and your interest in my work. Your sharp comments and different perspective made me think deeper about findings. Thank you for providing me with access to the resources at the Donders Institute by accepting me into your research group.

The members of the *manuscript committee, prof. McQueen, prof. van den Bosch, and prof. Kester,* evaluated the dissertation and provided helpful, thoughtful feedback. Thank you! I also thank the other *opponents, prof. Bekkering, dr. Molenaar, dr. Verkoeijen, and dr. Strating*, for participating in the defence. I look forward to discussing your comments on the thesis.

I would like to thank a number of researchers who contributed directly to the work reported in this dissertation: *Hedderik van Rijn* for allowing me to use the SlimStampen algorithm, your involvement in the design of experiments and for your advice, including on mixed model analyses. It was a pleasure to work with you and learn from you! Thanks are also due to *Carola Wiklund-Hörnqvist, Linnea Karlsson, and Lars Nyberg* from Umea University for the invitation to a fabulous symposium and their persistence during the publication of our literature review. I learned a lot from our exchange.

I am grateful to *Jeffrey Karpicke* for inviting me to visit his research group at Purdue University, where I learned a lot about testing effect research and postdoc life in the United States. Testing effect researchers at other universities in the Netherlands have also kindly included me in interesting events and symposia. A word of thanks to *Huib Tabbers and Nicole Goossens* for this!

The work reported in this thesis would not have been possible without more than 650 anonymous *participants*, including 316 *students* from different high schools. Special thanks to the teachers who accommodated the experiments in their lessons and the participating students for their enthusiastic responses and good ideas. I am also grateful to my master students *Manon, Nathalie, Lotte, and Wendy* for their help with the data collection.

Radboud University is a great working environment with a high density of smart, interesting people who make working (and taking a break) more fun. Beste (oud-) collega's, dank voor jullie advies, de gezelligheid en inspirerende gesprekken, de fijne NmG cursus, epische lunches met ukulele, voor het pilottesten van experimenten en figurant zijn in films, powerpoint slides, etentjes, uitjes, brainstormsessies, en alle tips over onderwijs, onderzoek en promoveren: *Caressa, Elise, Suzanne, Henriette, Frauke, Linda, Sabine, Nathalie, Eva, the word learning journal club: Carmen, Loes, Roza, Nicole; Evelien (v.W.), Monique, Suzan, Cindy, Marco, Sanne, Liza, Sophie, Carolien, Evelien (M.), Joyce, Helen, Joep, Barbara, Gitta, Neomie, Brigitte, Roy, Kim, Arjan, Mark, Lian, Stijn, Merel, Iske, Marianne, Marjolein, Esther, Mathijs, Lex, Anneke, Janneke, Lanneke, Lonneke, Christel, Anne-Els, en Katja.* I also drew much inspiration and motivation from working with fantastic colleagues in the university's works council, especially *Dorian, Daniela, Bart, Ezra, and Frank,* and from the Memrise project with *Anke Marit, Ruud, Marlieke, Boris, Nils and Paul,* which was a great creative outlet!

Special thanks to everyone who proofread this dissertation – *Claudia, Thomas, Carmen, Daniela, Henriette, Sanne, Cindy, Marco, and Evelien*. If there are any spelling errors left, then I probably added them after you read the manuscript. Dankjewel, *Merel*, voor het vertalen van mijn 1001 ideeën naar een fantastisch coverontwerp en *Joska* voor de prettige professionele samenwerking bij de vormgeving. Ein Dankeschön auch an die Grafik-Co-Nymphe *Christine* für alle Tipps zu Indesign und Lay-Out. Den Mitgliedern des fröhlichen Womba (*wimbo*!) Teams, *Martina, André, Carmen und Nanette*, danke ich für die Hilfe beim Erstellen der 200 Stimuli für Kapitel 5.

One of the best traditions in Dutch academia is that PhD candidates have *paranymphs* at their side during the defence and, if they're lucky like me, during the preparations of the dissertation and festivities. *Dankjewel, Lucy, Henriette, en Daniela* for taking on this task so brilliantly! Dankzij jullie fijne ondersteuning heb ik veel plezier met de voorbereidingen!

I am not the type to print declarations of love but I want to thank my friends for the support in the past years, for checking in, thinking along, and for celebrating small (and large) successes together. Our creative projects, travels, and evenings of laughter, games, food and dancing take my mind off work and make me happy and productive. A big warm thank you goes out to you, *Lucy en Wouter (en Nelis en Ada), Henriette, Daniela (Patru), Caressa, Annika, Daniela (Post), Catrin, Daniela (B.), Sven, Laura, Vanja and the other ladies of our academic knitting squad, Gesa ('the other one'), Christine, Doro, Christian, Esther, Alex, Susanne, and the Lindy Hop dancers in Nijmegen*!

Zum Schluss danke ich meiner Familie für das Vertrauen, die geduldige Unterstützung und viele lustige, entspannte Momente. *Regina* und *Thomas* bin ich gerade besonders dankbar für meinen entspannten Rücken; meiner Großmutter *Elsa* für das liebevolle Daumendrücken; *Ulrike* für den Witz und die Gelassenheit, die sie schon sehr lange in meinem Leben verströmt ("Lass sie!"); *Egon* für die Anrufe um sieben Uhr morgens mit frischen Nachrichten aus der Tageszeitung; und *Carmen* für viel Unterstützung in sehr vielen Dingen. Liebe Carmen, danke! Ohne dich wäre ich weniger wohlgenährt und weniger entspannt gewesen während dieses langen Projekts. *Paul*, dir danke ich für ein Schmunzeln und Lasagne im richtigen Moment - und für vieles andere. Ich freu mich über uns und auf unsere Zukunft!

PS: This section is dedicated to all PhD students who read it instead of finishing their papers.

AUTHOR BIOGRAPHY

Gesa van den Broek (Kleve, 1985) holds a Bachelor's degree in Psychology and a Research Master's degree in Behavioural Science from Radboud University. During her studies, she became interested in connecting fundamental learning sciences and educational practice. She specialized in educational neuroscience, supported by a Huygens talent grant from the Dutch Ministry of Education, Culture, and Science. Gesa graduated summa cum laude in 2011 and then worked for the Organization for Economic Co-operation and Development in Paris on a project on innovative learning environments. She continued to work on this international project as an external consultant for several years after her return to Nijmegen, where she began a PhD project in 2011. Gesa carried out most of the research for her PhD project at the Behavioural Science Institute. A Mohrmann prize by the Radboud University and a Language Learning dissertation grant allowed her to visit other research groups abroad, including a stay of three months at the Cognition and Learning Laboratory of Purdue University. During her PhD project, Gesa taught in the educational sciences programme and supervised research projects by Bachelor and Master students at Radboud University. She was also an elected member of the University Council from 2013 to 2015. In this role, she represented employee interests in discussions with the university's Executive Board on topics like assessment standards and PhD employment conditions. Gesa has been invited to present her research at various international conferences and symposia. Her research has been covered in different news outlets, most recently because she won an international science competition for a vocabulary learning method that she developed together with colleagues. At the time of writing, Gesa is about to start a new position as postdoctoral researcher at Utrecht University. More information can be found on her homepage: www.gesavdbroek.net.

PUBLICATIONS

van den Broek, G. S. E. (2012). *Innovative Research-Based Approaches to Learning and Teaching* (OECD Education Working Papers No. 79). http://dx.doi.org/10.1787/5k97f6x1kn0w-en

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

van den Broek, G. S. E., Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? Insights from immediate and delayed retrieval speed. *Memory*, *22*(7), 803–812. https://doi.org/10.1080/09658211.2013.831455

van den Broek, G. S. E., Segers, E., & Verhoeven, L. (2014). Effects of text modality in multimedia presentations on written and oral performance. *Journal of Computer Assisted Learning*, *30*(5), 438–449. https://doi.org/10.1111/jcal.12058

van den Broek, G. S.E., Segers, E., Takashima, A., Verhoeven, L. (2014). Het testeffect en het brein. *Didactief, 4*, 22-23. Retrievable via http://goo.gl/HmuPNL van den Broek

van den Broek*, G. S. E., Takashima*, A., Wiklund-Hörnqvist, C., Karlsson Wirebring, L., Segers, E., Verhoeven, L., & Nyberg, L. (2016). Neurocognitive mechanisms of the “testing effect”: A review. *Trends in Neuroscience and Education*, *5*(2), 52–66. https://doi.org/10.1016/j.tine.2016.05.001
*equal contributions

Behavioural
Science
Institute

Radboud University