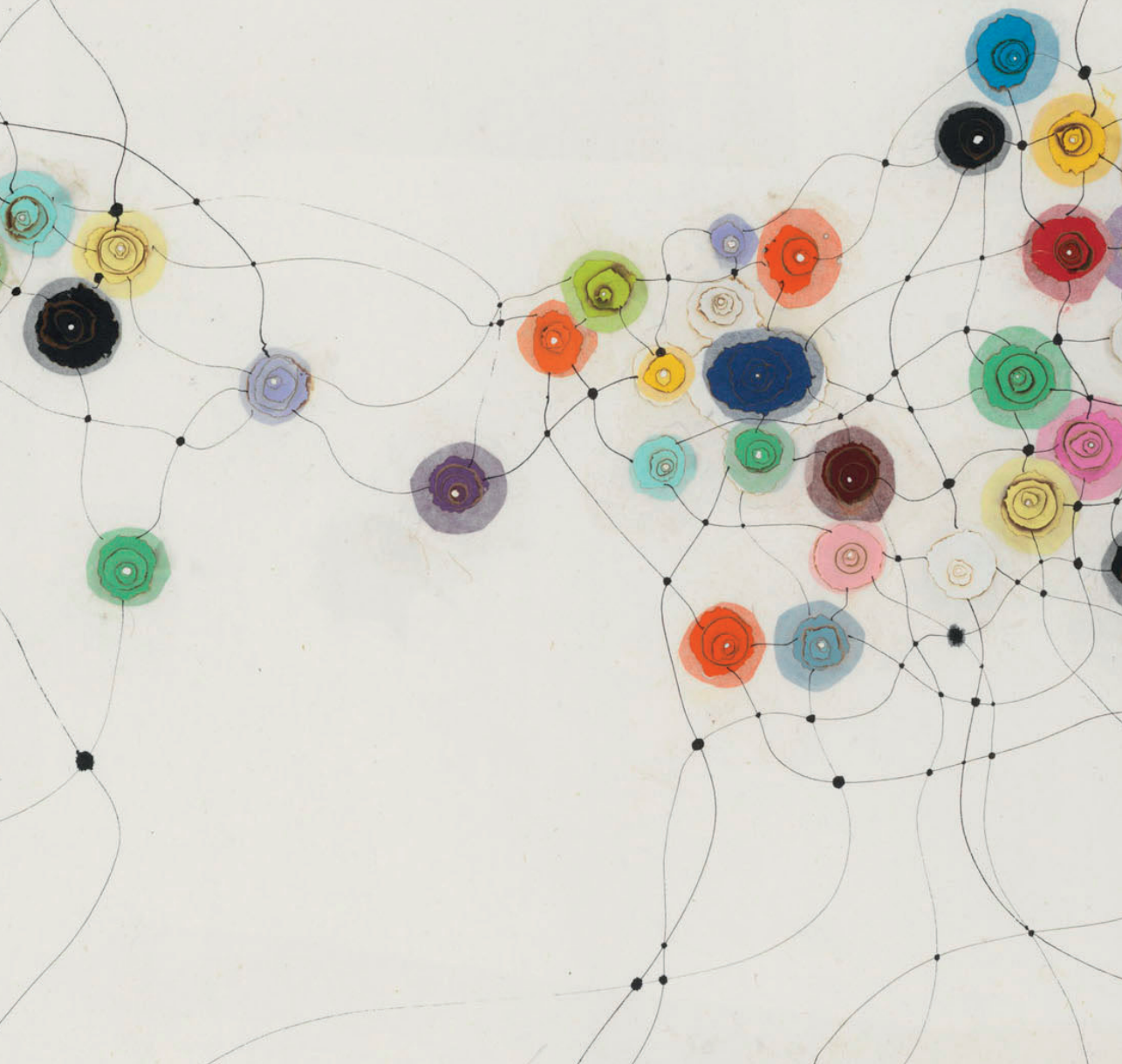
Limited Effectiveness of Retrieval Practice for the Transfer of Knowledge

Gerdien van Eersel



Stellingen behorende bij het proefschrift

Limited Effectiveness of Retrieval Practice for the Transfer of Knowledge

Gerdien van Eersel

1. Het uit het geheugen ophalen van een eerder geleerde tekst leidt niet tot een betere prestatie op een onmiddellijke of uitgestelde transfertaak dan herlezen. (dit proefschrift)
2. Het beantwoorden van tussentijdse vragen over een eerder geleerde tekst en het vervolgens inzien van de goede antwoorden (feedback) leidt tot een betere prestatie op een uitgestelde transfertaak dan tussentijds herlezen. (dit proefschrift)
3. Het uit het geheugen ophalen van een eerder geleerde informatieve tekst leidt niet tot een betere prestatie op een onmiddellijke of uitgestelde retentietaak dan herlezen. (dit proefschrift)
4. Het positieve effect van een oefentest op een retentie- of transfertaak wordt versterkt door het aanbieden van feedback. (dit proefschrift)
5. Er is onvoldoende bewijs dat het uit het geheugen ophalen van eerder bestudeerde stimuli de semantische geheugensporen ervan meer versterkt dan herbestuderen. (dit proefschrift)
6. The underdetermination of a theory by empirical data must not be considered as a problem, but as a property of the theory that can be employed for optimising empirical research. (Jan-Willem Romeijn)
7. It is easy to lie with statistics, but easier without them. (Frederick Mosteller)
8. The totality of our so-called knowledge or beliefs ... is a man-made fabric which impinges on experience only along the edges. Or, to change the figure, total science ... is so undetermined by its boundary conditions, experience, that there is much latitude of choice as to what statements to re-evaluate in the light of any single contrary experience. (Willard Van Orman Quine)
9. Het is onrechtvaardig dat kinderen met hoogopgeleide ouders hoger eindigen op de onderwijsladder dan gelijk getalenteerde kinderen met laagopgeleide ouders. (zie Staat van het Onderwijs 2015/2016, Onderwijsinspectie)
10. De briljante serie The Wire (seizoen 4) laat zien dat onderwijs vaak niet kan compenseren voor sociale ongelijkheid.
11. Angst is de duizeling van de vrijheid. (Søren Kierkegaard)

Limited Effectiveness of Retrieval Practice for the Transfer of Knowledge

Gerdien van Eersel

ISBN: 978-94-6299-843-8

Printing: Ridderprint BV - www.ridderprint.nl
Lay-out: Nikki Vermeulen - Ridderprint BV

Cover illustration: *Predestination* (2015) by Minjung Kim (Gwangju, 1962)

The research presented in this dissertation was funded by the Netherlands Organisation for Scientific Research (NWO project number: 411-10-912)

Limited Effectiveness of Retrieval Practice for the Transfer of Knowledge

De geringe effectiviteit van retrieval-gestuurd leren voor de transfer van kennis

Proefschrift

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

De openbare verdediging zal plaatsvinden op
donderdag 29 maart 2018 om 13:30

door

Gerdien van Eersel
geboren te Rotterdam.

Promotiecommissie

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

Overige leden:
Prof. dr. F. Paas
Prof. dr. L. Kester
Prof. dr. L.R. Arends

Copromotor:
Dr. P.P.J.L. Verkoeijen

Voor El Charoy

It is a mistake to suppose that acquisition of skills in reading and figuring will automatically constitute preparation for their right and effective use under conditions very unlike those in which they were acquired.

John Dewey

Contents

1

General introduction

In 2014, my dear nephew El Charoy had to give a presentation at school. He chose to talk about his favorite animals: penguins. First, he designed his PowerPoint presentation and then he wrote out the text. Subsequently, he started to learn the text by heart, which intuitively he did by reading and rereading the text. However, as a cognitive researcher, I advised him not to reread, but to *retrieve* the text from memory without looking at it. To improve retrieval even more, we made a note with keywords that he could use during retrieval. Furthermore, whenever he doubted his memory, he looked back at the text in order to correct himself. The most successful moments of learning occurred when he practiced his speech out loud with the possibility to look back at the text, in order to correct possible errors. In other words, it turned out that the best and fastest way for him to remember the text in the long term was by performing *retrieval practice with feedback*. The strategy worked perfectly well: El Charoy knew his text in no time and gave an awesome presentation, which was rewarded with a 9.5 (A+)!

Research into the retrieval practice effect

The learning strategy that El Charoy employed is termed retrieval practice, which can be described as retrieving information from memory after an initial learning phase. The beneficial effect of this strategy over restudy on long-term learning has been shown to be robust and is dubbed the (retrieval practice) *testing effect* (for reviews, Delaney, Verkoeijen, & Spirgel, 2010; Karpicke, Lehman, & Aue, 2014; Roediger & Butler, 2011; Roediger & Karpicke, 2006a; Rowland, 2014). In a typical retrieval practice experiment, participants repeat a set of initially studied stimuli either by restudying or by retrieval practice, with exposure time equated for the two conditions. After a certain retention interval, they receive a final criterion test. When no feedback on their memory performance is provided during the intervening test, performance for restudied stimuli is generally better than, or comparable to, performance for tested stimuli after a short interval. However, after a longer interval (i.e., one day or more), retrieval practice is typically more effective than restudy, giving rise to a cross-over interaction effect between study condition and retention interval (e.g., Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011). Several studies have also found retrieval practice effects after a short interval (e.g., Carpenter, 2009; Carpenter & DeLosh, 2005, 2006; Karpicke & Zaromb, 2010; Rowland & DeLosh, 2015), especially when a restudy opportunity was provided after the retrieval practice phase (i.e., feedback) (e.g., Bishara & Jacoby, 2008; Carpenter, Pashler, Wixted, & Vul, 2008; Carrier & Pashler, 1992; Jacoby, Wahlheim, & Coane, 2010; Kang, 2010; Kornell, Hays, & Bjork, 2009; Wartenweiler, 2011) or when retrieval success in the initial test was relatively high (Karpicke & Zaromb, 2010; Smith, Roediger, & Karpicke, 2013).

Although the first research into the advantages of retrieval practice already took place at beginning of the 20th century (e.g., Gates, 1917; Jones, 1923–1924), the topic regained large scientific interest after the publication of a review (Roediger & Karpicke, 2006) and an empirical study about the retrieval practice effect (Roediger & Karpicke, 2006, Experiment 1). In the latter, participants had to read two prose passages about the sun and about sea otters. They then wrote down as much of the material as they could remember from one of the passages (free recall retrieval practice), and reread the other passage (restudy). The final free recall test was administered after five minutes, two days or one week. It turned out that restudy led to better performance than free recall in the five-minutes condition, while a free recall retrieval practice advantage emerged in the two-days and one-week conditions.

The retrieval practice effect has been demonstrated across a wide range of practice tests, such as cued-recall, recognition, free recall, fill-in-the-blank, and short answer questions (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). Retrieval practice formats such as free recall and cued recall are generally more beneficial for performance than fill-in-the-blank, multiple-choice and recognition tests (e.g., Butler & Roediger, 2007; Dunlosky et al., 2013; Glover, 1989; McDaniel, Anderson, Derbish, & Morrisette, 2007; McDaniel, Roediger & McDermott, 2007; Roediger & Karpicke, 2006; Rowland, 2014). Furthermore, the magnitude of the retrieval practice effect depends on initial retrieval success and on the possibility of re-exposure to the material after retrieval practice (e.g. Rowland, 2014). If only a limited number of items is retrieved in the initial test, the advantage of retrieval practice will be limited, especially when there is no restudy opportunity. Indeed, a meta-analysis (Rowland, 2014) showed that the magnitude of the retrieval practice effect increased with initial retrieval performance, and that no retrieval practice effect occurred when initial performance was below 50%. Studies that did include a phase of re-exposure to the material after retrieval practice (i.e. *feedback*) delivered the largest retrieval practice effects, regardless of retrieval success.

A substantial part of the retrieval practice research has focused on relatively simple study materials, such as words or word pairs (e.g., Coppens, Verkoeijen, Bouwmeester, & Rikers, 2016; Hogan & Kitsch, 1971; Wheeler, Evans, & Buonanno, 2003), simple facts (e.g., Butler, Karpicke, & Roediger, 2008; Carpenter, Pashler, & Cepeda, 2009; Carpenter, Pashler, Wixted, & Vul, 2008; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011), locations on maps (e.g., Carpenter & Pashler, 2007; Rohrer, Taylor, & Sholar, 2010), animations (Johnson & Mayer, 2009), and symbols (Coppens, Verkoeijen, & Rikers, 2011). There has been increasing interest in materials that are more educationally relevant, like expository texts (e.g. Glover, 1989; Kang, Roediger, & McDermott, 2007; Nungester & Duchastel, 1982). Rowland (2014) found in a meta-analysis that associated word pairs yield the largest retrieval practice effects, followed by prose and then followed by single words.

Benefits of retrieval practice have emerged on different types of criterion tests, such as recognition, multiple choice, cued recall, and free recall (for overviews, see Roediger & Karpicke, 2006; Rowland, 2014). The literature suggests that retrieval practice effects might be smaller with recognition or multiple-choice tests than with cued recall or free recall final tests (Halamish & Bjork, 2011; Rowland, 2014). Only a small number of studies has shown retrieval practice to produce better performance on a test that measures related but new knowledge, i.e., *transfer* of knowledge (e.g., Blunt & Karpicke, 2014; Butler, 2010; Eglington & Kang, in press; Foos & Fisher, 1988; Hinze, Wiley, & Pellegrino, 2013; Johnson & Mayer, 2009; Karpicke & Blunt 2011; McDaniel, Howard, Einstein, 2009; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013). Transfer can be broadly defined as the ability to apply previously learned knowledge or skills in a novel context and to solve new problems (e.g., Salomon & Perkins, 1989), and might be considered to be the aim of learning (Carpenter, 2012; Rohrer, Taylor, & Sholar, 2010).

Explanations of the retrieval practice effect

Several theories (for overviews, see Delaney, Verkoeijen, & Spirgel, 2010; Karpicke, Lehman, & Aue, 2014; Rowland, 2014) have been proposed to explain the beneficial effect of retrieval practice. Note that these theories are not mutually exclusive.

A first possible explanation might be found in the *bifurcation framework* (Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011). According to this framework, a test bifurcates the distribution of items' memory strength: memory traces of non-retrieved items remain low in strength while the memory traces of retrieved items become high in strength, resulting in a gap between the two sets of items. Furthermore, items that are restudied are strengthened more in memory than non-retrieved items (but less than retrieved items). Because strong memories last, retrieval practice will result in better performance than restudying after an interval that is long enough for only the strongest memories (i.e., the memory representations of items that were retrieved during retrieval practice) to survive. Together this also implies that when a small number of items is retrieved in the initial test, the benefit of retrieval practice will be limited.

A second theory explaining the retrieval practice effect is based on the idea that when an item is successfully retrieved from memory, the representation of that item in memory is strengthened (Bjork, 1975). Moreover, the more effort it requires to retrieve the item initially, the stronger the resulting memory trace. Evidence for this *retrieval effort hypothesis* was provided by e.g., Pyc and Rawson (2009), who showed that the more difficult the initial (successful) retrieval, the higher was the final test performance. Because retrieval requires more effort than restudy, retrieved items are strengthened more in memory than restudied items. This results in better retrieval at a later point in time, especially after a long interval. The retrieval effort hypothesis is consistent with

several empirical findings (e.g., Carpenter & DeLosh, 2006; Karpicke & Roediger 2007), for example that initial free recall tests –which are assumed to require more effort– tend to produce larger retrieval practice effects than do recognition or multiple-choice tests (e.g., Butler & Roediger, 2007; Dunlosky et al., 2013; Glover, 1989; McDaniel, Anderson, Derbish, & Morrisette, 2007; McDaniel, Roediger & McDermott, 2007; Roediger & Karpicke, 2006; Rowland, 2014). The retrieval effort hypothesis does not make specific assumptions about the mechanisms through which retrieval effort produces the retrieval practice effect, but it is possible that retrieval effort adds an extra information element to a memory trace (i.e., encoding variability), or that retrieval effort activates an elaborative structure of related concepts (Delaney, Verkoeijen, & Spirgel, 2010).

A third category of explanations for the retrieval practice effect can be labelled as the *elaboration framework* (e.g. Carpenter, 2009, 2011; Carpenter & Yeung, 2017; Pyc & Rawson, 2010; Rawson, Vaughn, Carpenter, 2015), which proposes that retrieval practice induces students to bring to mind semantically related words, i.e., semantic elaboration. When retrieving a target, information that is semantically related to the cue is activated, and becomes linked to the target. As a result, the number of retrieval cues is larger for tested items than for restudied items, which in turn leads to a retrieval practice benefit on a final memory test. Specifically, the information that is activated through retrieval practice is likely to have the form a mediator, linking the cue to the target (Carpenter, 2011; Pyc & Rawson, 2012). For example, when attempting to retrieve the target word *bread* when given the (weakly) associated cue word *basket*, several words associated with the cue are activated, such as *flour* and *breakfast*, which then form semantic mediators between *basket* and *bread*. On a later test, these mediators then serve as additional retrieval cues, enhancing the retrieval of tested items relative to restudied items. Also, weakly associated cues are assumed to trigger more semantic elaboration than strongly associated cues because they require a more extensive search for the target (Carpenter, 2009), and therefore lead to larger retrieval practice effects. The elaboration theory can explain the typical interaction effect between study condition and retention interval, with the advantage of retrieval practice emerging only after a longer retention interval. Because of the *semantic* organization of long-term memory (e.g., Bartlett, 1932) versus the more perceptual organization of short-term memory (e.g., Baddeley, 1976), semantic information is more likely to serve as a long-term retrieval cue than perceptual information. As retrieval practice is assumed to activate semantically related words, retrieved information will be better remembered in the long term than restudied information (Carpenter, 2011).

A specific account within the elaboration framework is the fuzzy trace explanation of the retrieval practice effect (Bouwmeester & Verkoeijen, 2011; Verkoeijen, Bouwmeester, & Camp, 2012). The central idea of the fuzzy trace theory is that information is stored on

two different types of memory traces: verbatim/surface and gist traces. Verbatim traces are representations of a memory target's literal, contextual, and item-specific surface features. Gist traces, on the other hand, are representations of semantic, relational, and other elaborative information about a target. According to the fuzzy trace explanation of the retrieval practice effect, restudying an item strengthens its verbatim memory traces more than retrieval practice does. By contrast, retrieval practice is assumed to activate the gist memory traces of the item, because people mainly use semantic cues to retrieve information from memory. This theory can explain that the superior memory performance for tested items typically emerges after a long retention interval. That is, several studies (e.g., Anderson, 1974; Kintsch, Welsch, Schmalhofer, & Zimny, 1990; Sachs, 1967) have shown that information on verbatim traces decays more rapidly from memory than information on gist traces. The fuzzy trace explanation for this finding is that verbatim traces are more sensitive to sources of interference than gist traces, and therefore do not consolidate as much as gist traces (Brainerd & Reyna, 2004). Hence, after a short retention interval of several minutes, people can retrieve information from memory based on surface traces, gist traces, or a combination of both. After a retention interval of multiple days, however, they need to rely almost exclusively on gist traces. Because retrieval practice strengthens gist traces more than restudy, the retrieval practice effect is stronger after a multi-day retention interval than after a short retention interval of several minutes.

As a fourth and final explanation of the retrieval practice effect, Karpicke and colleagues (2014) proposed *the episodic context account*. First, they assume that information about an item is stored in memory as a representation with lexical/semantic item features and temporal context features. When learners later retrieve this knowledge, they reinstate the prior (temporal) context in which the information was learned. If retrieval is successful, the context associated with that knowledge is updated to include features of the original study context and features of the present test context. When learners again try to retrieve the information on a later test, the updated context representation serves as an effective retrieval cue, and also allows learners to restrict their search set. This results in better memory performance as compared to situations in which people did not practice retrieval. The episodic context account can also explain why *effortful* retrieval tasks lead to high memory performance, for example the finding that weakly related cues produce larger retrieval practice effects than strongly related cues, and the finding that initial free recall tests generally lead to better performance than do multiple-choice or recognition tests (Glover, 1989; Roediger & Karpicke, 2006). That is, according to the episodic context account, effortful retrieval tasks are also the ones that require learners to engage in higher levels of context reinstatement (Karpicke, Lehman, & Aue; Whiffen & Karpicke, 2017).

Note that Karpicke and colleagues (2014) explain the interaction effect between study condition and retention interval in terms of an item-selection effect. In the restudy condition, learners are re-exposed to more items than in the retrieval practice condition. If retrieval success were nearly perfect, however, item exposure would be comparable across restudy and retrieval practice conditions, and no benefit of restudy in the short term would occur. For this reason, Karpicke and colleagues consider this interaction effect not to be in need of further theoretical explanation.

Across the boundaries?

Nunes and Karpicke (2015) stated that "to integrate findings from cognitive science with educational practice, at a minimum, researchers must use authentic educational materials, tasks that would be plausible in educational settings, and assessments that are relevant to real-world learning outcomes." As for the first educational goal of using authentic educational materials, an increasing number of retrieval practice studies has focused on authentic classroom material, like informative video (Cranney, Ahn, McKinnon, Morris, & Watts, 2009), interactive teaching sessions (Larsen, Butler, & Roediger, 2009), art history lectures (Butler & Roediger, 2007), and a teacher's lesson with a middle school textbook chapter (McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013). Furthermore, in a non-exhaustive literature review, we found twenty-five studies that demonstrated a benefit of retrieval practice over restudy with *expository text* (see Chapter 5). However, already in 1917, Gates wrote that "the advantage of recitation over reading is greater in learning senseless, non-connected material than in learning senseful, connected material" (p. 23). More recently, Van Gog and Sweller (2015) also claimed that the retrieval practice effect decreases when the complexity of the study material increases. Complexity is defined in terms of element interactivity: material is complex when its individual elements are related and must therefore be processed simultaneously. However, in a critical response, both Karpicke and Aue (2015) and Rawson (2015) concluded that several studies have shown that the retrieval practice effect arises frequently with coherent/integrated materials as well. Taken together, there seems to be some discussion about whether retrieval practice is particularly useful for learning coherent material like expository text.

The second goal set by Nunes and Karpicke (2015) was to measure outcomes that are relevant to real-world learning settings. Now, most retrieval practice research involved final tests that only assessed retention (Roediger & Butler, 2011; Rowland, 2014), whereas less is known about the effect of retrieval practice on tests that measure *transfer* of knowledge (e.g., Barnett & Ceci, 2002; Butler, 2010; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013). According to Mayer (1996), transfer of knowledge reflects knowledge *understanding,* and involves integrating and organizing the studied

information into a coherent and meaningful mental representation. Understanding means to grasp the way in which one's beliefs in the propositions of interest cohere with other propositions one believes (Kvanvig 2003), as well as the structural relationships between the central pieces of information (Kvanvig, 2009).

As mentioned above, only a small number of studies has shown retrieval practice to produce better performance on a final transfer test (Blunt & Karpicke, 2014; Butler, 2010; Eglington & Kang, in press; Foos & Fisher, 1988; Hinze, Wiley, & Pellegrino, 2013; Johnson & Mayer, 2009; Karpicke & Blunt 2011; McDaniel, Howard, Einstein, 2009; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013). This might suggest that the retrieval practice effect is weaker with final transfer tests (e.g., Tran, Rohrer & Pashler, 2014). Possibly, because transfer tests mainly follow from coherent/integrated study material, for which the advantage of retrieval practice may be limited, the retrieval practice effect might be limited for transfer final tests as well.

In this thesis, we will further explore whether retrieval practice effects are obtained with expository text and final tests that measure transfer. Moreover, we will look into the additional effect of providing a re-exposure opportunity after retrieval practice (i.e., feedback). Feedback enhances the beneficial effect of retrieval practice (e.g., Rowland, 2014) because it enables learners to correct errors and to improve metacognitive monitoring (e.g., Butler, Karpicke, & Roediger, 2008; Pashler, Cepeda, Wixted, & Rohrer, 2005). Such feedback will make subsequent study more effective.

Thesis outline

The present thesis consists of five experimental studies that addressed the following research questions. Firstly, does a retrieval practice benefit emerge with expository text on a final test that measures transfer? This question was the aim of Chapter 3, 4, and 6. Secondly, does a retrieval practice benefit occur with the same expository texts on a final pure memory text? This second question was the focus of Chapter 5. Thirdly, does a phase of re-exposure to the material after retrieval practice (i.e., *feedback*) boost the beneficial effect of retrieval practice? This third question was addressed in Chapter 5 and 6. As an additional research goal, in Chapter 3, retrieval practice was compared to *self-explanation* on a final transfer test. Finally, in Chapter 2, as a second additional research goal, the fuzzy trace theory of the retrieval practice effect (Verkoeijen, Bouwmeester, & Camp, 2012) was assessed, using simple words as study material and a final test measuring (near) transfer.

The central idea of the fuzzy trace theory is that information is stored on two types of memory traces: verbatim and gist traces. Also, rereading mostly strengthens verbatim traces while retrieval practice mostly strengthens gist traces. Therefore, the fuzzy trace theory predicts that a retrieval practice effect will emerge when people cannot use

verbatim/surface cues in a final test, and have to rely exclusively on semantic/gist cues instead. In **Chapter 2**, this prediction was tested by gradually reducing the surface features overlap between cues in the learning phase and the final recognition test over five experiments. The experimental final tests consisted of scrambled words, words in a new context, scrambled words in a new context, synonyms, or images. Such final tests, with only semantic cues available, can be considered to measure (near) transfer, because the surface cues that were present during learning were absent in the final test. To recognize an item, the learner had to rely on cues that (partly) differed from the cues that were available in the learning phase, which can be regarded as a form of transfer.

In **Chapter 3**, we compared retrieval practice to self-explanation on a final transfer test. As text material we used an argumentative text that had been part of the Dutch secondary school exams. Participants first read the text and then completed the read-recite-review (RRR) condition, the self-explanation condition, or the baseline control condition. Participants in the RRR-condition first read a paragraph, then recited as much as possible, and afterwards read the paragraph again. Participants in the self-explanation condition had to clarify and explain the central ideas of each of the paragraphs. In the baseline control condition, participants only read the text for the first time like in the other conditions, and then immediately performed the final comprehension test. Participants in the RRR-condition and in the self-explanation condition received this final test immediately after their study phase. The comprehension test comprised of twelve open-book multiple-choice questions that assessed the extent to which learners had grasped the explanatory connections and logical implications of the argument at issue.

The aim of **Chapter 4** was to investigate whether two different types of retrieval practice would benefit performance as compared to rereading on a final transfer test. Participants read four expository texts and then engaged in either verbatim free recall, constructive recall, or rereading. In the verbatim free recall condition, participants were asked to type verbatim everything they could remember from the texts. In the constructive recall condition, participants were instructed to type in their own words what they had comprehended from the content of the texts. The final test consisted of sixteen closed-book short-answer inference questions, administered immediately and after a one-week delay. These inference questions were aimed at measuring inferences going beyond what was stated in the text; participants had to apply the acquired information to a new situation (i.e., transfer of knowledge).

In **Chapter 5**, participants read two of the four expository texts that were also used in Chapter 4. As a final test we used a free recall test that measured pure free recall memory. With this kind of memory test, the beneficial effect of retrieval practice over restudying has been well established. After first reading the texts, participants reread

one of the texts and performed free recall retrieval practice on the other text, the latter amounting to retrieving as much as possible from the text. Immediately or after a one week delay, the final test was administered. Importantly, in the first experiment, no re-exposure opportunity was provided after free recall retrieval practice. In the second experiment, however, a re-exposure opportunity did follow the retrieval practice phase.

In **Chapter 6**, we examined whether a different type of retrieval practice, namely cued recall questions with feedback, would benefit transfer performance as compared to rereading. The first experiment was a direct replication of the third experiment by Butler (2010), which had been the only study so far to show a retrieval practice testing effect on a final transfer test tapping onto a different knowledge domain. Participants studied expository texts and then either reread them three times or went through three cycles of cued recall questions (i.e., retrieval practice) with feedback. The second experiment was similar to the first, but now with an extra reread-plus-statements condition. In this condition, participants only received focused exposure to the key information after they had reread a text. This key information was identical to the feedback in the cued recall condition. In this way we investigated whether this feedback could – partly – account for the retrieval practice effect found in Butler (2010).

Chapter 7 summarizes and discusses the main findings of this thesis in relation to existing literature. Finally, some directions for future research are provided.

2

Does retrieval practice depend on semantic cues? Assessing the fuzzy trace account of the testing effect

This chapter has been published as:
Van Eersel, G. G., Bouwmeester, S., Verkoeijen, P. P. J. L., Tabbers, H. K. & Rikers, R. M. J. P. (2017). Does retrieval practice depend on semantic cues? Assessing the fuzzy trace account of the testing effect. *Journal of Cognitive Psychology, 29,* 583–598.
doi: http://dx.doi.org/10.1080/20445911.2017.1300156

Abstract

Retrieval practice enhances long-term retention more than restudying; a phenomenon called the testing effect. The fuzzy trace explanation predicts that a testing effect will already emerge after a short interval when participants are solely provided with semantic cues in the final test. In the present study, we assessed this explanation by gradually reducing the surface features overlap between cues in the learning phase and the final recognition test. In all five experiments, participants in the control/word condition received as final test cues the same words as in the learning phase. The experimental final test cues consisted of scrambled words, words in a new context, scrambled words in a new context (Experiment 1), synonyms (Experiment 2), or images (Experiments 3, 4a, 4b). A short-term testing effect was only observed for the image final test cues. These results do not provide strong support for the fuzzy trace explanation of the testing effect.

The testing effect occurs when retrieving information from memory after an initial study phase enhances long-term retention more than restudying does (for reviews, see Delaney, Verkoeijen, & Spirgel, 2010; Karpicke, Lehman, & Aue, 2014; Roediger & Butler, 2011; Rowland, 2014). In a typical testing effect experiment, participants learn a set of words during an initial study phase either by restudying or by testing (i.e., retrieval practice). After a certain retention interval, they receive a final test. When no feedback on their memory performance is provided during the intervening test, performance for restudied stimuli is generally better than, or comparable to, performance for tested stimuli after a short interval of five minutes (exceptions can be found in, for example, Carpenter, 2009; Halamish & Bjork, 2011). However, after a long interval (generally one week), retrieval practice is more effective than restudy, giving rise to an interaction effect of study method and retention interval on memory performance (e.g., Hogan & Kintsch, 1971; Roediger & Karpicke, 2006; Wheeler, Evans, & Buonanno, 2003). This testing effect has been demonstrated under a variety of practice tests, such as cued-recall, recognition, free recall, fill-in-the-blank, and short answer questions (Dunlosky, Rawson, Marsh, Nathan, & Willingham 2013).

Several theories (for overviews, see Delaney, Verkoeijen, & Spirgel, 2010; Karpicke, Lehman, & Aue, 2014; Rowland, 2014) have been proposed to explain the mechanism underlying testing effect. One category of explanations can be classified as elaboration theories (e.g. Carpenter, 2009, 2011; Pyc & Rawson, 2010), which propose that retrieval practice induces more semantic elaboration of a memory trace than restudy. When retrieving a target, information that is semantically related to the cue is activated, and becomes linked to the target. As a result, the number of retrieval routes is larger for tested items than for restudied items, which in turn leads to a testing benefit on a final memory test administered after a long retention interval.

Verkoeijen, Bouwmeester, and Camp (2012; see also Bouwmeester & Verkoeijen, 2011) postulated another explanation of the testing effect within this category, based on the fuzzy trace theory (Brainerd & Reyna, 2004). The central idea of the fuzzy trace theory is that information is stored on two different types of memory traces: verbatim/surface and gist traces. Verbatim traces are representations of a memory target's literal, contextual, and item-specific surface features. Gist traces, in contrast, are representations of semantic, relational, and other elaborative information about a target. The empirical support for the distinction between verbatim and gist traces is extensive (for an overview, see Brainerd & Reyna, 2004). According to the fuzzy trace explanation of the testing effect (Verkoeijen, Bouwmeester, & Camp, 2012), restudying an item strengthens its verbatim memory traces more than retrieval practice does. By contrast, retrieval practice is assumed to activate the gist memory traces of the item, because people mainly use semantic cues to retrieve information from memory.

The fuzzy trace account can explain an important boundary condition of the testing effect, namely that the superior memory performance for tested items typically emerges after a long retention interval. Several studies (e.g., Anderson, 1974; Kintsch, Welsch, Schmalhofer, & Zimny, 1990; Sachs, 1967) have shown that information on verbatim traces decays more rapidly from memory than information on gist traces. The fuzzy trace explanation for this observation is that verbatim traces are more sensitive to sources of interference than gist traces, and therefore do not consolidate as much as gist traces (Brainerd & Reyna, 2004). Hence, after a short retention interval of several minutes, people can retrieve information from memory based on surface traces, gist traces, or a combination of both. After a retention interval of multiple days, though, they need to rely almost exclusively on gist traces. Because, according to the fuzzy trace explanation, retrieval practice is assumed to strengthen gist traces more than restudy, the testing effect is stronger after a multi-day retention interval than after a short retention interval of several minutes.

An interesting prediction that follows from the fuzzy trace account of the testing effect is that a testing effect will be obtained after a short retention interval (i.e., 'short-term testing effect') when people cannot use verbatim/surface cues in a final test, and have to rely exclusively on semantic cues instead. Verkoeijen, Bouwmeester, and Camp (2012) tested this prediction. A group of 64 Dutch psychology undergraduates was asked to study 12 Dutch Deese-Roediger-McDermott word lists (DRM: Deese, 1959; Roediger & McDermott, 1995) either by restudying or by testing. Each list consisted of eight words, each of which had a strong backward association with one semantically related distractor. Immediately after the learning phase, the participants took a final yes-no recognition test in Dutch (within-language condition) or in English (across-language condition). The participants were bilingual with respect to the English final-test words in the across-language condition. It was assumed that participants in the within-language condition were cued with both semantic and verbatim/surface information (i.e., the visual appearance of a word) of the studied words. By contrast, in the across-language condition, surface features of the previously studied words were unavailable, so participants were only cued with semantic information. As indicated, according to the fuzzy trace account of the testing effect, recognition of restudied items depends more strongly on surface cues than the recognition of tested items. The recognition of tested items, on the other hand, hinges more strongly on semantic cues. For that reason, the fuzzy trace theory predicts a testing effect to arise in the across-language but not in the within-language condition. The results of Verkoeijen and colleagues' experiment (2012) were in line with these predictions. The proportion of correctly recognized items was higher for tested items (.78) than for restudied items (.67) in the across-language condition, but did not differ between tested items (.78) and restudied items (.81) in the

within-language condition. In other words, there was a short-term testing effect in the across-language condition but not in the within-language condition.

As outlined above, the fuzzy trace account proposes that the short-term testing effect found by Verkoeijen and colleagues (2012)[1] emerged because the final test recognition of restudied items, but not of tested items, suffered from the lack of surface features overlap between the items in the learning phase and the cues in the final test. In the present study, we assessed this fuzzy trace account by gradually reducing the degree of surface features overlap between the items in the learning phase and the items in final test over five experiments. In this way it was possible to examine whether a testing effect would occur when there was small surface features overlap. In line with the fuzzy trace account, we expected that the smaller the surface features overlap, the larger the benefit of testing over restudy.

In all five experiments, participants studied a list of unrelated words through restudying or through testing. The crucial manipulation took place at the final yes/no recognition test, which was administered five minutes after the learning phase. In the first study, the factor 'surface features overlap' differed in the extent to which there was a surface features overlap between the words in the learning phase and the cues in the final test. The factor had four levels: word, scrambled, background, and background scrambled. In the word condition, the final test cues were the studied targets (plus distractors) presented in exactly the same manner as in the learning phase, thereby guaranteeing a maximum overlap of surface cues between the learning phase and test phase. This condition was similar to the within-language condition of Verkoeijen and colleagues (2012). In the scrambled condition, scrambled versions of the words were presented, and here the surface features overlap between the words from the learning phase and the items in the final test was still considerable. In the background condition, words were vertically presented in a different font, at the top right of the computer screen on a colorful, flowered background. In the background scrambled condition, scrambled versions of the words were vertically presented in a different font, at the top right of the screen, on a colorful and flowered background. The surface features overlap in the latter condition was smaller than in the other three conditions. In the second experiment, there were two versions of the final recognition test: words and synonyms. In the synonym condition, synonyms of the target words were shown. In the last three experiments, the final recognition test consisted of either the target words or images of the target words, with the latter being purely semantic cues.

In general, we predicted that the smaller the surface features overlap between the words in the learning phase and the cues in the final test, the larger the benefit

1 Carpenter (2011) and Rawson, Vaughn, and Carpenter (2015) have also used semantic final test cues yet within a cued-recall setting, which differs from the recognition memory framework that is of interest in the current study.

of testing compared to restudying. For this reason, we expected that no testing effect would emerge in any of the five word conditions. Furthermore, in the first experiment, there was still some surface features overlap between the targets in the learning phase and the cues in the final tests. We therefore predicted the advantage of testing in the background condition and the scrambled condition to be small or absent, possibly just like in the background scrambled condition, where the manipulation was a bit stronger but still subtle. In the final tests of the synonym condition (Experiment 2) and the image conditions (experiment 3, 4a, 4b), the surface features of the studied targets were absent, so we predicted a large benefit of testing to actuate in these two conditions. In addition, we expected that the smaller the surface features overlap, the more difficult the final test, and the lower the overall performance on the targets as well as on the unrelated distractors.

Experiment 1

Method

Participants and design

One hundred eighty-three participants were recruited online through Amazon's Mechanical Turk (AMT) (http://www.mturk.com). Twelve participants were excluded on the basis of one of the three following criteria, resulting in a total of 171 participants. Firstly, a score lower than zero on the equation 'percentage correct on the targets minus percentage incorrect on the distractors'. In this case, participants choose the option 'old' (i.e., presented during the learning phase) more often when a word was new than when it was old. This indicates that they did not pay full attention to the task or that they (coincidentally) switched the response buttons. A second criterion for exclusion was a score of less than 30% correct on the distractors in the final test, since such a score of at least 20 percent points lower than chance level is also an indication of not giving full attention to the task or switching response buttons. Thirdly, participants were excluded when the log files showed that they had performed retrieval practice during the 2-min distractor task, because practicing when they were supposed not to study would have an undesired effect on the final test scores. If a log file contained one or more words from the last learning phase instead of numbers counted backwards, all data from that participant were discarded. For more information on the demographic characteristics of the AMT population, see Paolacci, Chandler, and Ipeirotis (2010), and Ross, Irani, Silberman, Zaldivar, and Tomlinson (2010). All participants were native English speakers and residents of the USA. They were paid $0.80 for their participation, which took about 25 minutes.

A 2 Study Method (restudy vs. testing) x 4 Surface Features Overlap (word vs. scrambled vs. background vs. background scrambled) mixed design was used with repeated measures on the first factor. Participants were randomly assigned to the levels of the between-subjects factor.

Materials

All material and data from this study can be retrieved from the Open Science Framework[2]. For the learning phase of Experiment 1, we selected 80 concrete, simple English nouns. Thirty-six words were used as 'targets' (i.e., they would later appear in the final recognition test), while the other 44 words in the learning phase were used as fillers (i.e., not appearing in the final test). Mean word frequency was determined using the SUBTLEXus database, and did not differ statistically between targets and fillers (1.38 ± 1.62 and 1.49 ± 0.46 lnLog per million, respectively). Also, mean word length did not differ statistically between targets (4.67 ± 1.16 letters) and fillers (4.68 ± 1.29 letters). There were ten lists of eight words, and these ten lists were randomly split into two sets of five lists. Then four study sequences were created by counterbalancing across presentation order of sets (set 1 first vs. set 2 first) and study method (testing first vs. restudy first), such that participants first received a set of five restudy lists and then a set of five test lists, or vice versa. The order of words within a list and the order of lists within a set were fixed.

The final recognition test consisted of 73 words: 36 target words and 37 unrelated distractors[3], the latter also being concrete English nouns. The words in the final test were randomly assigned to the serial positions, and the resulting test sequence was administered to all participants. In the word condition of the final test, words were presented in the same way (i.e., same font, letter size, letter type, and screen position) as they were during the learning phase. Hence, in this condition there was complete surface features overlap between the words in the learning phase and the words in the final test. In the scrambled condition, scrambled versions of the words were displayed. For example, the word "black" was presented as "cklba". Participants were asked to mentally unscramble the word and then to indicate whether the word had been presented during the learning phase. In the background condition, words were vertically presented in a different font at the top right of the computer screen on a colorful, flowered background. In the background scrambled condition, scrambled versions of the words were vertically presented in a different font, at the top right of the screen on a colorful and flowered background (see Appendix A for an example in black and white). As a result, the surface features overlap in the latter condition was minimal.

2 https://osf.io/nx3zm/

3 Due to a few programming errors, the numbers of targets and distractors were not always equal, and also slightly differed across test sessions and experiments.

Procedure

The experiment was programmed and presented in the Qualtrics survey research suite (http://www.qualtrics.com). Participants were first informed that they would be presented with ten lists of eight words to memorize. They then started with an initial study phase in which words were presented in the center of the computer screen at a 3.75-s rate. After each list, a free recall test was conducted (testing), or the list was presented again using the same procedure (restudy). In the free recall test, participants were asked to type in all the words that they could remember from the preceding study list. This free recall test took 30 seconds in total, which was equal to the duration of the restudy condition. The learning phase was followed by a 2-minutes distractor task in which participants counted backwards on a sheet of paper in steps of three from a given number. Subsequently, participants completed the final recognition test. This task was varied according to the levels of the factor 'surface features overlap': word, scrambled, background, and background scrambled. In all conditions, the final test required participants to indicate whether the word was old or new (i.e., presented in the previous learning phase or not). Words were presented one by one on the computer screen. In the scrambled condition and the background scrambled condition, participants were asked to first mentally unscramble the word and then to indicate whether the word was old or new. The final test was self-paced, and a new test item appeared after the participant had clicked on the next item button.

Results

The three outcome variables are the responses to the immediate free recall test, the unrelated distractors in the final recognition test and the targets in the final recognition test. Because these are all binomial count variables (correct=1 / incorrect=0), they were entered into a logistic regression analysis with a random intercept to deal with the dependence of the repeated measures. With this type of outcome variable, a regular ANOVA on the proportion of correct responses can lead to spurious findings because it might attribute probability mass to impossible values (i.e., values below 0 or above 1) and because the assumption of homogeneity is easily violated (Jaeger, 2008). The level of significance was set at $a = .05$.

Immediate test

There were no statistical[4] differences in the mean proportion correctly retrieved tested items in the learning phase between the word condition ($M = .73$, $SD = .15$), the scrambled condition ($M = .74$, $SD = .15$), the background condition ($M = .72$, $SD = .16$) and the background scrambled condition ($M = .73$, $SD = .18$), *Wald* (156) = 1.76, $p = .620$.

4 Following Kline (2004, Chapter 3) and Cumming (2014), we use the term 'statistically' instead of 'significantly', because the latter is often erroneously understood as meaning 'important'.

Final test performance

A 2 Study Method (restudy vs. testing) x 4 Surface Features Overlap (word vs. scrambled vs. background vs. background scrambled) logistic regression on the binomial count targets (see Table 1) did not reveal a statistical study method x surface features overlap interaction, *Wald* (162) = 4.63, p = .200. We did not find a statistical main effect of study method, *Wald* (162) = 1.33, p = .250. The regression coefficient (*b*) was -0.16, which is the log (ln) of the odds ratio between the restudy and the testing condition. The corresponding 95% confidence interval was [-0.42, 0.11], and the odds ratio for a correct answer was 0.85. This is a small effect size. An odds ratio of 0.85 means that the odds of a correct answer in the restudy condition is 0.85 times the odds of a correct answer in the test condition. The closer an odds ratio is to 1, the smaller the effect. Furthermore, there was a main effect of surface features overlap, *Wald* (162) = 46.41, $p <$ .001. Recognition performance was M = .80 (SD = .13) in the word condition, it was M = .72 (SD = .12) in the scrambled condition (odds ratio = 0.55), it was M = .80 (SD = .13) in the background condition (odds ratio = 0.86), and M = .61 (SD = .11) in the background scrambled condition (odds ratio = 0.34), with the word condition acting as the baseline category for the odds ratios, which are all small. Additionally, the mean proportion of correctly classified distractors differed between the word condition (M = .79, SD = .16), the scrambled condition (M = .75, SD = .15, odds ratio = 0.79), the background condition (M = .79, SD = .16, odds ratio = 0.96), and the background scrambled condition (M = .70, SD = .14, odds ratio = 0.61), *Wald* (144) = 46.08, $p <$ .001, with the word condition taken as the baseline category for the odds ratios, which are all small.

Table 1. Mean Proportion of Correctly Recognized Targets in Experiment 1 by Surface Features Overlap and Study Method. Standard Errors are between Brackets

	Surface Features Overlap			
Study Method	**Words**	**Scrambled**	**Background**	**Background Scrambled**
Restudy	.79 (.02)	.72 (.02)	.82 (.02)	.60 (.02)
Testing	.81 (.02)	.72 (.02)	.79 (.03)	.62 (.02)

Discussion

In our first experiment, we performed a subtle manipulation of the final test cues in order to assess the fuzzy trace explanation of the testing effect. According to this explanation, testing strengthens the gist traces of stimuli in memory, while restudying strengthens the surface traces. In Experiment 1, we varied the overlap between the surface features of presented words in the learning phase and the targets in the final test. By doing so, we sought to examine to what extent the surface features overlap had to fade in order for a short-term testing effect to emerge. We expected to find no advantage of testing in

the word condition, a small or no advantage of testing in the background condition and the scrambled condition, and a relatively larger advantage of testing in the background scrambled condition. However, we did not observe an interaction effect of study method and surface features overlap, that is, the difference between the recognition of tested and restudied items was very small in all four conditions. Apparently, there was still sufficient surface features overlap between the learning phase and the final tests of all four conditions. We therefore conducted a second experiment in which the surface manipulation of the final test cues was stronger, namely synonyms, resulting in a small surface features overlap between the learning phase and the final test words.

Experiment 2

Method

Participants and design

A total of 96 native English speaking participants were recruited online through AMT. They were paid $0.80 for participating, which required about 25 minutes. Five participants were excluded from this experiment on the basis of one of the criteria mentioned in the method section of Experiment 1, resulting in a total number of 91 participants.

A 2 Study Method (restudy vs. testing) x 2 Surface Features Overlap (words vs. synonyms) mixed design was used with repeated measures on the first factor. Participants were randomly assigned to the levels of the between-subjects factor.

Materials

For the learning phase of this experiment, we selected 80 new English nouns and adjectives. Thirty-six words were used as targets and the other 44 were used as fillers. The synonyms in the final test were selected on the basis of the Edinburgh Associative Thesaurus word association norms (http://www.eat.rl.ac.uk/), for example movie/film and pants/trousers. After initial selection, we verified on Merriam-Webster's Learner's dictionary whether the words were indeed regarded as synonyms. Mean word frequency was determined using the SUBTLEXus database, and did not differ statistically between targets and fillers (1.49 ± 0.72 and 1.64 ± 0.69 lnLog per million, respectively). Also, mean word length did not differ between targets (5.11 ± 1.43 letters) and fillers (4.61 ± 1.21 letters). The counterbalancing method was the same as in Experiment 1. The final recognition test consisted of 36 target words and 36 unrelated distractors. In the word condition, words were presented identical to the way they were presented in the learning phase. In the synonym condition, synonyms of the words were shown.

Procedure

The procedure was identical to that in Experiment 1, except that the scrambled condition, the background condition, and background scrambled condition were replaced by one synonym condition. In this condition, participants were asked to indicate whether a synonym of the word on the screen had been in one of the studied lists ('old or new'). Participants in the synonym condition were given the following instruction: "Next you will receive a test that consists of 72 words. For each word, you have to indicate whether a synonym of the word on the screen was in one of the lists you have just studied (yes) or not (no). For example, in the following test you see the word 'act'. If you have seen a synonym of 'act' in one of the lists you've studied, for example the word 'play', you answer yes. If you have not just studied a synonym of the word 'act', you answer no." We had ensured that none of the distractors in the final task was a synonym of one of the studied words.

Results

The three outcome variables and their analyses are the same as in Experiment 1.

Immediate test performance

The mean proportion correctly retrieved tested items during the learning phase did not statistically differ between the word condition (M = .70, SD = .21) and the synonym condition (M = .72, SD = .14), *Wald* (78) = 1.39, p = .240, regression coefficient -0.08, 95% confidence interval of the regression coefficient [-0.23, 0.06], odds ratio = 0.92. This is a small effect size.

Final test performance

A 2 Study Method (restudy vs. testing) x 2 Surface Features Overlap (words vs. synonyms) logistic regression on the binomial count targets (see Table 2) did not reveal a statistical interaction effect, *Wald* (86) = 0.62, p = .430, regression coefficient = -0.13, 95% confidence interval [-0.47, 0.20], odds ratio = 0.88. This is small effect size. This odds ratio of 0.88 means that in the word condition, the difference in odds of a correct answer between the tested and restudied words is 0.88 times this difference in the synonym condition. We did not find a statistical main effect of study method, *Wald* (86) = 3.22, p = .073, regression coefficient = 0.20, 95% confidence interval [-0.02, 0.43], odds ratio = 1.22. This is a small effect size. There was a statistical main effect of surface features overlap, *Wald* (86) = 12.88, p < .001, regression coefficient = 0.78, 95% confidence interval [0.35, 1.22], odds ratio = 2.18, with the proportion of recognized words being higher in the word condition (M = .77, SD = .17) than in the synonym condition (M = .67, SD = .16). This effect size is small. In addition, the mean proportion of correctly classified

distractors was higher in the word condition (M = .81, SD = .17) than in the synonym condition (M = .73, SD = .14), $Wald$ (70) = 28.55, p < .001, regression coefficient 0.45, 95% confidence interval [0.28, 0.62], odds ratio = 1.57. This effect size is small.

Table 2. Mean Proportion of Correctly Recognized Targets in Experiment 2 by Surface Features Overlap and Study Method. Standard Errors are between Brackets

	Surface Features Overlap	
Study Method	**Words**	**Synonyms**
Restudy	.78 (.03)	.69 (.03)
Testing	.77 (.03)	.65 (.03)

Discussion

In the second experiment, we expected an advantage of tested words compared to restudied words in the synonym condition but not in the word condition, since only in the former the surface cues of the studied words were unavailable. The results of the experiment were incongruent with these expectations. We did not find an interaction effect between the factors surface features overlap and study method, and even numerically there was no tendency in the hypothesized direction. These results were surprising, because the synonym condition was conceptually identical to the across-language condition in Verkoeijen, Bouwmeester, and Camp (2012). In the latter an interaction effect did occur, thereby substantiating the fuzzy trace account of the testing effect.

There is a possibility that the final test synonym cues did activate the surface representations of the studied words after all. Support for this idea might be found in studies using the lexical decision task (LDT). This task measures how fast participants can classify letter strings as words or nonwords, which can be used to show a *priming effect* — the implicit memory effect that exposure to one word influences the response time to another word. Several authors have claimed that the LDT mainly relies on *orthographic* or lexical processes (e.g., De Groot, 2002; Zeelenberg & Pecher, 2003). Now some studies (e.g., Perea & Rosa, 2002) have shown a masked priming effect for related synonym pairs on the LDT. When a word was presented between 66 and 166 ms (i.e., the prime) and then followed by its synonym (i.e., the target) in the lexical decision task, participants respond faster to the target than when the prime and the target were not related (Perea, & Rosa, 2002). However, studies have failed to find *across-language* repetition priming effects on the LDT, that is, when the targets are translations of the primes (e.g., Gerard & Scarborough, 1989; Kirsner, Brown, Abrol, Chadna, & Sharma, 1980; Scarborough, Gerard, & Cortese 1984; Zeelenberg & Pecher, 2003). This distinction might be due to the LDT primarily depending on orthographical or lexical processes. In support of this

claim, Zeelenberg and Pecher (2003) showed that when a *semantic* classification task was used instead of the LDT, cross-language priming did occur. Together these studies indicate that cross-language cues directly active their semantic representations, while synonyms activate their orthographic representations. This hypothesis would explain the discrepancy between the results of our second experiment and the cross-language testing effect observed by Verkoeijen, Bouwmeester, and Camp (2012).

In sum, it is possible that the final test words of Experiment 2 did activate the orthographic representations of their studied synonyms after all. We therefore conducted another experiment with *non-verbal* cues as final test cues, namely images. Research has shown that images primarily activate their semantic representations (e.g., Johnson, Paivio, & Clark, 1996). We consider the image cues to be the strongest of all manipulations, probably even stronger than the across-language condition in Verkoeijen, Bouwmeester, and Camp (2012), since the latter condition is still of a verbal nature. In the image final test condition, there is no surface features overlap between the studied words and the final test cues.

Experiment 3

Method

Participants and design

A total of 152 native English speaking participants were recruited online through AMT. They were paid $0.40 for their participation, which required approximately 25 minutes. Twelve participants were excluded on the basis of one of the criteria mentioned in the method section of Experiment 1, resulting in a number of 140 participants.

A 2 Study Method (restudy vs. testing) x 2 Surface Features Overlap (word vs. image) mixed design was used with repeated measures on the first factor. Participants were randomly assigned to the levels of the between-subjects factor.

Materials

For the learning phase we used the same words as in Experiment 1, except for six words that could not easily be translated into images. We replaced these six words by six other concrete nouns, resulting in a total number of 36 targets and 44 fillers. Mean word frequency was determined using the SUBTLEXus database, and did not differ statistically between targets and fillers (1.37 ± 0.50 and 1.47 ± 1.62 lnLog per million, respectively). Moreover, mean word length did not differ statistically between targets (4.66 ± 1.26 letters) and fillers (4.73 ± 1.20 letters). The counterbalancing method was the same as in Experiment 1. The final recognition test consisted of 35 or 36 target words –depending on the test session– and 38 unrelated distractors. The word-image combinations were

validated by asking six PhD candidates what the 73 images depicted (by free association, so without offering them any possible alternatives). Only if all six candidates mentioned the same object, the image was used. The images were obtained from the following website: http://users.skynet.be/taal/pictos/Page.html (see Appendix B for an example).

Procedure

The procedure of Experiment 3 was identical to that in Experiment 1, except that the scrambled condition, the background condition, and background scrambled condition were replaced by one image condition. In the image condition, images of the words were shown, and participants were asked whether the word that was represented by the image was old or new.

Results

The three outcome variables and their analyses are the same as in Experiment 1.

Immediate test performance

The mean proportion correctly retrieved tested items during the learning phase did not statistically differ between the word condition (M = .74, SD = .15) and the image condition (M = .72, SD = .17), *Wald* (78) = 2.16, p = .140, regression coefficient = -.09, 95% confidence interval [-0.21, 0.03], odds ratio = .91. This is a small effect size.

Final test performance

A 2 Study Method (restudy vs. testing) x 2 Surface Features Overlap (images vs. words) logistic regression on the binomial count targets (see Table 3) did not yield a statistical study method x surface features overlap interaction effect, *Wald* (135**) =** 0.62, p = .430, regression coefficient = -0.11, 95% confidence interval [-0.41, 0.18], odds ratio = 0.89. This odds ratio is of a small size. The analysis further showed a statistical main effect of surface features overlap, *Wald* (135) = 15.38, p < .001, regression coefficient = 0.61, 95% confidence interval [0.30, 0.92], odds ratio = 1.84. This effect size is small. The mean proportion of recognized words was higher in the word condition (M = .83, SD = .11) than in the image condition (M = .74, SD = .14). We did not find a statistical main effect of study method, *Wald* (135) = 2.63, p = .11, regression coefficient = 0.17, 95% confidence interval [-0.04, 0.39], odds ratio = 1.19. This effect size is small. In addition, the mean proportion of correctly classified distractors was higher in the word condition (M = .80, SD = .16) than in the image condition (M = .71, SD = .18), *Wald* (73) = 59.47, p < .001, regression coefficient = 0.50, 95% confidence interval [0.38, 0.63], odds ratio = 1.66. This is a small effect size.

Table 3. Mean Proportion of Correctly Recognized Targets in Experiment 3 by Surface Features Overlap and Study Method. Standard Errors are between Brackets

	Surface Features Overlap	
Study Method	**Words**	**Images**
Restudy	.82 (.01)	.72 (.02)
Testing	.83 (.01)	.75 (.01)

Discussion

In Experiment 3, a study method x surface features overlap interaction effect did not occur: there was a numerical advantage of testing compared to restudying in the image condition and also in the word condition. However, both simple effects were too small to be statistically significant. Accordingly, the overall final test performance did not differ between the restudy condition and the testing condition. These outcomes are not in line with the findings by Verkoeijen and colleagues (2012). They observed a benefit of testing over restudy with purely semantic final test cues in their across-language condition, but no difference between testing and restudy in their within-language condition.

The question then is what could be underlying the discrepancy between the results from our Experiment 3 and those from the across language condition in Verkoeijen and colleagues' study (2012). One possibility is that the findings differed because the participant pools, settings, and procedures differed as well. That is, the study by Verkoeijen et al. (2012) was performed by Dutch psychology undergraduates in the laboratory at the Erasmus University Rotterdam, rather than by Amazon's Mechanical Turk workers anonymously at home, as in our Experiment 3 (as well as in our Experiments 1 and 2). Furthermore, there were three small differences between the procedure by Verkoeijen and colleagues (2012) and the procedure on AMT in our first three experiments (see the procedure below). However, it should be noted that there are no theoretical reasons as to why any of these differences –or a combination of some of these differences– should influence the testing effect. Nevertheless, we decided to repeat Experiment 3 in the laboratory with Dutch psychology undergraduates and Dutch materials, using exactly the same procedure as Verkoeijen, Bouwmeester and Camp (2012). The only difference was in the final test cues, which were images instead of non-cognate translations. Since the outcomes of the first laboratory experiment (Experiment 4a) supported the fuzzy trace theory of the testing effect, we repeated this experiment (Experiment 4b) to see if its results were robust.

Experiment 4a and 4b

Method

Participants and design

The participants were 60 (Experiment 4a) and 61 (Experiment 4b) Dutch undergraduates from the Erasmus University Rotterdam, the Netherlands, who were rewarded with course credits or €5.

A 2 Study Method (restudy vs. testing) x 2 Surface Features Overlap (words vs. images) mixed design was used with repeated measures on the first factor. Participants were randomly assigned to the levels of the between-subjects factor.

Materials

We used the same words as in Experiment 3, except that we replaced six words that were not easily translatable into Dutch. There were 36 targets and 44 fillers. Mean word frequency was determined using the Dutch CELEX database, and did not differ statistically between targets and fillers (1.41 ± 0.62 and 1.41 ± 0.62 lnLog per million respectively). Also, mean word length did not differ statistically between targets (4.3 ± 1.03 letters) and fillers (4.2 ± 1.00 letters). The counterbalancing method was the same as in Experiment 1. The final recognition test consisted of 36 target words, and 37 (Experiment 4a) or 36 (Experiment 4b) unrelated distractors.

Procedure

Experiment 4a and 4b were programmed and presented in E-Prime software and conducted at the Erasmus University Rotterdam. The procedures were identical to that of Verkoeijen, Bouwmeester, and Camp (2012). Participants were first informed that they would be presented with ten lists of eight words, and they were asked to memorize these words. They then started with an initial study phase in which words were presented in the center of the computer screen at a 4-s rate with a 1-s interstimulus interval (cf. the 3.75-s rate of Experiment 1, 2, and 3). In this initial study phase, participants were instructed to type in each word and memorize it (cf. Experiments 1, 2, and 3, where typing in the words was not required). After each list, participants engaged in free recall or restudy. The restudy phase was identical to the initial study phase. In the free recall phase, participants were asked to type in all words that they could remember from the preceding study list. Free recall time was divided into eight periods of four seconds, with a 1-s interval between periods (cf. Experiment 1, 2, and 3, where all remembered words were typed during one uninterrupted period). In this way, free recall time was equally distributed over the words, to make the procedure more similar to the restudy condition. The total test time added up to forty seconds in total, which was equal to the

time-on-task in the restudy condition. Participants completed the 2-minutes distractor task and afterward the final recognition test, which was similar to the final task in the previous experiments. Participants were asked whether the word was old or new. A new test item appeared after the participant gave a response (instead of after clicking on a button, as in Experiments 1, 2, and 3).

Results Experiment 4a

The three outcome variables and their analyses are the same as in Experiment 1.

Immediate test performance

The mean proportion of correctly retrieved tested items during the learning phase did not statistically differ between the word condition (M = .74, SD = .15) and the image condition (M = .76, SD = .10), *Wald* (54) = 1.36, p = .240, regression coefficient = -0.11, 95% confidence interval [-0.31, 0.08], odds ratio = 0.90. This is a small effect size.

Final test performance

A 2 Study Method (restudy vs. testing) x 2 Surface Features Overlap (images vs. words) logistic regression on the binomial count targets (see Table 4) showed a statistical study method x surface features overlap interaction effect, *Wald* (55) = 7.09, p = .008, regression coefficient = 0.62, 95% confidence interval [0.16, 1.08], odds ratio = 1.86. This is a small effect size. Specifically, there was a recognition advantage of testing over restudying in the image condition, *Wald* (27) = 9.82, p = .002, regression coefficient = 0.50, 95% confidence interval [0.19, 0.82], odds ratio = 1.65 (small effect size), but not in the word condition, *Wald* (27) = 0.45, p = .500, regression coefficient = -0.11, 95% confidence interval [-0.44, 0.21], odds ratio = 0.90 (small effect size). In addition, we found a main effect of study method, *Wald* (55) = 9.92, p = 0.002, regression coefficient -0.51, 95% confidence interval [-0.82, -0.19], odds ratio = 0.60. This is a small effect size. The mean proportion of recognized words was higher after testing (M = .83, SD = .12) than after restudying (M = .80, SD = .15). The main effect of surface features overlap did not reach statistical significance, *Wald* (55) = 0.21, p = .650, regression coefficient = -0.11, 95% confidence interval [-0.59, 0.37], odds ratio = 0.89. This is a small effect size. Additionally, the mean proportion of correctly classified distractors was higher in the word condition (M = .85, SD = .10) than in the image condition (M = .80, SD = .13), *Wald* (56) = 8.29, p = .004, regression coefficient = 0.33, 95% confidence interval [0.11, 0.55], odds ratio 1.39. This is a small effect size.

Table 4. Mean Proportion of Correctly Recognized Targets in Experiment 4a by Surface Features Overlap and Study Method. Standard Errors are between Brackets

	Surface Features Overlap	
Study Method	**Words**	**Images**
Restudy	.84 (.02)	.77 (.03)
Testing	.82 (.03)	.84 (.02)

Results Experiment 4b

The three outcome variables and their analyses are the same as in Experiment 1.

Immediate test performance

The mean proportion correctly retrieved tested items during the learning phase differed between the word condition (M = .68, SD = .12) and the image condition (M = .74, SD = .12), *Wald* (59) = 9.94, p = .002, regression coefficient -0.28, 95% confidence interval [-0.46, -0.11], odds ratio = 0.76. This is a small effect size. However, the Pearson correlation coefficient between immediate test performance and the final test difference scores (correctly classified tested words - correctly classified restudied words) was r = 0.02, p = .860, which indicates that the differences on the immediate test did not confound the final test results.

Final test performance

A 2 Study Method (restudy vs. testing) x 2 Surface Features Overlap (images vs. words) logistic regression on the binomial count targets (see Table 5) only showed a trend toward a statistical study method x surface features overlap interaction effect, *Wald* (56) = 3.18, p = .075, regression coefficient 0.42, 95% confidence interval [-0.04, 0.88], odds ratio = 1.52. This is a small effect size. The odds ratio of 1.52 means that in the word condition, the difference in odds of a correct answer between the tested and restudied words is 1.52 times this difference in the image condition. There was no main effect of study method, *Wald* (56) = 1.09, p = .300, regression coefficient = -0.16, 95% confidence interval [-0.45, 0.14], odds ratio = 0.85. This is a small effect size. The main effect of surface features overlap did not reach statistical significance, *Wald* (56) = 2.61, p = 0.11, regression coefficient = 0.38, 95% confidence interval [-0.09, 0.84], odds ratio = 1.45. This is a small effect size. Additionally, the mean proportion of correctly classified distractors was higher in the word condition (M = .83, SD = .10) than in the image condition (M = .79, SD = .16), *Wald* (59) = 6.38, p = .012, regression coefficient = 0.26, 95% confidence interval [0.06, 0.46], odds ratio = 1.30. This is a small effect size.

Table 5. Mean Proportion of Correctly Recognized Targets in Experiment 4b by Surface Features Overlap and Study Method. Standard Errors are between Brackets

	Surface Features Overlap	
Study Method	**Words**	**Images**
Restudy	.88 (.02)	.77 (.03)
Testing	.85 (.02)	.79 (.03)

Discussion

In Experiment 4a, we found an interaction effect of study method x surface features overlap on recognition performance. In the word condition, there was no statistical difference between tested items and restudied items, while in the image condition a testing effect did emerge. In Experiment 4b, although there was no statistical interaction effect, numerically there was a tendency in the expected direction. In addition, there was a main effect of study method in Experiment 4a but not in Experiment 4b. When zooming in on the results of Experiment 4a, it appears that this main effect of study method resulted from the relatively high testing score in the image condition, which at the same time gave rise to the interaction effect in Experiment 4a.

What can we conclude from these outcomes? In Experiment 4a the *p*-value of the critical interaction effect was smaller than .05, while in Experiment 4b it was larger than .05 (i.e., $p = .075$). However, identical replication studies are likely to produce different outcomes as a result of random sampling fluctuation, especially for sample sizes that are typically used in psychological research (e.g., Coursey, Hovis, & Schulze, 1987; Gámez, Diaz, & Marrero, 2011; Lakens & Etz, 2017; Morey & Lakens, 2016). It is therefore best to evaluate the results of replication studies on more than just the criterion of statistical significance. That is, when the replication attempt is imprecise, the conclusion based on statistical significance might be opposite to what the evidence warrants (Simonsohn, 2015). It is possible that a replication study obtains an effect size similar to that of the original study, but still produces a nonsignificant finding because the replication estimation is noisy or underpowered. On the other hand, two effect sizes can differ to a large extent but both lead to statistically significant outcomes. In the latter case, the replication attempt cannot be said to be successful. When evaluating the findings of replication studies, it is therefore important to also check whether the effect sizes and the confidence intervals are similar (Cumming, 2014). Now, the effect sizes of the interaction effects in Experiments 4a and 4b are comparable (i.e., odds ratios of 1.86 and 1.52, resp.), and the 95% confidence intervals of their regression coefficients largely overlap. This indicates that the interactions effects in these experiments were consistent with one another.

Furthermore, we conducted a 2 Experiment (4a vs 4b) x 2 Study Method (restudy vs. testing) x 2 Surface Features Overlap (images vs. words) logistic regression on the binomial count targets, thus combining the data of Experiments 4a and 4b. This analysis yielded a statistical interaction effect between study method and surface features overlap, *Wald* (114) = 9.27, $p = .002$, odds ratio = 1.88, which is a small effect. Specifically, there was a recognition advantage of testing over restudying in the image condition, *Wald* (58) = 8.56, $p = .003$, regression coefficient = 0.32, 95% confidence interval [0.10, 0.54], odds ratio = 1.37 (small effect size), but not in the word condition, *Wald* (57) = 2.18, $p = .140$, regression coefficient = -0.18, 95% confidence interval [-0.42, 0.06], odds ratio = 0.83 (small effect size). Moreover, we did not observe a statistical three-way interaction effect, *Wald* (114) = 1.05, $p = .300$, odds ratio = 0.79 (small effect), which means that there is no indication that the interaction effect between surface features overlap and study method statistically differed between Experiments 4a and 4b. Again, this suggests that the outcomes of Experiments 4a and 4b are comparable. Both experiments clearly reinforce each other and together they provide evidence that the short-term testing effect is larger for images than for words.

Summary of all findings

In Figure 1, the regression coefficients are plotted, together with their confidence intervals, corresponding to the log (ln) of the odds ratio between restudy and testing within the surface features overlap conditions of the five experiments. In this plot, a positive regression coefficient signifies an advantage of testing on the final recognition test (i.e., a testing effect), while a negative coefficient denotes an advantage of restudying. In general, the overlap between the twelve confidence intervals is large, suggesting that the differences between conditions are small (and/or that the parameter estimations are imprecise). At the top of the figure the regression coefficients of the five *word* conditions are presented, which were predicted to be close to zero. Figure 1 shows that they are indeed centred around zero, signifying the absence of a testing effect in these conditions. Below are the regression coefficients of the seven *non-word* conditions. Because there was still some surface features overlap in the final test conditions of Experiment 1 (scrambled, background, background scrambled), we expected a (very) small advantage of testing in the scrambled condition and the background condition, and possibly also in the background scrambled condition. We predicted larger advantages of testing to occur in the synonym and the image conditions, since these cues were purely semantic. It turned out that the subtle manipulations in Experiment 1 did not yield a testing effect in any of the surface features overlap conditions. In the synonym condition of Experiment 2, we did not find a testing benefit either. However,

Figure 1 clearly shows that the results from the three image conditions in Experiments 3, 4a and 4b stand out. Contrary to all other surface features overlap conditions, the image conditions consistently produced a mean recognition benefit of tested items over restudied items in the studied samples. Although the 95% confidence intervals indicate that there was only a two-tailed statistical testing effect in Experiment 4a, the results as a whole provide evidence that using image cues in the final test can give rise to a short-term testing effect in recognition.

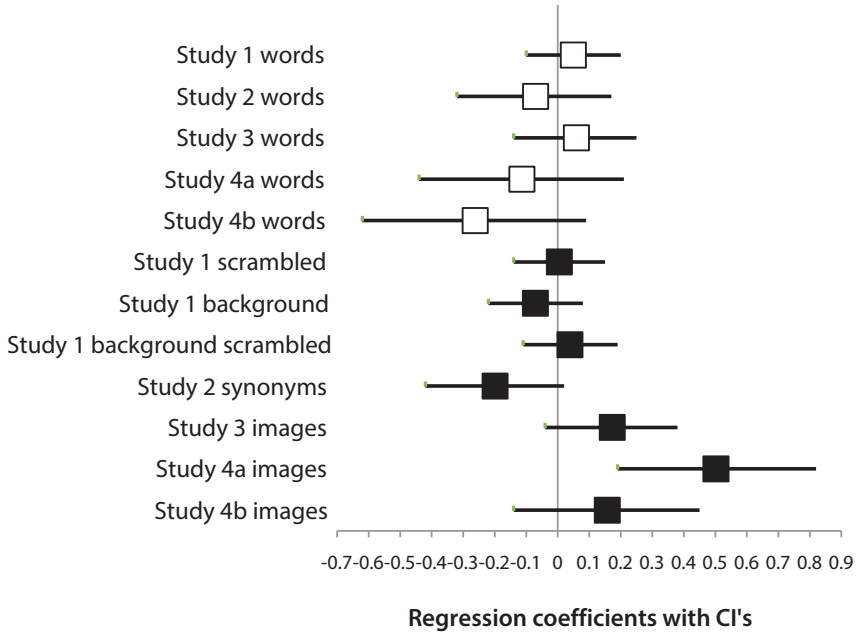


Figure 1. The regression coefficients, with their 95% confidence intervals, corresponding to the log of the odds ratio between restudy and testing within the twelve different surface features overlap conditions of the five experiments. The white squares correspond to the five word-cue conditions, and the black squares correspond to the conditions where the surface features of the final test cues were altered as compared to the learning phase.

General discussion

The fuzzy trace account of the testing effect predicts that a short-term testing effect will emerge when there is a low degree of surface features overlap between the items in the learning phase and the final test. In the present study, we assessed the fuzzy trace account by gradually reducing the availability of surface cues in the final test. In Experiment 1, the four surface features overlap conditions consisted of words, scrambled words, words vertically presented in a different font on a colorful flowered background

("background"), or a combination of the latter two conditions. In Experiment 2, the surface features overlap conditions contained either the same words or synonyms of the studied words. Experiment 3, 4a and 4b had two surface features overlap conditions: words and images. In Experiment 1 and 2, the reduction of the availability of surface cues in the final tests did not result in a benefit of testing over restudying, which is not congruent with the fuzzy trace theory. The findings in Experiments 3, 4a, and 4b, however, differed markedly from the findings in Experiments 1 and 2. These experiments showed an (numerical) advantage of testing as compared to restudying in the image conditions, which is in keeping with the fuzzy trace theory. Moreover, in the word conditions of Experiments 4a and 4b, no benefit of testing occurred. Experiments 4a and 4b were identical in their methods and subject pools, and overall produced corresponding results. However, although these testing benefits show that image cues can produce short-term testing effects in recognition memory, we hasten to add that more research is needed to examine the robustness of the short-term testing effect with image cues, because the results in the image conditions were small and quite variable. All in all, the present study provides only weak evidence in support of the fuzzy trace theory of the testing effect.

How can we explain the difference in results between the image studies on the one hand and Experiment 1 and 2 on the other? Apparently, the images trigger a distinctive response as for the effect of testing versus restudying on recognition. In Experiment 1, the manipulation of surface features overlap was more subtle than the manipulations in the other four experiments. Possibly, a considerable number of surface cues was still present in the final test of Experiment 1, such that the recognition of restudied words was not sufficiently impaired. The results in the synonym condition of Experiment 2, however, were surprising, because it was conceptually identical to the across-language condition in Verkoeijen, Bouwmeester, and Camp (2012). A potential explanation for these deviating outcomes might be that the synonyms in the final test did in fact activate the surface features of their intermediate test equivalents. Evidence for this idea comes from lexical decision studies that have demonstrated cross-synonyms priming (e.g., Perea & Rosa, 2002), but no cross-language repetition priming (e.g., Gerard, & Scarborough, 1989; Kirsner, Brown, Abrol, Chadna, & Sharma, 1980; Scarborough, Gerard, & Cortese 1984; Zeelenberg & Pecher, 2003). This difference might be due to lexical decision tasks depending primarily on orthographic processes (e.g., De Groot, 2002; Zeelenberg & Pecher, 2003). Now if it is true that the surface features of the synonyms were in fact activated, then this would explain the difference between the findings in the synonym and the image conditions. However, this idea is speculative, and future research could focus on the differences between the memory representations of translation equivalents and synonyms.

In our last two experiments, as well as in the Verkoeijen, Bouwmeester, and Camp study (2012), the tasks were performed by Dutch college undergraduates at our laboratory. In the first three studies, on the other hand, AMT workers performed the task anonymously. This complicates the comparison between experiments 1/2/3 versus 4a/4b. It might be possible that the AMT population has some distinctive characteristics that the Dutch undergraduates population lacks, which in turn interacted with the study method x surface features overlap effect in the present experiments. However, although there are known differences between the ATM population and a typical undergraduate pool (e.g., Paolacci, Chandler, & Ipeirotis, 2010), there are no theoretical reasons as to why these differences should produce a three-way interaction in present study. In addition, when looking at task performance measures, the ATM participants were very similar to the psychology undergraduates. That is, on the immediate test scores and the scores on the distractors we obtained highly comparable results across our ATM experiments and our experiments with psychology undergraduates. Also, the final test scores and the standard deviations were fairly similar across experiments. Moreover, many replication studies have shown that the behavior of the AMT population resembles the behavior of laboratory participants (e.g., Buhrmester, Kwang, & Gosling; Casler, Bickel, & Hackett, 2013; Horton, Rand, & Zeckhauser, 2011; Paolacci, Chandler, & Ipeirotis, 2010; Rand, 2012). Furthermore, Klein et al. (2014) assessed the replicability of a number of studies and found that very little of the variability in effect sizes could be attributed to whether the data collection occurred online or in the laboratory. All things considered, it seems unlikely that there were *relevant* differences in this study between the lab population and the AMT population.

However, perhaps procedural differences between Experiment 1/2/3 and Experiments 4a/4b might underlie the deviating results. Specifically, it might be possible that typing the words during the initial learning phase (Experiment 4a and 4b), and/or typing the words in separate periods during the free recall phase (Experiments 4a and 4b) versus typing them all in one go (Experiments 1, 2, and 3), made a difference to the outcomes. For example, typing responses during the initial learning phase (Experiments 4a and 4b) could have strengthened the verbatim traces to a higher extent than not typing. However, if this were true, one would not have expected any testing effects in these last two experiments. Moreover, Verkoeijen, Bouwmeester, and Camp (2012) and Coppens, Verkoeijen and Rikers (2011) also asked participants to type their responses during the initial study phase, as well as to type the words in separate periods during free recall. In the latter study, an advantage of testing over restudy did emerge after seven days, again suggesting that in the testing condition, the gist traces had been strengthened more than the verbatim traces. Taken together, we think it is unlikely that the procedural differences between Experiment 1/2/3 and Experiments 4a/4b led to variability in outcome patterns.

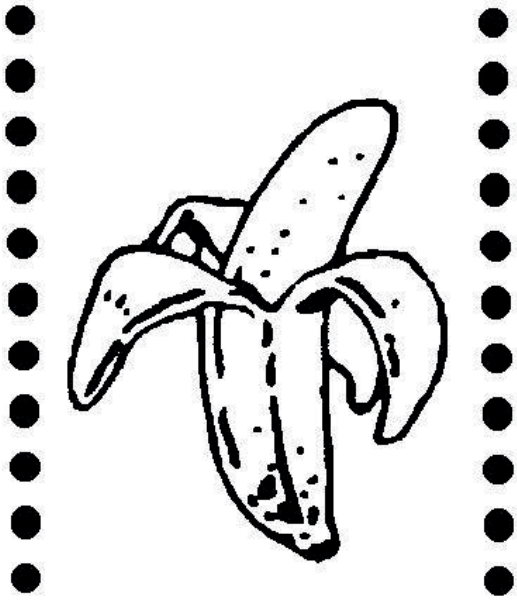
A different kind of explanation for the opposing outcomes evidently concerns the fuzzy trace theory itself. It is likely that the central notion that testing activates semantically related information does not fully correspond to reality. From a broader perspective, this would also mean that this type of elaborative retrieval accounts (e.g. Carpenter, 2009, 2011; Pyc, & Rawson, 2010) is not corroborated, which is in line with a number of other studies (e.g., Karpicke, Lehman, & Aue, 2014; Lehman, & Karpicke, 2016).

A different theory that might explain our findings is the bifurcation model (Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011). This theory predicts that the more difficult the final test, the larger the benefits of testing. According to this framework, items that are successfully recalled during testing are strengthened more in memory than items that are restudied. This implies that when the final test is sufficiently difficult, tested items will more often meet the criterion for retrieval in the final test than restudied items. Applied to the present study, the reduction in surface features overlap did not require participants to rely more on gist than on verbatim, but simply made the final test more difficult. However, the findings in Experiment 2 do not square well with the bifurcation framework. In the condition with the lowest average performance (the synonym condition), no benefit of testing emerged. Also, according to the bifurcation framework, one would have expected the benefit of testing to increase with a decreasing level of performance in the different final test conditions of Experiment 1. However, the data do not show such a pattern. In the two most difficult final test conditions, the difference between testing and restudying is absent (the scrambled condition) or statistically nonsignificant (the background scrambled condition). Furthermore, performance in the image condition was lower in Experiment 4b than in Experiment 4a, while the advantage of testing compared to restudy was somewhat larger in Experiment 4a than in Experiment 4b. Taken together, the bifurcation model cannot account for the findings in the present study.

All things considered, the outcomes of this study do not provide strong support for the fuzzy trace theory of the testing effect. The theory predicts that a short-term testing effect will arise when the overlap in surface cues between the learning phase and the final test is limited. This idea was substantiated in Experiment 4a and 4b, and partly in Experiment 3. Because the effect size estimates in these experiments were small and somewhat variable, it would be interesting to conduct a large-scale replication study to obtain more precise estimates, and shed more light on the question whether the fuzzy trace theory reveals one of the mechanisms underlying the testing effect.

Appendix A

Appendix B



3

How to comprehend a text: Retrieval practice versus self-explanation

This chapter is in preparation as:
Van Eersel, G. G., Verkoeijen, P. P. J. L., & Rikers, R. M. J. P. (in preparation).
How to comprehend a text: Retrieval practice versus self-explanation.

Abstract

Generative learning occurs when a new mental representation of the information is built by mentally reorganizing and integrating it with prior knowledge. In this study, two generative learning strategies and a control condition were compared on an immediate comprehension test. First participants read an argumentative text and then completed the read-recite-review (RRR) condition, the self-explanation condition, or the baseline control condition. Participants in the RRR-condition first read a paragraph, then recited as much as possible, and afterwards read the paragraph again. Participants in the self-explanation condition clarified and explained the central ideas of each of the paragraphs. In the baseline control condition, participants only read the text for the first time and then immediately performed the final open-book multiple-choice comprehension test. No differences between conditions were found on the final test, which suggests that self-explanation and RRR are not beneficial for comprehension, at least not when using this type of argumentative text and open-book final test. However, these findings should be interpreted with caution due to the low Cronbach's alphas of the final test.

Text comprehension is one of the most important skills that students acquire during their years in school. However, comprehension instructions are often minimal or ineffective (Snow, 2002). In the field of educational psychology, many efforts have been made to discover strategies that optimize text comprehension and learning in general. In a review, Dunlosky, Rawson, Marsh, Nathan, and Willingham (2013) examined the effectiveness of ten general learning techniques (see also Fiorella & Mayer, 2015) that are easy to apply and for which learners do not need supervision. For each technique, an overall utility judgment was provided based on the generalizability of the technique to different learning conditions, populations, materials, and criterion tasks. Two techniques rated as having a moderate or high utility can be applied when learning from text: retrieval practice and self-explanation. Because the effects of both techniques on text comprehension have been assessed mostly in comparison with rereading, it remains an open question what their effect will be when compared to each other. This will be investigated in the present study.

According to the interactive-constructive-active-passive (ICAP) framework (Chi, 2009; Fonseca & Chi, 2011) four types of learning activities can be distinguished. *Passive* learning occurs when a learner is not engaging in any overt activity related to the learning task, like listening to a lecture or reading a text. *Active* learning means to be engaged in some form of overt action that does not go beyond what is stated in the material, such as writing verbatim notes. A *constructive* learning activity is executed when a learner produces some additional output that contains information beyond what is provided in the studied material, like self-explanation. Finally, *interactive* learning occurs when a partner is present with whom there is a dialogue about the material. Reasoning from this framework, as students move from passive to interactive learning, the higher the learning gains, especially with meaningful learning outcomes like comprehension. This prediction is based on the strength and depth of the hypothesized cognitive processes underlying the four activities, and it is substantiated by empirical evidence (Fonseca & Chi, 2011). Furthermore, according to Fiorella and Mayer (2015), constructive and interactive learning are forms of *generative* learning, which involves building a new mental representation of the information by integrating it with prior knowledge.

As mentioned above, an example of an active and generative learning strategy is self-explanation (i.e., Nokes, Hausmann, VanLehn, & Gershman, 2011; Richey & Nokes-Malach, 2015; Roy & Chi, 2005). Generating self-explanations is a promising technique that fosters learning across different materials, age groups, and criterion tasks (Dunlosky et al., 2013). In their review, Dunlosky and colleagues (2013) focused on self-explanation prompts that do not mention any content of the studied material, but contain only general instructions and questions, such as 'what information does this sentence provide for you?' Most studies employed procedural or problem-solving tasks,

like mathematical problems (e.g., Atkinson, Renkl, & Merrill, 2003; Renkl, 1997; Rittle-Johnson, 2006; Schworm & Renkl, 2006) computer programming (e.g., Bielaczyc, Pirollo, & Brown, 1995), playing chess (De Bruin, Rikers, & Schmidt, 2007), worked examples in physics (e.g., Nokes-Malach, VanLehn, Belenky, Lichtenstein, & Cox, 2012) analytical reasoning tasks (e.g., Neuman & Schwarz, 1998), and clinical reasoning (Chamberland et al., 2011). The criterion tasks that were used varied from standard memory tests to measures of comprehension (e.g., De Koning, Tabbers, Rikers, & Paas, 2011), like transfer tests asking students to solve problems that to some extent differed from the practice problems (e.g., Nokes-Malach et al., 2012).

Only a few self-explanation studies involved learning from text (e.g., Ainsworth & Burcham, 2007; Magliano, Trabasso, & Graesser, 1999; McNamara, 2004). For example, Chi, De Leeuw, Chiu, and LaVancher (1994) asked twenty-four eight-grade students to complete a test measuring knowledge of the human circulatory system. Afterwards, they read an expository text on this subject. Fourteen students were asked to read each sentence out loud and explain what it meant (prompted group). An experimenter provided them with this instruction after every sentence, and asked for clarification if the explanation was not clear. Ten other students were asked to read the text twice (unprompted group). On the final one-week delayed comprehension test, the gain in percentage correct answers from pre-test to post-test was statistically larger in the prompted group than in the unprompted group. Chi and colleagues (1994) concluded that self-explaining supported the integration of new information with already existing knowledge, resulting in a deep understanding of the text.

In most self-explanation studies with text, self-explanation is compared to reading out loud or rereading the material (an exception can be found in, e.g., O'Reilly, Symons, & Maclatchy-Gaudet, 1998, who compared self-explanation to elaborative interrogation when studying isolated biological facts, and found that the former outperformed the latter on cued recall and recognition). Especially rereading is a method regularly used by students (Karpicke, Butler, & Roediger, 2009), but both reading out loud and rereading the material can be considered passive learning activities, while self-explanation is a constructive and generative one (Fonseca & Chi, 2011). Given the ICAP framework, it is not surprising that generating self-explanations yields superior learning outcomes when compared to rereading or reading aloud. Hence, it is still an open question whether self-explanation will improve text comprehension when it is compared to another constructive and generative learning activity. An activity that fits this description is retrieval practice.

The (retrieval practice) testing effect entails that retrieval practice after an initial study phase enhances long-term retention more than restudying the material (for reviews, see Delaney, Verkoeijen, & Spirgel, 2010; Karpicke, Lehman, & Aue, 2014; Roediger & Butler,

2011; Roediger & Karpicke, 2006a). The testing effect has been demonstrated under a variety of practice tests, materials, and age groups, and was rated as having a high practical utility (Dunlosky et al., 2013). A number of studies has examined the effect of practice testing in educationally relevant settings (e.g., McDaniel, Anderson, Derbish, & Morrisette, 2007), or with more complex classroom materials, like expository texts (e.g., Kang, McDermott, & Roediger, 2007; Roediger & Karpicke, 2006b). In studies where texts were used, the criterion tasks mostly concerned retrieval of the verbatim text (e.g., Chan, McDermott, & Roediger, 2006), but sometimes inferential knowledge was assessed as well (e.g., Blunt & Karpicke, 2014; Butler, 2010; Foos & Fisher, 1988; Johnson & Mayer, 2009; Karpicke & Blunt 2011; McDaniel, Howard, Einstein, 2009; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; Van Eersel, Verkoeijen, Povilenaite, & Rikers, 2016).

In most studies with text material, retrieval practice has been compared to restudying; only a few studies have explored the effect of retrieval practice in comparison to a generative method.[1] For example, Karpicke and Blunt (2011) and Blunt and Karpicke (2014) compared retrieval practice (i.e., free recall) to elaborative studying with concept mapping. Blunt and Karpicke (2014, Experiment 2) asked participants to read a text and then perform free recall or to create a concept map, either with or without the text present. In the free recall condition, participants wrote down all that they remembered from the text. In the concept mapping condition, participants were explained what a concept map is (i.e., a diagram in which concepts are represented as nodes that are linked together with words and phrases), then saw an example of a concept map, and eventually were asked to recall the text by creating a concept map on paper. Both learning activities were performed with or without the texts provided. Afterwards participants read the text again and then repeated the learning activity. On the final short-answer inference test administered after one week, there was no difference between free recall and concept mapping. However, performance on the final test was better with the text absent than with the text present, regardless of the learning condition. Blunt and Karpicke (2014) concluded that recalling the material was the locus of this retrieval practice effect, with the exact format not making much difference.

The present study

This study directly compares self-explanation and retrieval practice on comprehension in an educational setting. As the retrieval practice format, we choose the read-recite-review strategy (McDaniel, Howard, & Einstein, 2009), which is a simplified version of Robinson's (1941) SQ3R (survey-question-read-recite-review) method. This strategy

1 Note that Larsen, Butler and Roediger (2013) did compare retrieval practice and self-explanation, but in their study the topics were taught in an "interactive didactic format" instead of in a pure text format.

requires students to first read a text, then recite as much as possible from the material, and afterwards read the text again. In this way, students receive feedback on their performance, which greatly enhances the effect of retrieval practice (e.g., Roediger & Butler, 2011). Also, the literature (e.g., Butler & Roediger, 2007; Dunlosky et al., 2013; Glover, 1989; McDaniel et al, 2007; Rowland, 2014) suggests that tests that involve more generative answers (i.e., recall or short answer) are more effective than tests calling for less generative answers (i.e., fill-in-the-blank or recognition; Dunlosky et al., 2013). This is another reason to opt for the RRR-strategy, which falls into the first category. An additional advantage of the RRR-strategy is that it is easy for students to implement when they are studying a text on their own.

Furthermore, we address some critical issues in the present study. Firstly, according to Dunlosky and colleagues (2013), a general shortcoming of the self-explanation literature is that only a few studies controlled for time on task, with self-explanation typically taking more time than the control activity. For example, in the study by Chi and colleagues (1994), the mean study time of the prompted group was 2 hr 5 min, while the mean study time of the unprompted group was 1 hr 6 min. As a result, it is impossible to determine whether the superior performance of the self-explanation group is to be attributed either to the activity of self-explaining, or to the increased time on task (Ploetzner, Dillenbourg, Preier, & Traum, 1999). We therefore control for time on task in this study. Secondly, in a strict sense, the activity employed by the students in the study by Chi and colleagues (1994) does not qualify as generating self-explanations, as the term self-explanation implies that the explanation is directed towards oneself instead of towards a partner. For that reason, in the present study the self-explanations are not aimed at an experimenter.

Pilot study

We first performed a pilot study in order to examine the materials and determine the optimal time-on-task in the two experimental conditions.

Participants

Participants were thirteen Dutch undergraduate students from Erasmus University Rotterdam. They received course credits for their participation. Their mean age was 20 years (*SD* = 2.52). Five of them were males, eight were females.

Design

A 3 Study Strategy (self-explanation vs. read-recite-review) between-subjects design was used. Participants were randomly assigned to one of the two conditions.

Materials

We used an argumentative text of 1223 words (13 paragraphs) on welfare and the knowledge economy that had been part of the Dutch national exams in 2009. Argumentative text is the official text type used in the Dutch secondary school exams to assess reading comprehension. All national exams are developed and audited by the official exam institute CITO (*Centraal Instituut voor Toetsontwikkeling*). The text was accompanied by twelve open-book multiple-choice comprehension questions that were also developed by CITO. Cronbach's alpha of this original set of questions when used in the Dutch national exams as reported by CITO was 0.41, which was comparable to the reported Cronbach's alphas of questions used in other years. Because Cronbach's alpha was very low, we used this pilot to construct a new set of questions that might result in a higher Cronbach's alpha.

Procedure

Participants were tested and interviewed in groups of 2–3 people. First, they were informed by the experimenter that they would participate in an experiment on reading comprehension, where they had to study a text and afterward answer multiple-choice questions. Then they received a booklet including the text and instructions, and they read the text for the first time. Next, they performed one of the two study strategies on paper, see below. Afterwards participants answered the twelve accompanying open-book questions. The procedure was self-paced, and the first author recorded the time that each participant spent on each part of the procedure. When all participants in one session had finished answering the questions, they were asked to explain whether the questions were clear and unambiguous. The text and the questions were then discussed with the experimenter for 15–30 min.

In the self-explanation condition, participants read a general instruction on how to self-explain, which was partly based on the instruction used by Ainsworth and Burcham (2007). The general instruction was as follows: "Next you will perform a study strategy called 'self-explanation'. You will clarify and explain the central concept(s) in the paragraph in your own words. Also try to relate the information in the paragraph to what you have read in previous paragraphs. If you have questions about the content, write them down as well. Note that a self-explanation is more than just a rewording or paraphrasing of what was stated in a sentence; it involves making connections between what you have read and other sections of the text or prior knowledge related to the topic. Don't worry about sounding eloquent; it is the self-explaining process, rather than the outcome, that will aid your learning." After reading this instruction, participants turned over the page to start self-explaining the individual paragraphs on paper. Participants could only view the paragraphs by consulting the full text. Every time before self-

explaining an individual paragraph, participants read the following short instruction of self-explanation on the top of the page: "What are the central ideas in this paragraph and how are they related to what you have previously read? Relate the information in this paragraph to the information in the previous ones."

Note that there were two reasons why we opted for self-explanation per paragraph instead of per sentence (cf. Ainsworth & Burcham, 2007; Chi et al., 1994). Firstly, in an argumentative text the connections between the central pieces of information in the text are more important than the separate sentences. Secondly, if we had chosen for self-explanation after each sentence, the self-explanation task would have taken much more time than the tasks in the other two conditions. By asking participants to self-explain after each paragraph, it was possible to keep the time-on-task equal between conditions (at least nominally, if not functionally).

In the read-recite-review (RRR) condition, participants first read a paragraph without seeing the rest of the text. Afterwards they turned over the page and were asked to recite as much as possible from the paragraph they had just read, without the possibility to consult the text. Finally, they were asked to turn over the page and read the paragraph again without seeing the rest of the text.

Time-on-task results

The average time that participants needed to read the text for the first time was 6:39 min (*SD* = 0:56), with a range from to 6:00 to 8:30 min. Furthermore, in the RRR-condition, participants used on average 2:03 min (*SD* = 0:44) to read a paragraph, 2:25 min (*SD* = 0:37) to recite, and 0:38 min (*SD* = 0:11) to review the paragraph. In the self-explanation condition, the average time participants took to self-explain a paragraph was 3:40 min (*SD* = 1:32), with a range from 1:18 to 8:00 min. Finally, the number of minutes that participants spent on answering the final questions ranged from 8:07 to 23:00 minutes, with an average of 13:45 min (*SD* = 5:32).

For the main experiment, we decided to fix the time for reading the text for the first time at 8:00 min. In the RRR-condition, we set the average time to read a paragraph at 2:00 min, the recite time at 3:00 min and the review time at 0:30. We then set the self-explanation time at 5:30 per paragraph. Finally, the time for answering the final test questions in the main experiment (i.e., in the self-explanation condition and in the RRR-condition) was fixed at 25 minutes.

Material evaluation

Only if all thirteen participants agreed that a question was clear and unambiguous, it was used in the final test of the main experiment. Following this procedure, we selected seven out of the twelve original exam questions, of which we clarified four. One of these

four questions was an open question, which we turned into a multiple-choice question by constructing four answer options. Two other questions had one answer option that was not well understood by several participants (both were incorrect answers), so we replaced these two answer options by new ones. Lastly, the right answer option of one other question contained a few old-fashioned and abstract words, so we replaced these by more concrete and contemporary words. Furthermore, one paragraph of 79 words was not necessary to keep the structure of the text, and neither was it useful for answering any of the final test questions. We therefore decided to remove this paragraph from the text in the main experiment.

Method main experiment

Design and participants

A 3 Study Strategy (self-explanation vs. read-recite-review vs. baseline control) between-subjects design was used. Participants were randomly assigned to one of the three study strategy conditions. We calculated the effect size for the comparison between the prompted and the unprompted group on the inference and transfer questions (Category 3 and 4) in the study by Chi and colleagues (1994, p. 453), which was r = .49. A power analysis in G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that a number of 54 participants was sufficient to detect an effect size of r = .49 for a comparison between three groups with a power of .95. Note that in the studies by Chi and colleagues (1994) two groups were contrasted (i.e., self-explanation vs. read-only), while in the present experiment three groups were compared. Because the time on task might have confounded the results in the study by Chi and colleagues, we lowered the estimated effect size by .10 to r = .39. G*Power showed that in order to find an effect of r = .39 for the difference between three groups, 87 participants was sufficient.

Participants were 99 Dutch undergraduates from the Erasmus University Rotterdam, the Netherlands, who were rewarded with course credits or a financial compensation. Two out of the 99 participants indicated that they had read the text before, so they were excluded from participation, leaving a total number of 97 participants. Then, before grading the final test, the first author read all the self-explanation protocols and excluded seven people from participation because they did not comply with the self-explanation instructions. That is, these participants had not written down anything for any of the paragraphs, or they had only copied a few parts of the text *literally*. The RRR-protocols were also scored, and here every participant had complied with the instructions. This resulted in a final number of 90 participants. Their mean age was 21.92 years (SD = 3.29). Fifty-seven of them were female, 33 were male.

Materials

All materials and data from this study can be retrieved from the Open Science Framework[2]. We used the same argumentative text of 1144 words (12 paragraphs) that we used in the Pilot study (so with the one paragraph removed), which had been part of the Dutch national exams. We also used seven accompanying open-book multiple-choice questions, of which we clarified four by replacing some words (see Pilot Study). We also constructed five new questions ourselves. Together these twelve questions measured the level of understanding of the central pieces of information in the text and the explanatory connections between them. All questions required participants to make inferences going beyond what was stated in the text. In two of the questions (both from the original set of exam questions) a second text was introduced, and participants were asked how the original text and the additional text were related. An example of a question was the following:

"The title of the text is '*The illusions of the knowledge economy*'. Several illusions are being discussed in the text. On which of these illusions is placed the strongest emphasis?

A. The illusion that life if fully predictable.

B. The illusion that all life-threatening dangers can be banished.

C. The illusion that technological advancements improve the quality of life.

D. The illusion that life can be fully controlled."

Note that the final test questions were open-book, for two reasons. Firstly, Chi and colleagues (1994) also allowed participants to consult the text, in order to place more weight on learning as opposed to memory (Chi et al., 1994, p. 451). Secondly, the questions that we used from the Dutch national exams had been constructed in such way that in order to answer them, it was necessary to look back at the text.

Procedure

Students were tested in groups of 5–15 participants. They were informed by the experimenter that they would participate in an experiment on reading comprehension and that they would read a text and answer multiple-choice questions afterward. Next, students received a booklet including the text and instructions. First, they read the text during 8 minutes, and then they performed one of the three study strategies. Finally, they received the open-book test containing the twelve multiple-choice comprehension questions. During the entire experiment, the experimenter kept track of time and notified participants when time was up.

For the self-explanation instructions, see Pilot Study. Participants self-explained the individual paragraphs on paper for 5:30 minutes per paragraph. Total time-on-task in this condition was 99 minutes: 8 minutes to read the text for the first time and 66

2 osf.io/9gm42

minutes to self-explain the paragraphs, and another 25 minutes to answer the final questions. The average number of minutes that participants in this condition spent on the final test ranged from 6:18 to 19:54, with a mean of 12:10 and a median of 12:06.

For the exact instructions in the RRR-condition, see Pilot Study. Participants first read a paragraph during 2:00 minutes on average (the exact reading time depended on the length of the paragraph). They were then asked to recite as much as possible from the paragraph for 3:00 minutes. Finally, they read the paragraph again for 0:30 minutes. Total time-on-task in this condition was 99 minutes: 8 minutes to read the text for the first time, 66 minutes to read, recite, and review the individual paragraphs, and 25 minutes to answer the questions. The average number of minutes that participants in this condition spent on the final test ranged from 6:24 to 9:18, with a mean of 12:38 and a median of 11:54.

The baseline control group served to check whether our two study strategies were useful as compared to no study at all, and was tested after the first two experimental conditions had been carried out. The participants in the baseline control condition were asked to read the text for 8 minutes, identical to the first phase in the other two conditions. Afterward, they immediately received the final test questions. The average amount of time that participants needed to answer the questions in the other two conditions was 12 minutes. In order to keep the functional time-on-task identical, participants in the control condition were also given 12 minutes to answer the questions. In total, time-on-task in this condition was 20 minutes. Note that it was possible for participants to skip final test questions, but in fact only one participant once skipped a question of the first ten questions (for questions 11 and 12, see below).

Results main experiment

The first outcome variable is the number of correct responses to the twelve final comprehension questions. Due to the expected lack of time in the control group, these twelve questions were answered by only 78 of the 90 participants. Eleven participants in the control condition, and 1 person in the self-explanation condition, did not answer the final two questions, which were then scored as incorrect (i.e., 0). The other dependent variable was the number of correct responses to the first ten final comprehension questions, which were answered by all participants.

Scale

The average number of correct responses to all twelve comprehension questions was 7.33 (*SD* = 2.23) and the average number correct to the first ten comprehension question was 6.50 (*SD* = 1.90). The Cronbach's alpha of the set of twelve final comprehension

questions was 0.53. In the self-explanation group Cronbach's alpha was 0.12, in the RRR-group it was 0.63, and in the baseline control group it was 0.65. Note that 12 of the 90 participants did not answer the last two questions. In line with the scoring procedure of CITO, we gave these missing answers a score of 0. The Cronbach's alpha of the first ten final questions, which were answered by all participants, was 0.47. In the self-explanation group Cronbach's alpha was .28, in the RRR-group it was .51, and in the baseline control group .60. In Table 1, the inter-item correlations are presented. In Table 2, the mean proportion correct, standard deviation, and alpha if item deleted for each question from the twelve-question scale are shown.

Table 1. The Inter-Item Correlations for the Questions of the Twelve-Question Scale

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
Q2	.07										
Q3	.08	.20									
Q4	.07	.03	.20								
Q5	.32	.19	-.18	.10							
Q6	.05	.21	.12	.04	-.01						
Q7	-.06	-.01	.00	.01	.30	.02					
Q8	.07	-.09	.05	.08	.04	-.04	-.04				
Q9	.20	-.04	.03	.18	.11	.09	-.00	-.07			
Q10	.16	.11	-.01	.19	.29	.17	.12	.06	.28		
Q11	.21	-.09	.03	.15	-.06	.01	-.08	.09	.41	.16	
Q12	.18	.16	.23	.19	.03	.20	-.07	-.09	.23	.00	.09

Table 2. The Mean Proportion Correct, Standard Deviation, and Alpha if Item Deleted for the Questions of the Twelve-Question Scale

	Mean	SD	Alpha If Item Deleted
Q1	.50	.50	.42
Q2	.70	.46	.45
Q3	.79	.41	.47
Q4	.87	.34	.44
Q5	.62	.49	.40
Q6	.38	.49	.46
Q7	.84	.36	.48
Q8	.52	.50	.51
Q9	.71	.46	.44
Q10	.57	.50	.38
Q11	.34	.48	.42
Q12	.49	.50	.45

Final test

A 3 Study Method (self-explanation vs. RRR vs. reread) ANOVA on the set of first ten questions did not reveal a statistical effect of study method, $F(2,87) = 0.84$, $p = .44$, $\eta_p^2 = .02$ (see Table 3). Because a non-significant *p*-value does not provide evidence for a null effect (e.g., Rouder, Speckman, Sun, Morey, & Iverson, 2009; Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2017), we also performed a 3 Study Strategy (self-explanation vs. RRR vs. baseline control) *Bayesian* ANOVA (e.g., Rouder, Morey, Speckman, & Province, 2012) on the set of first ten questions in the software program JASP (Love et al., 2015; Wagenmakers et al., 2016) with a default Cauchy prior width of $r = 0.50$ for effect size on the alternative hypothesis (for arguments for the Bayesian approach, see, e.g., Dienes, 2011). This analysis yielded a Bayes Factor of BF01 = 5.18, indicating that the likelihood of the data under the null hypothesis was 5.18 times larger than the likelihood of the data under the alternative hypothesis that postulates the presence of an effect. Following Wetzels and Wagenmakers (2012), Bayes Factors between 3 and 10 indicate substantial evidence in favor of the relevant hypothesis, in this case the null hypothesis of no difference between the three study strategies on the number of correct answers to the first ten final comprehension questions.

We also performed a 3 Study Method (self-explanation vs. RRR vs. reread) ANOVA on the set of all twelve questions, which again did not reveal a statistical effect of study method, $F(2,87) = 1.66$, $p = .197$, *Partial* $\eta_p^2 = .04$ (see Table 3). A 3 Study Strategy (self-explanation vs. RRR vs. baseline control) Bayesian ANOVA with a default Cauchy prior width of $r = 0.50$ on the set of all twelve questions yielded a Bayes Factor of BF = 2.72, indicating that the observed data were 2.72 more likely under the null hypothesis than under the alternative hypothesis. This is anecdotal evidence in favor of the null hypothesis of no difference between the three study strategies on the twelve final comprehension questions (Wetzels & Wagenmakers, 2012).

Table 3. Number of Correct Answers by Study Strategy for each Dependent Variable. Standard Deviations are Between Brackets

Study Strategy	Ten questions	Twelve Questions
Baseline control	6.13 (2.10)	6.73 (2.43)
Self-Explanation	6.70 (1.62)	7.63 (1.67)
Read-Recite-Review	6.67 (1.95)	7.63 (2.44)

Discussion

According to Fiorella and Mayer (2015), generative learning occurs when a new mental representation of the information is built by mentally reorganizing and integrating it with prior knowledge. In the present study, we examined two generative learning strategies that are useful for learning from text: retrieval practice (RRR) and self-explanation. Because the effects of these strategies have been assessed mostly in comparison with rereading, it was an open question what the results on a comprehension measure would be when weighed against each other. In the present study, participants first read an argumentative text and then completed the self-explanation condition, the RRR-condition or the baseline control condition. Participants in the self-explanation condition had to clarify and explain the central ideas of each of the paragraphs, and relate this information to that in previous paragraphs. Participants in the RRR- condition first read a paragraph, then recited as much as possible, and afterwards read the paragraph again. In the baseline control condition, participants only read the text for the first time during eight minutes as in the other conditions, and then immediately performed the final comprehension test. This comprehension test comprised of twelve open-book multiple-choice questions. The results showed that the three learning strategies did not differ on the final comprehension test.

Given the large number of studies showing the advantages of retrieval practice (for reviews, see, e.g., Karpicke, Lehman, & Aue, 2014;) and self-explanation (e.g., Ainsworth & Burcham, 2007; Chi et al., 1994; Dunlosky et al., 2013), what could underlie these unexpected findings? Firstly, the lack of differences between conditions could be driven by the low Cronbach's alpha of the final test, especially in the self-explanation condition. If Cronbach's alpha of the final test is as low as in the self-explanation group, the variance between scores represents mostly measurement error. Because this measurement error is assumed to be random, it will not lead to reliable differences on the test in question. This might explain why we did not observe an effect of self-explanation on our final comprehension measure.

Although we constructed a new set of questions that resulted in a somewhat higher Cronbach's alpha than that of the original set of questions (which was 0.41), our alpha was still too low. Because Cronbach's alpha increases with the number of items, our low alpha might partly be due to the relatively small number of items in our final test, which was chosen for practical reasons. However, the problem of the low Cronbach's alpha is a general problem for the Dutch national exams that aim to measure text comprehension. Indeed, Kamalski, Sanders, Lentz, and Van den Bergh (2005) presented four hundred secondary school students with four argumentative texts and four methods measuring text comprehension: a cloze test, a sorting task, a mental model task, and a set of multiple-choice (and open) questions designed by CITO, the latter being comparable

to our questions. They found that these CITO questions indeed had a low Cronbach's alpha, and did not correlate with the three other comprehension measures (although no specific correlation coefficient was reported). Moreover, the test scores of the CITO method depended almost exclusively on which of the four texts was used. Meuffels and Van den Bergh (2006) therefore advice to use several texts when assessing text comprehension. Hence, future research into text comprehension might use multiple texts, and take the Cronbach's alphas into consideration.

Another possible explanation for our findings is related to the large time-on-task differences between conditions. Time-on-task in the RRR-condition and the self-explanation condition was 99 minutes, while it was only 20 minutes in the baseline control condition. It is possible that fatigue and boredom effects played a role here (Furr & Bacherach, 2014), reducing motivation in the two experimental groups. Specifically, because the final comprehension test required participants to look back at parts of the text, it might be possible that boredom and fatigue limited the advantages of the experimental conditions in comparison with the baseline control condition. That is, participants in the baseline control condition had only read the text for eight minutes, so in order to answer the questions they had no other option than look up the relevant parts of the text. Participants in the experimental conditions, on the other hand, had already read the text repeatedly and extensively. This might have resulted in a reduced motivation and inclination to look back at the text, which was nevertheless necessary to answer the final questions correctly. However, this idea is speculative, and future research could examine whether possible fatigue and boredom leads to reduced performance on this type of final test.

On the other hand, it is surprising that neither of the experimental conditions produced benefits on the final test, given that participants in both groups spent much more time studying the text than participants in the baseline control group. Perhaps self-explanation and retrieval practice are not very efficient strategies to improve text comprehension. This would be in line with other studies showing that self-explanation is "not worth the time" (McEldoon, Durkin, & Rittle-Johnson, 2012). The locus of the self-explanation technique is that a new mental representation of the text is built by integrating the different pieces of information within the text and by integrating it with prior knowledge (Fiorella & Mayer, 2015). Apparently, this activity did not lead to a final test benefit given our specific text materials and open-book final test. This result might be due to the nature of our text material. That is, our text was an argumentative one, in which the author makes a personal argument about the conditions under which the development of knowledge leads to an increase in general welfare. Many other self-explanation studies, however, used an expository text in which a certain systematic process was explained, for example the human circulatory system (e.g., Ainsworth

& Loizou, 2003; Chi at al., 1994). This involves a causal explanation of the interacting components within a system, which seems to be perfectly suited for self-explanation. On the other hand, an argumentative text like ours does not consist of an explanation of such a systematic process with a number of fixed steps, but instead of a more abstract and personal argumentation, which could make self-explanation less obvious.

Moreover, self-explanation might work better for graphics and diagrams than for text in general. Evidence for this claim was provided by Ainsworth and Loizou (2003), who presented twenty participants with information about the human circulatory system and an instruction to self-explain. Ten participants received this information in text and ten participants in diagrams. No differences between these groups were observed on the pre-test, but on the post-test only the diagram condition had improved significantly. Furthermore, the diagram groups generated more self-explanations than the text group. According to Roy and Chi (2005), such results suggests that the low expressivity of graphics and diagrams –as compared to texts– demand more self-explanations in order to fill in the missing information and construct a coherent representation of the information. To further test this hypothesis, Roy and Chi (2005) performed a small literature review of studies reporting the amount of self-explanations with different information formats. As expected, they found that the amount of self-explanation was lowest in text-only contexts, higher in multimedia contexts, and highest with diagrams-only. Furthermore, learning gains were equal for text and multimedia learning, but for diagrams-only it was substantially higher. In addition, overall performance was lowest for the text-only studies and highest for the diagrams-only contexts. Chi and Roy (2005) conclude that because there is more information to generate in learning from diagrams than in learning from text, diagrams result in more self-explanations and deeper learning.

Similarly, it is possible that there was no benefit of retrieval practice because retrieval practice research into higher level learning (i.e., comprehension, inference making, and transfer) has shown mixed results (e.g., Blunt & Karpicke, 2014; Butler, 2010; Eglington & Kang, in press; Foos & Fisher, 1988; Hinze, Wiley, & Pellegrino, 2013; Johnson & Mayer, 2009; McDaniel, Howard, Einstein, 2009; McDaniel et al., 2013; Tran, Rohrer & Pashler, 2014; Van Eersel, et al., 2016; Van Eersel, Verkoeijen, Tabbers, Van Mierlo, Paas, & Rikers, submitted). By contrast, testing effects seem to be reliably observed when pure memory final tests are employed (for a review, see Rowland, 2014). It might therefore be possible that retrieval practice works best when pure retention is assessed, but that the retrieval practice effect is weaker with final comprehension tests.

A different factor that may explain why we did not observe an advantage of retrieval practice is that we did not include a long-term retention interval. Although there is agreement in the literature that the retrieval practice effect arises after a long retention

interval (i.e., longer than one day), mixed support has been found for short intervals (Rowland, 2014). However, in the present study participants reviewed the text after the recite phase, which can be considered as a form of feedback. Now in several studies where feedback was provided after retrieval practice, testing effects did emerge after a short retention interval (e.g., Bishara & Jacoby, 2008; Carpenter, Pashler, Wixted, & Vul, 2008; Carrier & Pashler, 1992; Jacoby, Wahlheim, & Coane, 2010; Kang, 2010; Kornell, Hays, & Bjork, 2009; Wartenweiler, 2011).

In conclusion, these results seem to tell us that the known benefits of self-explanation and retrieval practice might not generalize to this type of argumentative text and open-book final test. However, the findings should be interpreted with caution due to the low Cronbach's alphas of the final test. It is therefore still an open question what the results on this kind of comprehension measure would be when self-explanation is compared to retrieval practice. Another pressing question following from this study is how to improve the reliability of the comprehension questions. Future research should shed some light on this important issue.

4

A comparison of study strategies for inference learning: Reread, verbatim free recall, and constructive recall

This chapter has been submitted as:
Van Eersel, G. G., Verkoeijen, P. P. J. L., Tabbers, H. K., Van Mierlo, S. A. A., Paas, F., & Rikers, R. M. J. P. (submitted). A comparison of study strategies for inference learning: Reread, verbatim recall, and generative recall.

Abstract

According to the Constructive Retrieval Hypothesis, retrieval practice that is focused on constructing a coherent text representation produces better inference learning than unguided free recall. The present study investigated whether this kind of constructive recall was indeed more beneficial than verbatim free recall for drawing inferences from a text, as measured by an inference test administered immediately and after a 1-week delay. Participants read expository texts and then engaged in either constructive recall, verbatim free recall, or rereading. In the constructive recall condition, participants were instructed to type in their own words what they had comprehended from the content of the text. In the verbatim free recall condition, participants were asked to type verbatim everything they could remember from the text. The final inference tests showed no differences between the three conditions. These results suggest that neither constructive recall nor verbatim free recall accommodates later inference learning.

A powerful way to boost long-term learning is by retrieving information from memory after an initial study phase; a phenomenon called the (retrieval practice) testing effect (for reviews, see Carpenter, 2012; Delaney, Verkoeijen, & Spirgel, 2010; Karpicke, 2012; Karpicke & Grimaldi, 2012; Karpicke, Lehman, & Aue, 2014; Karpicke & Roediger, 2006a; Roediger & Butler, 2011; Rowland, 2014). The testing effect has been demonstrated with a variety of practice tests and materials (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). However, most testing effect research involved final tests that only assessed retention (Roediger & Butler, 2011; Rowland, 2014), whereas less is known about the effect of retrieval practice on tests that measure related but new knowledge, i.e., *transfer* of knowledge (e.g., Barnett & Ceci, 2002; Butler, 2010; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013).

Transfer can be broadly defined as the ability to apply previously learned knowledge or skills in a novel context (e.g., Salomon & Perkins, 1989), and can be considered to be the main aim of learning (Carpenter, 2012; Rohrer, Taylor, & Sholar, 2010). One type of transfer is to make inferences about what has been learned from a text by connecting multiple concepts from the learned text to solve a new problem. Only a small number of studies has shown retrieval practice to produce better performance on a final inference test (e.g., Blunt & Karpicke, 2014; Butler, 2010; Johnson & Mayer, 2009; Karpicke & Blunt 2011; McDaniel, Howard, & Einstein, 2009; McDaniel et al., 2013; Van Eersel, Verkoeijen, Povilenaite, & Rikers, 2016). For example, in a study by Butler (2010), participants first read six prose texts, and then reread three of the passages (reread condition) and answered short-answer questions with feedback on the other three passages (retrieval practice condition). On a delayed 1-week final inference test, performance was better after retrieval practice than after rereading.

In his study, Butler (2010) used short-answer questions as the retrieval practice format. This is in line with the general finding that retrieval practice that involves elaborative recall (i.e., free recall or cued recall/short-answer questions) is more beneficial for long-term retention than retrieval practice calling for less elaborative recall (i.e., fill-in-the-blank or recognition tests; e.g., Butler & Roediger, 2007; Duchastel, 1981; Dunlosky et al., 2013; Glover, 1989; McDaniel, Anderson, Derbish, & Morrisette, 2007). Moreover, some researchers (e.g., Hinze, Wiley, & Pellegrino, 2013) have argued that a testing effect on *inference* questions might only emerge after a retrieval practice format that encourages the construction of a *situation model* of the text (Kintsch, 1994, 1998).

In his theory of text comprehension, Kintsch (1998) distinguished three levels that form the different mental representations of a text. The *surface level* contains the surface features of the text, representing the words and their syntactic relations (McNamara & Magliano, 2009). The *textbase level* can be seen as an abstraction of the exact words in propositional form, representing the meaning of the text as it is explicitly expressed by

the text (Kintsch & Rawson, 2005). The situation model, finally, is a mental model of the situation described by the text, in which information from the text is integrated with prior knowledge. This model contains both explicit and implicit relations between the ideas in the text. Inferences on this level are about relations that are implied in the text, but not explicitly mentioned (Wiley, Griffin, & Thiede, 2005).

Some studies have indeed found that elaborative forms of retrieval practice lead to better performance on an inference test that directly targets the situation model. For example, Blunt and Karpicke (2014, Experiment 2) asked participants to read a text and then perform one of two learning activities with feedback: free recall or creating a concept map. In the free recall condition, participants wrote down all that they remembered from the text. In the concept mapping condition, participants were explained what a concept map is (i.e., a diagram in which concepts are represented as nodes that are linked together with words and phrases), then saw an example of a concept map, and eventually were asked to recall the text by creating a concept map on paper. Both free recall and concept mapping were performed with or without the texts provided. Afterwards, participants read the text again and then repeated the learning activity. On the final short-answer inference test administered after one week, there was no difference between free recall and concept mapping. However, performance on the final test was better with the text absent than with the text present, regardless of condition. That is, both retrieval practice conditions showed a retrieval practice effect. Blunt and Karpicke (2014) concluded that recalling the material was the locus of this retrieval practice effect, and that the exact format did not matter much.

More evidence for the beneficial effect of elaborative retrieval practice for inference learning can be found in the study by McDaniel, Howard, and Einstein (2009, Experiment 2). They asked participants in their three study conditions to read two expository texts. Subsequently, participants in the read-recite-review condition (3R) recited as much as possible from a text, and then read the text again. Participants in the note-taking condition were instructed to read a text again and to take notes while reading. On both the final free recall test and the final short-answer inference test, administered immediately and after one week, the 3R and the note-taking groups performed better than the reread condition.

In sum, elaborative forms of retrieval practice can be helpful when it comes to generating inferences on the level of the situation model. However, in most studies that examined the effect of retrieval practice on inference performance (e.g., Blunt & Karpicke, 2014; Karpicke & Blunt, 2011; McDaniel et al., 2009), participants reread the text after performing retrieval practice, which can be considered as receiving feedback. Many studies have found that in general, providing feedback enhances the testing effect (see Rowland, 2014). However, providing feedback also confounds the direct

effect of retrieval practice with the effect of feedback (e.g., Van Eersel et al., 2016). The direct effect of retrieval practice amounts to the strengthening of the memory traces of information that has been recalled. To demonstrate the direct effect on text recall, in a classic study, Roediger and Karpicke (2006b) compared retrieval practice in the form of free recall without feedback to rereading. They found that retrieval practice without feedback led to better performance than restudying on a free recall test after one week, but not on an immediate free recall test.

To the best of our knowledge, there is only one study that has investigated this direct effect of retrieval practice with texts on a long-term inference test. Hinze and colleagues (2013, Experiment 3) had participants read three expository texts and then perform a rereading condition or one of two different retrieval practice conditions without feedback. According to their Constructive Retrieval Hypothesis, retrieval practice instructions that trigger more constructive processing will produce better learning than unguided free recall. Therefore, the instruction in their '*explain*' condition was as follows: "Practice writing an explanation of the text. Use your own words to communicate the text's explanation for how [vision] works. Your response will be scored on how completely and accurately you can explain the topic." The instruction in their *free recall* condition was the following: "Practice retrieving the content from the [vision] text. You may use your own words or those from the text. Your response will be scored on how much of the text you can recall." Your response will be scored on how completely and accurately you can explain the topic." One week later, participants received two multiple-choice tests, one measuring verbatim details from the texts and the other measuring inferences. It turned out that after a one week delay, participants in the explain condition scored higher than the other two conditions on both measures, whereas the free recall condition and the reread condition did not differ. Hence, Hinze and colleagues concluded that only retrieval practice that focused on constructing a coherent text representation seemed to benefit long-term inference performance.

Interestingly, on the multiple-choice *detail* final test in Hinze and colleagues' (2013) third experiment, which measured pure verbatim text memory, the explain condition also lead to better performance than free recall and reread, while the latter two conditions did not differ. This latter outcome is remarkable, since the free recall testing effect has been shown to be robust for final pure memory tests (for overviews, see Kornell, Bjork, & Garcia, 2011; Roediger & Karpicke, 2006a; Rowland, 2014). One would therefore also have expected an advantage of free recall compared to rereading on the detail final test in Experiment 3 by Hinze and colleagues (2013). Hence, it is possible that performance in their free recall retrieval practice condition was for some reason deprived, and did therefore not produce the standard memory benefit. This may have translated into suboptimal performance on their final inference test as well. In that case,

the benefit of the explain condition vis-à-vis the free recall condition on both final tests was not the result of the explain condition scoring relatively high, but free recall scoring relatively low.

Moreover, it could be argued that some of the multiple-choice *inference* questions used by Hinze and colleagues measured verbatim text recall rather than inference performance. For example, the first inference question on vision was the following: "For most people, what is the purpose of the lens in a pair of eye glasses? (a) They help bend the light rays for the lens of the eyes. (b) They contain a certain amount of rods and cones if a retina is low in these cells. (c) They help the iris open and close. (d) They help the pupil open and close." The part of the text necessary to answer this question was as follows: "To see things close up, the lenses in our eyes need to be thick to bend the light more. To see things far away, the lenses need to be thin to bend the light less. When people need glasses it is because their lenses aren't able to bend the light to the correct part of the eye by themselves". The correct answer to this inference question, i.e., alternative A, only requires (verbatim) recall of this part of the text, so to answer this question it was not necessary to draw an inference from the text.

Thus, in the present study, we investigated whether the conclusion from Hinze and colleagues that only retrieval practice focused on constructing a coherent text representation benefits long-term inference performance was warranted. The present study can be regarded as a conceptual replication of Experiment 3 of the study by Hinze and colleagues (2013), but with different materials, a different final test, and the final test also being administered immediately after the learning phase (besides a *delayed* final test, as used by Hinze and colleagues). For our final test, we constructed a set of open final test questions aimed at measuring inferences going beyond what was stated in the text. Verbatim text memory was not sufficient to answer these questions; participants had to apply the acquired information to a new situation (i.e., transfer of knowledge). Furthermore, we used slightly different instructions for the retrieval practice conditions. In our verbatim free recall condition, participants were asked to type verbatim everything they could *remember* from the studied text. In the constructive recall condition, which was based on the Constructive Retrieval Hypothesis (Hinze et al., 2013), participants were instructed to type in their own words what they had *comprehended* from the content of the text, allowing more room for elaboration and making inferences within and beyond the text. Feedback was not provided, in order to be able to attribute possible differences between conditions to retrieval practice alone. The dependent measure was performance on the inference test, which was presented immediately and after a 1-week interval (i.e., delayed). We predicted that after a delay, constructive recall would lead to better performance on the final inference test than verbatim free recall and reread, and that verbatim free recall would produce higher inference test scores

than reread. On the immediate test, we did not expect any differences between the conditions, in line with the findings from Roediger and Karpicke (2006b).

Baseline experiment

Before our main experiment, we conducted a baseline study where participants answered the inference questions without having read the texts that we used in our main experiment. This study made it possible to assess baseline performance on these questions.

Participants

Participants were twenty-three Dutch undergraduate students from Erasmus University Rotterdam. They received course credits or a monetary compensation for taking part in the study. Their mean age was 20.13 years (*SD* = 3.38). Five of them were males, eighteen were females.

Materials

Four expository texts in Dutch with a length of 266-306 words were used to construct sixteen inference questions. The topics of the texts were tropical cyclones, bats, snakes, and solar systems. The first and second texts were shortened and translated versions of the texts used by Butler (2010), the third and the fourth texts were created on the basis of two texts taken from a Dutch website for education in science (www.kennislink.nl). Four short-answer inferential questions per text were created, so sixteen questions in total. These questions were beyond what was stated in the texts; participants had to apply the acquired information to a new situation. For example, an inference question about the bats text was the following: "A bat that dies while hanging upside down does not fall to the ground. Please explain why". The order of questions was randomized per participant (cf. Experiment 3 by Hinze and colleagues (2013), where the order of the questions was not randomized).

Procedure

Participants were individually tested, with the experimenter being present in the same room during the experiment. They were asked to answer the inference questions on paper at their own pace. They did not read the texts in advance.

Results

One research assistant scored the answers to the inferential questions. Each correct answer to a question contained two or three elements. Participants received one point

per element, so they scored between zero and two or three points per question, with a maximum score of forty points on the total set of sixteen questions.

The mean score on the sixteen inference questions was 9.04 (*SD* = 3.97), which comes down to a proportion of .23 (*SD* = 0.10). There were no statistical differences between the four texts on the mean scores on the set of inference questions.

Method main experiment

Participants

Participants were seventy-four Dutch undergraduate students from Erasmus University Rotterdam. The data from two participants were removed because we had coincidentally used the wrong counterbalance version. Participants received course credits or a monetary compensation for taking part in the study. Their mean age was 21.07 years (*SD* = 2.37). Twenty-six of them were males, forty-six were females.

Materials and design

For materials, see the Baseline Study. The experiment had a 3 Study Method (constructive recall vs. verbatim free recall vs. reread) x 2 Retention Interval (immediate vs. delayed) mixed design with repeated measures on the second factor. Participants were randomly assigned to the levels of the between-subjects factor. The reading order of the four texts was counterbalanced by using a Latin Square method, creating eight counterbalance versions. Participants completed the inferential questions on two of the texts immediately (depending on the counterbalance version), while the questions on the two remaining texts were completed after a 1-week delay. The texts and the inference tests were presented in Dutch on a computer screen using E-Prime software. The dependent variable was the proportion of correct answers to the sixteen final inferential questions. To compute this proportion, we divided each participant's inference test score by the maximum score.

Following Simmons, Nelson, and Simonsohn (2011), we report all conditions and all measures in this experiment. For another research project, two additional dependent variables were measured in this experiment (always *after* the inference test). Because we considered these variables not to be of importance for our hypotheses, we decided *a priori* not to include them in the present paper.

Procedure

Participants were individually tested with the experimenter being present in the same room. They were told that during the first part of the experiment they would read four texts and perform a task directly after reading each text. A short practice trial was used to make participants acquainted with the procedure from the main experiment.

During a study phase, participants had five minutes to study a text at their own pace. Directly after reading a text, participants received different instructions, depending on the condition. Afterwards, they repeated the study phase for the other three texts. The instruction in the reread condition was that participants had five minutes extra study time. In the verbatim free recall condition, participants were given the following instruction: "Type as much as you can remember from the verbatim text that you have just read". The instruction in the constructive recall condition was as follows: "Describe what you have comprehended from the content of the text that you have just read. Try to do this in your own words as much as possible". For the latter two conditions, participants had 5 minutes.

After each of the four study phases, participants indicated on a 9-point Likert scale how much mental effort (e.g., Paas, Tuovinen, Tabbers, & Van Gerven, 2003) it had taken them to perform the study method, how much prior knowledge they had of the texts' topics, and to what extent they considered the text to be interesting.

During the test phase, participants received the self-paced inference tests for two of the texts immediately after the study phase. One week later they returned to complete the final inference tests on the two remaining tests. At the end of this session, participants were asked to indicate on a 5-point Likert scale how often they had thought and heard about the two remaining texts' topics during the past week.

Results main experiment

Following Kline (2004, Chapter 3) and Cumming (2014), we use the term 'statistical' instead of 'significant' for all statistical analyses, since the latter is often erroneously understood as meaning 'important'. All data from this study can be retrieved from the Open Science Framework[1].

Scoring

Two research assistants independently scored 28% of the answers to the inference questions. The intraclass correlation coefficient (with absolute agreement) between their scores was $r = .79$. The remaining responses were scored by the one of the research assistants alone.

Scale

The mean proportion correct on all sixteen inference questions (with a maximum score of 40) was.54 ($SD = .15$). Cronbach's alpha of the set of questions was 0.74, but with this sample size, the estimation of Cronbach's alpha is imprecise. We therefore also report the mean proportion correct and standard deviation per question in Table 1, and the inter-item Spearman's ρ correlation coefficients in Table 2.

1 https://osf.io/hr9vw/

Table 1. The Mean Proportion Correct and Standard Deviation per Final Test Question

	Mean	SD
Q1	.59	.30
Q2	.54	.40
Q3	.81	.27
Q4	.76	.31
Q5	.53	.33
Q6	.38	.30
Q7	.48	.34
Q8	.44	.44
Q9	.72	.33
Q10	.41	.40
Q11	.53	.40
Q12	.25	.32
Q13	.43	.26
Q14	.72	.35
Q15	.76	.29
Q16	.37	.33

Retrieval practice tests

To measure the quality of the recall protocols, we compared the written protocols by using a method based on Wiley and Voss (1999), who classified sentences as borrowed, transformed or added. A sentence was classified as *borrowed* when it contained literal of paraphrased information from the presented material. A sentence was coded as *transformed* when it contained information from the original text combined with novel information, or when two or more pieces of information were combined that were not connected in the original text. A sentence was classified as *added* when it contained only novel information. An extra category that we included in the current study was *incorrect*. Because participants did not have the text available during writing (as was the case in the Wiley and Voss study), it was possible that information from the text was retrieved erroneously, or that incorrect inferences were drawn. Therefore, a sentence was coded as *incorrect* when the information in a sentence was not in line with the original text. We expected that participants would write more borrowed sentences when asked to recall the passage verbatim (verbatim free recall condition) as compared to participants who were asked to write what they had comprehended from the content of the text (constructive recall condition). We also predicted that participants in the constructive recall condition would produce more transformed sentences than participants in the verbatim free recall condition.

Table 2. The Spearman's ρ Correlation Coefficients Between the Sixteen Final Test Questions

	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16
Q1	.06	.00	.08	.00	.26	.34	.07	.16	.18	.02	.50	.01	.06	.01	.12
Q2		.02	.26	.03	.16	.95	.18	.96	.83	.38	.99	.77	.86	.11	.47
Q3			.24	.24	.17	-.01	.09	.12	.06	.06	.02	.18	.17	.28	-.08
Q4				.21	.13	.03	.28	.05	.06	.24	-.06	.11	.28	.30	.19
Q5					.18	.21	.21	.07	.21	.38	-.02	.17	.19	.11	.18
Q6						.24	.30	.29	.36	.29	.06	.31	.31	.03	.04
Q7							.28	.14	.27	.39	.15	.01	.10	-.09	.07
Q8								.19	.26	.19	.03	.07	.43	.04	.41
Q9									.11	.11	.02	.11	.29	.03	.01
Q10										.39	.17	.13	.05	-.04	.13
Q11											.19	.22	.18	-.15	.05
Q12												.24	-.10	.19	.05
Q13													.25	.11	.18
Q14														.09	.21
Q15															.31

Independent t-tests showed a difference between verbatim free recall and constructive recall on the proportion of borrowed sentences, t (46) = 3.98, p < .001, r = .51, 95% CI of the difference [.07, .22], and on the proportion of transformed sentences, t (46) = -3.55, p = .001, r = .46, 95% CI of the difference [-.21, -.06], but not on the proportion of added sentences, t (46) = -1.36, p = .181, r = .20, 95% CI of the difference [-.02, .00], and not on the proportion of incorrect sentences, t (46) = -.13, p = .899, r = .02, 95% CI of the difference [-.04, .03]. Specifically, there were more borrowed sentences in verbatim free recall protocols (M = .54, SD = .14) than in the constructive recall protocols (M = .39, SD = .12), and more transformed sentences in constructive recall protocols (M = .50, SD = .13) than in the verbatim free recall protocols (M = .36, SD = .14). This indicates that our retrieval practice manipulations were successful in the sense that constructive retrieval practice resulted in a qualitatively different recall than verbatim free retrieval practice.

Furthermore, to determine the amount of information that participants retrieved during retrieval practice, the proportion of idea units in the written protocols was scored. Two research assistants independently scored 22% of the protocols, and the intraclass correlation coefficient between the mean proportions of idea units was .74. There was no statistical difference in the proportion of generated idea units between the verbatim free recall condition (M = .56, SD = .14) and the constructive recall condition (M = .54, SD = .11).

Final inference tests

A 3 Study Method (constructive recall vs. verbatim free recall vs. reread) x 2 Retention Interval (immediate vs. delayed) Repeated Measures ANOVA on the proportion of correct answers to the inferential questions did not yield a statistical interaction effect between study method and retention interval, F(2, 69) = .96, p = .39, *Partial* η^2 = .03 (see Table 3). There was no statistical main effect of study method, F(2, 69) = .62, p = .54, *Partial* η^2 = .02, but we did find a statistical main effect of retention interval, F(1, 69) = 9.53, p = .003, *Partial* η^2 = .12, 95% CI of the difference [.03, .15]. The proportion correct answers to the final inferential questions was larger in the immediate condition (M = .59, SD = .19) than in the delayed condition (M = .50, SD = .19).

Because a non-significant p-value does not provide evidence for a null effect (e.g., Rouder, Speckman, Sun, Morey, & Iverson, 2009; for more arguments for the Bayesian approach, see Dienes, 2011), we also performed a Bayesian Repeated Measures ANOVA (Rouder, Morey, Speckman, & Province, 2012; Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2017) in the software program JASP (Love et al, 2015; Wagenmakers et al., 2016). We used a Cauchy prior width of r = 0.30 for effect size on the alternative hypothesis because we expected a small effect size. The Bayes Factor for the interaction effect between study method and retention interval is BF01 = 2.29, indicating that

the observed data were 2.29 more likely under the null hypothesis that postulates the absence of the interaction effect than under the alternative hypothesis that postulates the presence of the effect. This is anecdotal evidence for the null hypothesis of the interaction effect (e.g., Wetzels & Wagenmakers, 2012). The Bayes Factor for the factor study method is BF01 = 3.45, indicating that the observed data are 3.45 more likely under the null hypothesis than under the alternative hypothesis. This is moderate evidence in favour of the null hypothesis of no difference between the three study methods on the proportion of correct answers to the final inference test. The Bayes Factor for the factor retention interval was BF10 = 17.85, meaning that the observed data are 17.85 times more likely under the alternative hypothesis than under the null hypothesis. This is strong evidence for the alternative hypothesis that postulates a difference between the two retention intervals on the final inference test, indicating forgetting over the one-week interval.

Exploratory analyses

Five extra variables had been measured on which we did not have a priori hypotheses, so we decided to analyse them in an exploratory matter. Participants indicated on a 9-point Likert scale how much mental effort it had taken to carry out the study method, how much prior knowledge they had of the texts' topic, and to what extent they considered the text interesting. Additionally, after the 1-week interval, participants indicated how often they had thought and heard about the text's topics during the past week on a 5-point Likert scale.

Table 3. Proportion of Answers Correct by Retention Interval and Study Method. Standard Errors are Between Brackets

	Retention Interval	
Study Method	**Immediate**	**Delayed**
Reread	.63 (.04)	.49 (.04)
Verbatim free recall	.53 (.04)	.49 (.04)
Constructive recall	.59 (.04)	.50 (.04)

An ANOVA on how much effort it took participants to perform the study method, measured on a 9-point Likert scale averaged over texts and ranging from 1 (almost no effort) to 9 (a very large amount of effort), yielded a difference between constructive recall (M = 4.75, SD = 1.05), verbatim free recall (M = 5.90, SD = 1.04), and reread (M = 3.83, SD = 1.25), $F(2, 69) = 20.49$, $p < .001$, *Partial* $\eta^2 = .37$. Bonferroni corrected pairwise comparisons revealed statistical differences between all three conditions. Constructive recall required less effort than verbatim free recall, $t(69) = -3.55$, $p = .002$, $r = .39$, 95% CI

of the difference [-1.94, -0.35], constructive recall took more effort than rereading, $t(69) = 2.84$, $p = .018$, $r = .32$, 95% CI of the difference [0.12, 1.71], and verbatim free recall took more effort than rereading, $t(69) = 6.39$, $p < .001$, $r = .61$, 95% CI of the difference [1.27, 2.85].

Furthermore, we did not find a statistical difference between the three conditions on participants' subjective prior knowledge, measured on a scale ranging from 1 (almost no prior knowledge) to 9 (a very large amount of prior knowledge), $F(2, 69) = 1.34$, $p = .268$, *Partial* $\eta^2 = .04$ (overall mean and standard deviation: $M = 3.30$, $SD = 1.24$). In addition, there was no statistical difference between conditions with regards to the extent that the texts were considered interesting, with a scale from 0 ('very, very uninteresting') to 9 ('very, very interesting'), $F(2, 69) = 0.05$, $p = .954$, *Partial* $\eta^2 = .00$ (overall mean and standard deviation: $M = 5.70$, $SD = 1.19$). Finally, participants were asked how often they had thought back and heard about the texts' topics during the one-week retention interval, with possible scores of 1 (1 to 5 times), 2 (6-10 times), 3 (11-15 times), 4 (16-20 times), and 5 (more than 20 times). Eighty percent of the participants in all three conditions choose option 1 or 2 when asked how often they had thought back to the texts' topics. In addition, more than 87 percent of the participants in all three conditions selected option 1 when asked how often they had heard about the topics during the one-week interval.

Discussion

According to the Constructive Retrieval Hypothesis (Hinze et al., 2013), the elaboration resulting from constructive retrieval practice promotes higher-level learning, such as drawing inferences from a text. Retrieval practice instructions that trigger more constructive processing are therefore expected to produce better higher-level learning than unguided free recall. The purpose of the present experiment was to examine whether constructive recall would indeed be more successful in drawing inferences from a text than verbatim free recall. Specifically, the present experiment can be considered a conceptual replication of Hinze and colleagues' third experiment, but with a different final inference test and slightly different instructions for the retrieval practice conditions. Specifically, we constructed a set of open final test inference questions because some final multiple-choice inference questions employed by Hinze and colleagues may have actually measured verbatim text recall rather than text inference.

The results from our retrieval-practice phase showed that there were more borrowed sentences in the verbatim free recall protocols than in the constructive recall protocols, and more transformed sentences in constructive recall protocols than in the verbatim free recall protocols. These results indicate that our study indeed led to different kinds of recall. However, the results showed that on the immediate and the delayed final tests,

neither verbatim free recall nor constructive recall fostered inference making more than rereading did.

Our findings deviate from those observed by Hinze, Wiley, and Pellegrino (2013, Experiment 3). Now, it might be possible that the two constructive retrieval practice instructions, i.e., "*explain in your own words"* (from the study by Hinze and colleagues) and "*indicate what you comprehend from the text"* (from the present study) triggered different cognitive processes. However, prima facie both instructions seem to encourage constructive processing and drawing inferences within and beyond the text, which is necessary to build a situation-model level representation. One would expect that our instruction to comprehend would stimulate students to give in an *explanation* of the process described in the expository text, thereby leading to similar outcomes on the final inference test as the explain instruction in Hinze and colleagues (2013, Experiment 3). An avenue for future research could therefore be a study in which their explain prompt and our constructive recall prompt are compared directly.

Generally, measuring inferences is complex (e.g., Barnett & Ceci, 2002), because often inference questions (partly) measure cued recall rather than straightforward inference making (e.g., Tran, Rohrer, & Pashler, 2014). Indeed, the multiple-choice questions used by Hinze and colleagues might have assessed verbatim text recall rather than pure inference drawing. In the present study, we therefore used a set of open final test questions aimed at measuring inferences going beyond what was stated in the text. Our study may therefore be viewed as an extended version of the third experiment by Hinze and colleagues (2013), although there is always a possibility that our final test did not fully succeed in measuring pure inferences either.

In conclusion, the present results seem to suggest that constructive recall is not particularly useful for higher level learning from text, at least when no feedback is provided. The results of studies using this kind of final test (e.g., Eglington & Kang, in press; Hinze et al., 2013; McDaniel et al., 2013; Tran, Rohrer, & Pashler, 2014) show mixed results, suggesting that retrieval practice effects might be weaker with final inference tests than with pure memory tests. However, Eglington and Kang (in press) and Tran, Rohrer, and Pashler (2014) used sets of related premises instead of pure expository text. Still, to the best of our knowledge, the study by Hinze, Wiley, and Pellegrino (2013) is the only one to demonstrate a long-term inference effect of constructive retrieval practice without feedback.

The present study showed no differences between a constructive recall condition, a verbatim free recall condition, and a reread condition on a final inference test. This raises the question what kind of recall retrieval practice instructions – if any – can enhance drawing inferences from text. Future research should therefore be aimed at further exploring if and how retrieval practice can facilitate long-term inference learning.

5

The retrieval practice effect for expository text: Small and only when feedback is provided

This chapter is in preparation as:
Van Eersel, G. G., Verkoeijen, P. P. J. L., De Jonge, M. O., & Rikers, R. M. J. P (in preparation). The retrieval practice effect for expository text: Small and only when feedback is provided.

Abstract

Previous studies (e.g., Van Eersel, Verkoeijen, Tabbers, Van Mierlo, Paas, & Rikers, submitted) did not find a difference on a final inference test between retrieval practice and rereading. Possibly, these findings were due to the nature of the final inference test. We therefore conducted an experiment with the same materials as in the earlier study by Van Eersel and colleagues, but now with a final free recall memory test. With this kind of memory test, the beneficial effect of retrieval practice over restudying has been well established. In Experiment 1, we observed a small benefit of reread over free recall after five minutes, but after one week, no difference between conditions was left. We then performed a study comparable to the first experiment, but with restudy (i.e., feedback) after the free recall phase. In this second experiment, only a small benefit of free-recall-plus-feedback emerged relative to reread. Together these findings suggest that feedback is required to obtain a beneficial effect of free recall retrieval practice on learning from expository text, but that the effect is small even when feedback is provided.

Taking a test on previously studied material is a powerful way to promote long-term learning, a phenomenon referred to as the retrieval practice (testing) effect (for reviews, see Delaney, Verkoeijen, & Spirgel, 2010; Karpicke, 2012; Karpicke, Lehman, & Aue, 2014; Roediger & Karpicke, 2006a; Roediger, & Butler, 2011; Rowland, 2014). In a typical retrieval practice study, participants learn a set of stimuli during an initial study phase either by restudying or by retrieval practice. After a short retention interval of five minutes, performance in the restudy condition is generally better than, or comparable to, performance in the retrieval practice condition. After a long retention interval (i.e., more than one day), however, retrieval practice is typically more effective than restudy, giving rise to an interaction effect of study method and retention interval on performance (e.g., Hogan & Kintsch, 1971; Roediger & Karpicke, 2006b; Wheeler, Evans, & Buonanno, 2003).

There are many different ways to design retrieval-based learning activities, but the literature suggests that free recall and cued recall are in general more beneficial for learning than other retrieval practice formats, like fill-in-the-blank and recognition tests (e.g., Butler & Roediger, 2007; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Glover, 1989; McDaniel, Anderson, Derbish, & Morrisette, 2007; Rowland, 2014). Moreover, free recall in particular is easy to implement; the only necessary thing to do is to instruct learners to attempt to retrieve in any order the information that was previously studied. Indeed, several studies have shown that free recall produces large advantages of retrieval practice as compared to restudy, especially after a long interval (for overviews, see Kornell, Bjork, & Garcia, 2011; Roediger & Karpicke, 2006a; Rowland, 2014).

A substantial part of the retrieval practice research has focused on simple materials, such as wordlists and paired associates. However, increasing attempts have been made to examine the retrieval practice effect with materials that are more educationally relevant, like expository texts (e.g., Glover, 1989; Kang, McDermott, & Roediger, 2007; Nungester & Duchastel, 1982). In a non-exhaustive literature review, we found twenty-five studies that demonstrated a benefit of retrieval practice over restudy with text materials. Eight of them used short answer questions (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Butler, 2010; Chan, McDermott, & Roediger, 2006; Dirx, Kester, & Kirschner, 2014; Kang et al., 2007; McDaniel et al., 2007; Weinstein, McDermott, & Roediger, 2010; Wooldridge, Bugg, McDaniel, Liu, 2014), four used multiple choice tests (Butler, & Roediger, 2008; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; Roediger, Agarwal, McDaniel, & McDermott, 2011; Spitzer, 1939), and three studies used fill-in-the-blank tests (Chan, 2010; De Jonge, Tabbers, & Rikers, 2015; Hinze, & Wiley, 2011). The remaining eleven studies employed free recall as their retrieval practice format, and in six of them some form of feedback was provided, for example a restudy opportunity (Blunt,

& Karpicke, 2014; Dobson & Linderholm, 2015; Karpicke, & Blunt, 2014; McDaniel, Howard, & Einstein, 2009; Karpicke, & Roediger, 2010; Smith, Blunt, Whiffen, & Karpicke, 2016).

To the best of our knowledge, there are only four studies with text materials that showed a long-term benefit of free recall over restudy without providing feedback after the retrieval practice phase (Glover, 1989, Experiment 4; Hinze, Wiley, and Pellegrino, 2013, Experiment 3; Karpicke, & Roediger, 2010, Experiment 2; Roediger, & Karpicke, 2006b, Experiment 1), and one study that did not show an advantage of free recall without feedback (Duchastel, 1981). Of these four studies, Karpicke and Roediger (2010, Experiment 2) and Roediger and Karpicke (2006b) used a one-week delayed free recall task as final test. Glover (1989, Experiment 4) used three final tests: recognition, cued recall, and free recall, which were administered four days after the first study phase and two days after the intervening test. Duchastel (1981) administered both a topical retention test and a short-answer test after a one-week retention interval. Finally, Hinze and colleagues (2013, Experiment 3) administered two multiple choice tests after one week, one measuring details and the other measuring inferences. Taken together, only one of these four studies employed a final inference test, measuring what had been learned by connecting multiple concepts from the text to solve a new problem. Now, the ability to apply previously learned knowledge or skills in a novel context (i.e., transfer) is an important aim of education (Carpenter, 2012; Rohrer, Taylor, & Sholar, 2010). From an educational perspective, it is therefore an important question whether free recall without feedback is indeed beneficial for solving new problems. As indicated, only the study by Hinze and colleagues (2013, Experiment 3) showed this to be the case, but their final test had a multiple-choice format and possibly measured verbatim text recall rather than inference performance (Van Eersel et al., submitted).

A recent study (Van Eersel et al., submitted) therefore explored the effect of different types of free recall instructions on a final inference test. This final inference test consisted of sixteen short-answer questions about the texts, and required participants to apply the acquired information from the texts to a new situation. Participants read four expository texts, and then reread, recalled the texts verbatim, or recalled the texts constructively, both these recall conditions without feedback. In the *verbatim free recall* condition, participants were asked to type in as literally as possible (verbatim) everything they could *remember* from the studied texts. In the *constructive recall* condition, participants were instructed to type in their own words what they had *comprehended* from the content of the text, allowing more room for elaboration and making inferences within and beyond the text (e.g., Hinze et al., 2013). In the *reread* condition, participants received five minutes extra study time. On the final test administered immediately and after one week, we observed forgetting over time. However, the three study conditions did not differ on either of the two measurements.

These results raise the question why Van Eersel and colleagues (submitted) did not find a free recall testing effect with expository text on a final inference test. There is a possibility that retrieval practice is not particularly useful for inference learning, especially when no feedback is provided. Only a small number of studies has shown retrieval practice to produce better performance on a final inference test (e.g., Blunt & Karpicke, 2014; Butler, 2010; Johnson & Mayer, 2009; Karpicke & Blunt 2011; McDaniel et al., 2009; Van Eersel, Verkoeijen, Povilenaite, & Rikers, 2016), and only one of these studies used free recall without feedback as the retrieval practice format (Hinze et al., 2013). Together these studies suggest that the retrieval practice effect might be weaker with final inference tests. However, a different explanation for the findings by Van Eersel and colleagues (submitted) is that the specific text materials that they used were not suitable to produce a testing effect, not even on a pure retention final test. We therefore decided to perform a study with the same materials as Van Eersel and colleagues, in which we compared a simple free recall condition to a reread condition. In contrast to Van Eersel and colleagues, instead of a final inference test we used a final free recall memory test. With this kind of memory test, the beneficial effect of retrieval practice over restudying has been well established (e.g., Rowland, 2014). So, if the results of Van Eersel and colleagues (submitted) were exclusively driven by the characteristics of the final inference test, then replacing this kind of final test with a final free recall memory test would (most likely) deliver a classic retrieval practice effect.

Note that our first experiment was modelled after the first experiment by Roediger and Karpicke (2006b), because their experiment has been the only one that included both a free recall condition without feedback, and a final free recall memory test after a short as well as a long retention interval. Roediger and Karpicke (2006b, Experiment 1) asked participants to read two prose passages and afterwards to reread one of the passages, and to recall the other passage by writing down everything they could remember. The final free recall test was administered after five minutes, two days or one week. Rereading produced a better final test performance than free recall retrieval practice after five-minutes, but an advantage of retrieval practice emerged in the two-days and one-week conditions, giving rise to the typical cross-over interaction effect between study condition and retention interval (e.g., Kornell et al., 2011). Our first experiment was comparable to this first experiment by Roediger and Karpicke (2006b), but with different text material, because we used the two texts of the study by Van Eersel and colleagues (submitted). We also ask participants to type instead of write responses during retrieval practice, and we did not have a two-days retention interval.

In the present study, participants read two text passages and then reread one of the passages and recalled the other passage. The final free recall test was administered after five minutes or after one week. In Experiment 1, no feedback was provided. In

Experiment 2, a restudy opportunity was provided after free recall, because feedback increases the beneficial effect of retrieval practice (e.g., Agarwal et al., 2008; Erdman & Chan, 2013; Hays, Kornell, & Bjork, 2013; Kang et al., 2007; Pashler, Cepeda, Wixted, & Rohrer, 2005; Pashler, Kang, & Mozer, 2013).

Experiment 1

Method

Following Simmons, Nelson, and Simonsohn (2012), we report how we determined our sample size, all data exclusions, all manipulations and all measures in this experiment. All materials and data from this study can be retrieved from the Open Science Framework[1].

Participants

The participants were sixty-four Dutch undergraduates from the Erasmus University Rotterdam, the Netherlands, who were rewarded with course credits. A power analysis in G*Power indicated that a total number of thirty-four participants was sufficient to detect an effect of size $\eta_p^2 = .24$, which was the smallest effect size reported in the first experiment by Roediger and Karpicke (2006b), namely for the factor Learning Condition. However, we decided *ex ante* to have at least thirty participants per between-subjects condition, which resulted in a total of sixty-four participants (given that the experiment had four counterbalance versions). Four participants dropped out because of a computer error, and another four dropped out because they did not show up for the second session after one week. Drop-outs were replaced by new participants to maintain a fully counterbalanced design. The mean age of the sixty-four participants in the final sample was 19.80 years (*SD* = 2.21). Fifty-one participants were women and thirteen were men.

Materials and design

The two expository texts we used had a length of 301 and 254 words. One text explained how the claw of a bat operates, and the other text explained how a snake climbs. The first text was a shortened version of a text used in Butler (2010), translated into Dutch. The second text was created on the basis of a text taken from a Dutch website for education in science (www.kennislink.nl). Texts were presented in Dutch on a computer screen using E-Prime software.

A 2 Study Method (reread vs. free recall) × Retention Interval (five minutes vs. one week) mixed design was used, with repeated measures on the first factor. Participants

1 https://osf.io/2jdxp/

were randomly assigned to one of the levels of the between-subjects factor. The order of study methods (reread or free recall) and the order of passages (snakes or bats) were counterbalanced. The dependent variable was the proportion of correctly recalled idea units during the final free recall test.

Procedure

The experiment was conducted on a computer using E-Prime software. Participants were individually tested during two sessions. In the first session, they read on the computer screen that the session consisted of four seven-minutes phases. All participants first read the first passage for seven minutes. Participants in the *reread* condition then restudied the passage for seven minutes. They were asked to continue reading until the time had passed, and to push Enter every time that they had read the whole text. In the *free recall* condition, participants were presented with a passage topic ("Bats" or "Snakes"), and were asked to type in as much as they could remember from the studied text during seven minutes, without concern for exact wording or correct order. Cumulative recall data were recorded by registering what was recalled per minute. After completion of one of these two study conditions (i.e., reread or free recall), participants read the other text for seven minutes, and then completed the second part of session 1. Between the four phases, participants solved multiplication problems for two minutes. Session 2 occurred after a five-minute or after a one-week retention interval. In the second session, participants were asked to recall the passages that they had learned in the first session. The recall instructions were identical to those given in session 1, and participants first recalled the passage that they had also studied first during session 1. Each final retention test lasted ten minutes.

Results

Scoring

The texts were divided into 27 (bats) and 26 (snakes) idea units, and the dependent variable was the proportion of correctly recalled idea units. One point was credited when an idea unit was fully stated, while 0.5 point was credited when an idea unit was only partly stated. Two research assistants independently scored all the final free recall protocols on the proportion of recalled idea units. The Pearson's correlation coefficient (*r*) between their scores was $r = .87$. The scores of one of the raters were used for the initial and final tests analyses.

Initial study

On average, participants read the first text 4.89 times (*SD* = 1.76) and the second text 4.98 times (*SD* = 2.86). In the reread condition, the average number of rereading was 5.25 (*SD* = 2.36). Furthermore, on the initial 7-min free recall test, participants recalled on average 11.78 idea units (*SD* = 3.73), which is 44.45% of the passage. There was no difference between the two retention interval conditions or counterbalance versions in the proportion of recalled idea units. Note that due to a computer error, the initial data of four participants were not included in these calculations.

Final tests

Following Kline (2004) and Cumming (2012), we use the term 'statistical' instead of 'significant' for all statistical analyses, since the latter is often erroneously understood as meaning 'important'. We checked whether the final retention phase was sufficiently long through the cumulative recall data, which showed that participants had exhausted their knowledge by the end of the phase (cf. Roediger & Karpicke, 2006b, p. 250). A 2 Study Method (reread vs. free recall) * 2 Retention Interval (five minutes vs. one week) Repeated Measures ANOVA on the proportion recalled idea units did not show a statistical main effect of study method, F (1, 62) = 2.88, p = .095, *Partial* η^2 = .04, 95% CI of the difference [-.01, .07]. The analysis did yield a statistical main effect of retention interval, F (1, 62) = 19.81, p < .001, *Partial* η^2 = .24, 95% CI of the difference [.08, .21], which showed forgetting over the one-week interval. The analysis further showed a trend towards a statistical interaction effect between study method and retention interval, F (1, 62) = 3.21, p = .078, *Partial* η^2 = .05 (see Table 1). Because this interaction effect was in line with our hypotheses, we performed two two-sided follow-up T-Tests showing that there was a statistical advantage of reread over free recall after five minutes, t (31) = 2.34, p = .026, *Cohen's d* = .41, 95% CI of the difference [.01, .12], but not after one week, t (31) = -0.07, p = .944, *Cohen's d* = .01, 95% CI of the difference [-.05, .05].

Because a non-significant *p*-value does not provide evidence for a null effect (e.g., Rouder, Speckman, Sun, Morey, & Iverson, 2009; Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2017), we also performed a Bayesian Repeated Measures ANOVA in the software program JASP (Love et al, 2015; Wagenmakers et al., 2016). We used a Cauchy prior width of r = 0.21 for effect size on the alternative hypothesis because we expected a small effect size (Rouder, Morey, Speckman, & Province, 2012) (for more arguments for the Bayesian approach, see, e.g., Dienes, 2011). The Bayes Factor for the factor study method was BF10 = 1.02, meaning that the likelihood of the data under the alternative hypothesis was 1.02 times the likelihood of the data under the null hypothesis. A Bayes Factor so close to 1 can be interpreted as no evidence for either of the two hypotheses. The Bayes Factor for the factor retention interval was BF10 =

313.44, which can be interpreted as decisive evidence (Wetzels & Wagenmakers, 2012) in favour of the alternative hypothesis of a difference between the two intervals on the final test, in this case an advantage of the five-minutes condition compared to the one-week condition. The Bayes Factor for the interaction effect between study method and retention interval on the proportion of recalled idea units was BF10 = 1.19, showing that the likelihood of the data under the alternative hypothesis that postulates the presence of the interaction effect was 1.19 times the likelihood of the data under the null hypothesis. A Bayes Factor so close to 1 indicates no support for either the null or the alternative hypothesis. In line with the frequentist analyses above, for each of the two retention intervals we conducted a two-sided Bayesian T-Test (Rouder et al., 2009) for the factor study method. We used a Cauchy prior width of $r = 0.30$ for effect size on the alternative hypothesis (which is equivalent to the prior width of $r = 0.21$ used for the Bayesian Repeated Measures ANOVA above, see Wagenmakers et al., 2016). In the five-minutes condition, the Bayes Factor of study method was BF10 = 3.54, which means that the likelihood of the data under the alternative hypothesis was 3.54 times the likelihood of the data under the null hypothesis. This is moderate evidence in favour of the alternative hypothesis, in this case a benefit of reread over free recall. In the one-week condition, the difference between reread and free recall yielded a Bayes Factor of BF01 = 2.64, indicating that the observed data were 2.64 times more likely under the null hypothesis than under the null alternative hypothesis. This is anecdotal evidence in favour of the null hypothesis of no difference between free recall and reread.

In sum, both the frequentist and the Bayesian approach show no clear evidence for an interaction effect and a large effect for the factor retention interval. Furthermore, both analyses demonstrate a modest advantage of reread over free recall after five minutes, but not after one week.

Table 1. Proportion of Recalled Idea Units in Experiment 1 by Retention Interval and Study Method. Standard Errors are Between Brackets

	Retention Interval	
Study Method	**Five minutes**	**One week**
Reread	.53 (.03)	.35 (.03)
Free Recall	.46 (.03)	.35 (.03)

Discussion

In a recent study by Van Eersel and colleagues (submitted), no difference occurred between a verbatim free recall condition, a constructive recall condition, and a reread condition on a final inference test. Possibly, this result was driven by the nature of the final inference test. Therefore, in our first experiment we compared a free recall

condition to a reread condition on a final free recall test, using the same materials as Van Eersel and colleagues. With this kind of memory test (i.e., free recall), the beneficial effect of retrieval practice over restudy has been shown to be robust (e.g., Rowland, 2014). If the outcomes of the study by Van Eersel and colleagues were due exclusively to the nature of the final test, then the present experiment would reinstate the typical retrieval practice interaction effect. However, we found a small benefit of rereading over free recall after five minutes, and no difference between conditions after one week. Although these findings were in the expected direction, the interaction effect was too small to be statistically significant. These results suggest that the outcomes as observed by Van Eersel and colleagues may not exclusively be explained by the nature of the final test.

Why did we not obtain this interaction, or an advantage of free recall retrieval practice after one week? A possible explanation is that our initial retrieval practice scores were relatively low, which may have limited the advantage of retrieval practice compared to rereading. According to the bifurcation framework (Halamish & Bjork, 2011; Kornell et al., 2011), memory traces of items that are successfully recalled during testing are strengthened more than memory traces of items that are restudied, which produces a final test advantage for tested items compared to restudied items. However, if only a limited number of items is retrieved in the initial test, this advantage might not occur, particularly when no feedback is provided. In line with this framework, a meta-analysis (Rowland, 2014) showed that the magnitude of the testing effect increased with initial retrieval practice performance and that no retrieval practice effect occurred when initial performance was below 50%. Indeed, in the first experiment by Roediger and Karpicke (2006b), participants recalled about 70% of the idea units, while in our Experiment 1 this percentage was only 44.45%. However, in a study by Wheeler and colleagues (2003), who used free recall of the learned words as the final measure and whose outcome pattern was comparable to ours (i.e., an advantage of restudy over free recall after five minutes, but no difference after two days (Experiment 1), and an advantage of free recall after one week (Experiment 2)), participants initially recalled 22% of the words in Experiment 1 and 28% in Experiment 2. These latter findings suggest that initial retrieval success can only partly explain why we did not observe an advantage of retrieval practice.

Another factor that may explain the results of our first experiment is that we did not provide feedback after free recall, while feedback is known to enhance the testing effect (e.g., Agarwal et al., 2008; Erdman & Chan, 2013; Hays et al., 2013; Kang et al., 2007; Pashler et al., 2005; Pashler et al., 2013). Specifically, in six out of the eleven testing effect studies with free recall as the retrieval practice format (see Introduction), feedback was provided (Blunt, & Karpicke, 2014; Dobson & Linderholm, 2015; Karpicke & Blunt, 2014; McDaniel et al., 2009; Karpicke & Roediger, 2010; Smith et al., 2016). For example,

McDaniel and colleagues (2009, Experiment 2) asked participants to read two texts on mechanical devices. Afterwards, participants in the read-recite-review condition (3R) recited as much as possible from the texts into a tape recorder, and they then read the text again (i.e., feedback). In the note-taking condition, participants were instructed to read each text twice and take notes on the passages while reading. Participants in the reread condition were simply instructed to read each text twice. On a final free recall test administered immediately and after one week, the 3R condition performed better than the other two conditions.

To further explore if we could find a retrieval practice effect on a free recall final test using the same materials as Van Eersel and colleagues (submitted), we decided to perform a second experiment that was similar to the first, but now with feedback provided after free recall. Note that in this second experiment, we shortened the time-on-task in the read and reread phases from seven to five minutes. Because participants in Experiment 1 on average read the texts 4.94 times for the first time and 5.25 times for the second time, they might have gotten fatigued and bored during the experiment. It is important, however, to create a testing situation that minimizes participant fatigue and boredom, because such states can reduce motivation and increase the chance of response biases (Furr & Bacherach, 2014). We therefore checked whether motivation decreased during Experiment 2 by using a questionnaire asking participants how interesting it was to read and reread the passages.

In Experiment 2, we expected a benefit of free recall with feedback compared to rereading to emerge after one week. Furthermore, because in several studies where feedback was provided, testing effects already emerged after a short retention interval (e.g., Bishara & Jacoby, 2008; Carpenter, Pashler, Wixted, & Vul, 2008; Carrier & Pashler, 1992; Jacoby, Wahlheim, & Coane, 2010; Kang, 2010; Kornell, Hays, & Bjork, 2009; Wartenweiler, 2011), we also predicted to observe an advantage of free recall after five minutes.

Experiment 2

Method

Following Simmons and colleagues (2012), we report all data exclusions, all manipulations and all measures in this experiment.

Participants

As in Experiment 1, the participants were sixty-four Dutch undergraduates from the Erasmus University Rotterdam, the Netherlands, who were rewarded with course credits. Ten participants dropped out because of a computer error during the first session. One

participant dropped out because she had recently seen the text about bats in another experiment, and one other participant dropped out because he cancelled the second session for personal reasons, and one participant dropped out because the experiment leader had coincidentally started the second session immediately after the first, while the participant was in the one-week condition. They were all replaced by participants in the same counterbalance condition. The mean age of the remaining sixty-four participants was 20.38 years (*SD* = 2.95). They were all replaced by participants in the same counterbalance condition. Forty participants were women, twenty-four were men.

Materials and design

The materials were the same as in Experiment 1, plus a questionnaire with six questions asking people to rate on a 7-point scale (*1 = very boring, 7 = very interesting*) how interesting it was to read the passage for the first, second and third time (reread condition), and how interesting it was to read the passage, retrieve it, and read it again (free-recall-plus-feedback condition) (cf. Roediger & Karpicke, 2006b, Experiment 2).

A 2 Study Method (reread vs. free-recall-plus-feedback) * Retention Interval (five minutes vs. one week) mixed design was used, with repeated measures on the first factor. Participants were randomly assigned to one of the levels of the between-subjects factor. The order of study methods (reread or free-recall-plus-feedback) and the order of passages (snakes or bats) were counterbalanced.

Procedure

The procedure was comparable to that in Experiment 1 except for some changes, which are mentioned here. In the first session, participants read on the computer screen that the session consisted of four phases of five to seven minutes. All participants first read the first passage for five minutes. Participants in the reread condition then first restudied the passage for seven minutes, and then restudied it again for five minutes. In the free-recall-plus-feedback condition, participants were asked to type in as much as they could remember from the studied text during seven minutes, and then to read the text again for five minutes. After completion of one of the two study conditions (i.e., reread or free-recall-plus-feedback), participants read the other text for seven minutes, and then finished the second study condition. At the end of the learning phase, all but the first fourteen participants received a questionnaire asking them how interesting it was to read the passages. Session 2 was identical to that in Experiment 1.

Results

Scoring

All the protocols were scored by the same research assistant that scored the first experiment.

Initial study

On average, participants read the first presented text 4.80 times (*SD* = 2.06) and the second presented text 4.48 times (*SD* = 2.56). However, due to a programming error, less than half of the participants were included in these calculations (N=25 and N=23, respectively). Furthermore, in the reread condition, participants read the text 3.88 times (*SD* = 2.35) for the second time (N=52), and 3.00 times (*SD* = 1.47) for the third time (N=49). In the free-recall-plus-feedback condition, the average number of times that participants reread the text after the free recall phase was 3.02 times (SD = 1.51) (N=50). In addition, on the initial 7-min free recall test, participants recalled 11.50 idea units on average (*SD* = 3.58) (N=64), which is 43.40% of the passage. There was no difference between the two retention interval conditions or counterbalance versions on the proportion of recalled idea units.

Final questionnaire

Note that these analyses are exploratory and based on fifty participants, as we introduced the questionnaire only after already having tested fourteen participants. The questionnaire given at the end of phase 1 showed an average rating of 4.78 (*SD* = 1.26) on the question how interesting it was to read either passage for the first time (*1 = very boring, 7 = very interesting*). In the reread condition, the second and third times reading were rated with average scores of 3.62 (*SD* = 1.24) and 1.90 (*SD* = 1.04), respectively. A Repeated Measures ANOVA showed a statistical decrease in interest ratings from the first to the second and third time reading in the reread condition, F (1.84, 90.29) = 150.76, $p < .001$, *Partial* $\eta^2 = .76$. Furthermore, in the free-recall-plus-feedback condition, the activity of free recall received a rating of 4.62 (*SD* = 1.23), and rereading (i.e., the feedback) was rated with an average score of 3.22 (*SD* = 1.80). A Repeated Measures ANOVA on the difference in interest rating showed a statistical effect between the first and the second time reading in the free-recall-plus-feedback condition, F (1, 49) = 36.29, $p < .001$, *Partial* $\eta^2 = .43$, CI of the difference [1.04, 2.08]. In sum, there was a significant decrease in interest ratings in both study conditions, which suggests that our choice to shorten the reading times from seven (Experiment 1) to five minutes (Experiment 2) was justified, at least from this motivational point of view. Moreover, a Repeated Measures ANOVA yielded a statistical difference in interest ratings between reading the text for

the last time in the reread condition (M = 1.90) versus in the free-recall-plus-feedback condition (M = 3.22), F (1, 49) = 34.18, p < .001, *Partial* η^2 = .41, CI of the difference [0.87, 1.77].

Final tests

The cumulative recall data showed that participants had exhausted their knowledge by the end of the retention phase and are not reported here. A 2 Study Method (reread vs. free-recall-plus-feedback) * 2 Retention Interval (five minutes vs. one week) Repeated Measures ANOVA on the proportion recalled idea units did not show a statistical interaction effect between study method and retention interval, F (1, 62) = .04, p = .847, *Partial* η^2 = .001 (see Table 2). The analysis did yield a statistical main effect of retention interval, F (1, 62) = 30.78, p < .001, *Partial* η^2 = .33, 95% CI of the difference [.10, .21], with better performance on the final free recall test after five minutes than after one week. There was also a statistical main effect of study method, F (1, 62) = 4.03, p = .049, *Partial* η^2 = .06, 95% CI of the difference [.00, .07], showing a small advantage of free-recall-plus-feedback compared to rereading.

Table 2. Proportion of Recalled Idea Units in Experiment 2 by Retention Interval and Study Method. Standard Errors are Between Brackets

	Retention Interval	
Study Method	**Five minutes**	**One week**
Reread	.49 (.03)	.33 (.03)
Free Recall	.52 (.02)	.37 (.02)

We also conducted a Bayesian Repeated Measures ANOVA in the software program JASP (Love et al, 2015) with a Cauchy prior width of r = 0.21 for effect size on the alternative hypothesis (Rouder et al., 2012). The Bayes Factor for the interaction effect between study method and retention interval was BF01 = 2.16, showing that the likelihood of the data under the null hypothesis that postulates the absence of the interaction effect was 2.16 times the likelihood of the data under the alternative hypothesis that postulates the presence of the effect. This is anecdotal evidence for the null hypothesis of no interaction effect (Wetzels & Wagenmakers, 2012). The Bayes Factor for the factor retention interval was BF10 = 10791.88, which is decisive evidence in favour of the alternative hypothesis of a difference, in this case an advantage of the five-minutes condition compared to the one-week condition. The Bayes Factor for the factor study method was BF10 =1.58, indicating that the observed data were 1.58 times more likely under the alternative hypothesis than under the null hypothesis, which is anecdotal evidence in favour of the alternative hypothesis.

In sum, both the frequentist and the Bayesian approach found no interaction effect and a large effect for the factor retention interval. Furthermore, the two approaches diverged with respect to the factor Study Method. In the frequentist analysis, the observed p-value was $p = .049$, which falls just below the (conventional) significance threshold of $p < .05$. Although from the frequentist perspective we would now conclude that there is a difference between study methods, the Bayes Factor for the factor study method was only BF10 =1.58, which is only anecdotal evidence in favour of the alternative hypothesis that postulates a difference between study methods. However, this difference between the two statistical approaches is not that surprising, as Bayes Factors and *p*-values almost always agree on the *direction,* but often not on the *strength* of the evidence (Wetzels, Matzke, Lee, Rouder, Iverson, & Wagenmakers, 2011), especially when the *p*-value falls just below the threshold. Moreover, on theoretical grounds it could be argued that feedback would enhance the beneficial effect of retrieval practice. In that case, a *one-sided* Bayesian T-Test would be permitted. With a Cauchy prior of 0.30 on the alternative hypothesis stating the benefit of retrieval practice over rereading, this one-sided Bayesian T-Test delivers a BF10 of 2.90, which is stronger but still anecdotal evidence in favour of the alternative hypothesis (Wetzels & Wagenmakers, 2012). Taken together, the results clearly show that the advantage of free-recall-plus-feedback compared to reread is modest.

Discussion

In Experiment 2, a reread and a free-recall-plus-feedback condition were compared on a final free recall test administered after five minutes and after one week. The results show a modest advantage of free-recall-plus-feedback relative to reread. These outcomes are in contrast with those of Experiment 1, where only a small benefit of reread occurred after five minutes. The pattern found in Experiment 2 is comparable to that observed by McDaniel and colleagues (2009, Experiment 2), who compared a read-recite-review (3R) condition (i.e., free-recall-plus-feedback) to note-taking and to rereading. On the final free recall tests administered immediately and after one week, the 3R condition outperformed the other two conditions. These results and those of our Experiment 2 are in line with the general finding that feedback bolsters the positive influence of retrieval practice (e.g., Rowland, 2014), even after a short retention interval (e.g., Bishara & Jacoby, 2008; Carpenter et al., 2008; Carrier & Pashler, 1992; Jacoby et al., 2010; Kang, 2010; Kornell et al., 2009; Wartenweiler, 2011).

Note that an exploratory analysis showed that participants found it more interesting to read the text for the second time after a free recall phase than to read the text for the third time after a reread phase. Although this might not seem surprising, it is interesting that at the end of the study phase, participants in the retrieval practice condition

showed more interest in their study activity than participants in the reread condition. This motivational aspect might add to the beneficial effect of retrieval practice.

General discussion

A recent study by Van Eersel and colleagues (submitted) asked participants to read expository texts and afterwards to engage in either constructive recall, verbatim free recall, or reread. On a final inference test one week later, no differences between the three study conditions emerged. The present study aimed to investigate why these two free recall conditions did not outperform the reread condition on the final inference test. The results of studies using this kind of final test (Eglington & Kang, in press; Hinze et al., 2013; Tran, Rohrer and Pashler, 2014; Van Eersel et al., submitted) show mixed results, suggesting that retrieval practice effects might be weaker with final inference tests. By contrast, testing effects seem to be reliably observed when pure memory final tests are employed. Therefore, we conducted a first experiment in which a free recall condition was compared to a reread condition on a final free recall test instead of an inference test. If we observed the standard testing effect pattern in this experiment, then the earlier results might have been due to the nature of the final test in this earlier study. Alternatively, a failure to find a typical testing effect pattern might suggest that a specific feature of the text stimuli, perhaps due to a relatively high level of complexity, may prevent a testing effect from occurring. We observed a small benefit of reread over free recall after five minutes, but after one week no difference was left between the two conditions. Moreover, the interaction effect was too small to be statistically significant. These results suggest that the outcomes as observed by Van Eersel and colleagues may not exclusively be explained by the nature of the final test. We then performed the same experiment, but with feedback after the free recall phase, as feedback enhances the positive effect of retrieval practice (e.g., Rowland, 2014). In this second experiment, only a small benefit of free-recall-plus-feedback emerged relative to reread, and no interaction effect between study method and retention interval occurred.

What can we conclude on the basis of these outcomes? Firstly, it is clear that the observed effects in the present study were small. This might be partly due to the low initial test scores in both Experiment 1 (44.45%) and Experiment 2 (43.40%). According to the bifurcation framework (Halamish & Bjork, 2011; Kornell et al., 2011), a test bifurcates the distribution of items' memory strength: memory traces of non-retrieved items remain low in strength while the memory traces of retrieved items become high in strength, resulting in a gap between the two sets of items. Furthermore, items that are restudied are strengthened more in memory than non-retrieved items (but less than retrieved items). Because strong memories last, testing will result in better performance

than restudying after an interval that is long enough for only the strongest memories (i.e., the memory representations of items that were retrieved during testing) to survive. Together this implies that when a small number of items is retrieved in the initial test, the benefit of testing will be limited. This theory might explain the results of our Experiment 1 in which the initial test scores were low, and where only a small advantage of reread over free recall emerged after five minutes. Moreover, providing feedback after testing does also enhance the memory strength of non-retrieved items, thereby preventing bifurcation to occur. In that case, testing with feedback strengthens the memory traces of all tested items, giving rise to a benefit of testing after both a short-term and a long-term interval. The results of Experiment 2 show such a pattern, although the effect sizes were small.

Furthermore, a number of studies (e.g., Carpenter & DeLosh, 2006; Karpicke & Roediger, 2007; Pyc & Rawson, 2009) suggest that the harder it is to retrieve an item initially, the better it will be remembered later. Accordingly, when retrieval effort during initial testing is low and no feedback is provided, the beneficial effect of retrieval practice will be limited. Together the bifurcation framework and the latter retrieval effort hypothesis entail that in order for a testing effect to occur, a sufficiently large number of items needs to be retrieved under circumstances that require a sufficient amount of effort. In other words, finding a cross-over interaction requires an optimal combination of retrieval success and retrieval effort, which are both hard to estimate a priori for a certain set of text material. This makes it difficult to deduce proper predictions when investigating free recall retrieval practice without feedback, because *any* outcome of such a study can be explained by a combination of the two theories. Furthermore, from an educational perspective, retrieval practice without feedback might not be very useful either. However, if retrieval practice is followed by feedback, then the memory traces of both non-retrieved and retrieved items are strengthened, and a testing effect emerges accordingly.

In order to understand the additional effect of feedback, it is important to make a distinction between direct and indirect effects of retrieval practice (Karpicke & Grimaldi, 2012; Roediger & Karpicke, 2006a). An *indirect* effect of retrieval practice means that the influence of retrieval practice is mediated through another factor, like motivation or feedback. The exploratory analyses of Experiment 2 indeed showed that the interest in rereading the text was higher in the free-recall-plus-feedback condition than in the reread condition. Moreover, providing feedback by having someone restudy the text can improve future study, for example because it enables students to correct errors, maintain correct responses, and improve metacognitive monitoring (e.g., Butler, Karpicke, & Roediger, 2008). Such feedback will make subsequent study more effective. A *direct* effect of retrieval practice entails that the retrieved knowledge itself is altered, thereby accommodating retrieval at a later point in time (Karpicke & Grimaldi, 2012).

Feedback may have also influenced performance in a direct way, by strengthening the memory traces of the practiced words through restudy.

However, it should be noted that even in Experiment 2, where free recall retrieval practice was followed by feedback, the retrieval practice advantage over restudying on the final test was small. It is possible that this is related to the degree of coherence of the texts that we used. Previous studies (see Congleton & Rajaram, 2012) showed that compared to restudy, free recall retrieval practice helps participants to structure/organize incoherent study materials. This, in turn, will produce an advantage on a final memory test because the structure formed during previous retrieval practice serves as a strong retrieval cue (Karpicke & Zaromb, 2010; Zaromb & Roediger, 2010). However, if the learning material is already coherent, like in expository text, the most effective retrieval cues are already available, namely the (semantic) associations between the individual elements. Hence, with coherent material, the function of free recall as organizer of the material becomes at least partly redundant, and the advantage of free recall over restudy becomes smaller or disappears (e.g., Bouwmeester & Verkoeijen, 2011; De Jonge et al., 2015; Van Gog & Sweller, 2015). In a critical response, however, both Karpicke and Aue (2015) and Rawson (2015) argued that several studies have shown the testing effect to arise frequently with complex materials as well. Moreover, one would expect that high text coherence would result in a high general recall performance, but the present study shows relatively *low* initial and final recall. Possibly, because of the relatively high coherence of the text, participants only remembered its gist, which forms only a small part of the whole text. Still, the present study suggests that the positive effect of free recall in comparison with restudy is small when expository text is concerned.

In conclusion, in the earlier study by Van Eersel and colleagues (submitted) with educationally relevant texts, no difference in performance on a final inference test emerged between reread, constructive recall without feedback, and verbatim free recall without feedback. Possibly, these findings were attributable to the nature of the final inference test. Indeed, the findings of the present study suggest that retrieval-based learning from text might be more useful when pure retention rather than inference learning is assessed, but the effects were small and occurred only when feedback was provided. This indicates that although the retrieval practice effect is robust, it might not generalize to all types of final tests, materials, and modalities (i.e., typing instead of writing). Future research could therefore further explore the boundaries of the testing effect in terms of stimuli and types of learning.

Acknowledgements

We thank Milou Wierenga, Thei Bongers and Jasmijn Klapwijk for their assistance in data collection and scoring the recall protocols.

6

The testing effect and far transfer: The role of exposure to key information

This chapter has been published as:
Van Eersel, G. G., Verkoeijen, P. P. J. L., Povilenaite, M., & Rikers, R. M. J. P. (2016).
The testing effect and far transfer: The role of focused exposure to key information.
Frontiers in Psychology, 7, 1977. doi: https://doi.org/10.3389/fpsyg.2016.01977

Abstract

Butler (2010: Experiment 3) showed that retrieval practice enhanced transfer to a new knowledge domain compared to rereading. The first experiment of the present study was a direct replication of Butler's third experiment. Participants studied text passages and then either reread them three times or went through three cycles of cued recall questions (i.e., retrieval practice) with feedback. As in Butler's experiment (2010), an advantage of retrieval practice on the final far transfer test emerged after one week. Additionally, we observed an advantage of retrieval practice on the final test administered after five minutes. However, these advantages might have been due to participants in the retrieval practice condition receiving focused exposure to the key information (i.e., the feedback) that was needed to answer the final test questions. We therefore conducted a second experiment in which we included the retrieval practice condition and the reread condition from our first experiment, as well as a new reread-plus-statements condition. In the reread-plus-statements condition, participants received focused exposure to the key information after they had reread a text. As in Experiment 1, we found a large effect on far transfer when retrieval practice was compared to rereading. However, this effect was substantially reduced when retrieval practice was compared to the reread-plus-statements condition. Taken together, the results of the present experiments demonstrate that Butler's (2010) testing effect in far transfer is robust. Moreover, focused exposure to key information appears to be a significant factor in this far transfer testing effect.

Retrieving information from memory after an initial learning phase enhances long-term retention more than restudying the material; an advantage referred to as the (retrieval practice) testing effect (for reviews, see Carpenter, 2012; Delaney, Verkoeijen, & Spirgel, 2010; Karpicke, 2012; Karpicke, Lehman, & Aue, 2014; Roediger & Butler, 2011; Rowland, 2014). The testing effect has been demonstrated with a variety of practice tests, materials, and age groups (Dunlosky et al., 2013). In most of the testing effect research, the materials in the intermediate test and the final test are identical (e.g., Roediger, Agarwal, McDaniel, & McDermott, 2011; Roediger & Karpicke, 2006b, Wooldridge, Bugg, McDaniel, & Liu, 2014). However, it is important to determine whether the testing effect still emerges when the final test is different from the intermediate test and measures related but new knowledge, i.e., whether *transfer* of knowledge takes place (McDaniel, Howard, & Einstein, 2009; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; Salomon & Perkins, 1989).

A relatively small but increasing number of studies has shown that retrieval practice benefits the transfer of knowledge (for a recent review, see Carpenter, 2012). Transfer may be broadly defined as the ability to apply previously learned knowledge or skills in a novel context (Carpenter, 2012). To assess transfer with respect to the testing effect, Carpenter (2012) makes a distinction between three dimensions along which the differences between learning and transfer contexts can be compared: temporal context, test format, and knowledge domain. Following Barnett and Ceci (2002), a transfer task can be evaluated on each of these dimensions in terms of the level of transfer (i.e., *near* vs. *far* transfer).

Many studies have found that the beneficial effect of retrieval practice transfers across temporal contexts, as the effect of retrieval practice usually occurs after a retention interval of some days (e.g., Carpenter, Pashler, & Cepeda, 2009; Chan, McDermott, & Roediger, 2006; Johnson & Mayer, 2009). However, in nearly all retrieval practice studies, the test formats and knowledge domains are comparable between the learning phase and the test phase (e.g., Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Carpenter & Delosh, 2006 (Exp. 1); Carpenter, Pashler, & Vul, 2006; Coppens, Verkoeijen, & Rikers, 2011). Only a small number of studies has shown transfer of a retrieval practice effect across different test formats (e.g., Carpenter, 2011; Halamish & Bjork, 2011; Karpicke & Zaromb, 2010; Rowland & Delosh, 2015; Sensenig, Littrell-Baez, & Delosh, 2011), and across both different test formats and temporal contexts (e.g., Blunt & Karpicke, 2014; Hinze & Wiley, 2011; Kang, McDermott, & Roediger, 2007; Lyle & Crawford, 2011; McDermott, Agarwal, D'Antonio, Roediger, & McDaniel, 2014; Rawson, Vaughn, & Carpenter, 2015; Rohrer, Taylor, & Sholar, 2010). Yet in all of the previous studies – and in fact in the vast majority of retrieval practice studies – the knowledge domain is the same in the learning phase and the final test. To the best of

our knowledge, there is only one study (Butler, 2010) in which the retrieval practice effect emerged on a final transfer test tapping onto a different knowledge domain. Specifically, Butler (2010: Experiment 3) found a positive effect of retrieval practice on a final test that consisted of questions pertaining to topics from a different knowledge domain than the intermediate test questions. Note that although in Carpenter's review (2012) several papers are mentioned in the section called "Transfer across knowledge domains", only the paper by Butler (2010) actually meets the condition stated in the title. In Butler's (2010) crucial third experiment, participants first read six prose texts, and then reread three of the passages and practiced cued recall on the other three passages. In the retrieval practice condition participants answered conceptual cued recall questions, and afterwards they received feedback in the form of the correct response. One week later, participants took the final transfer test. This test consisted of questions from different knowledge domains than the questions presented during the practice phase. On the final test, participants performed better after cued recall than after rereading. Because Butler's (2010) third experiment has been the only one to demonstrate that retrieval practice fosters transfer to a different knowledge domain, it is important to investigate whether Butler's results are robust. In addition, Butler's sample size was small (twenty participants), with imprecise parameter estimations as a result. Now given the importance of replication in (psychological) science (e.g., Cartwright, 1991; Ioannidis, 2005; Klein et al., 2014; Open Science Collaboration, 2012; Pashler & Wagenmakers, 2012), the purpose of our first experiment was conduct an exact replication (see Schmidt, 2009) of Butler's third experiment.

Furthermore, the observed beneficial effect of retrieval practice may have been partly due to another factor, namely the focused exposure to key information (i.e., the feedback) in the retrieval practice condition versus the reread condition. To illustrate this point, consider this example of a cued recall question from the retrieval practice condition: "Some bats use echolocation to navigate the environment and locate prey. How does echolocation help bats to determine the distance and size of objects?" The answer, which was taken from the text and presented as feedback, was the following: "Bats emit high-pitched sound waves and listen to the echoes. The distance of an object is determined by the time it takes for the echo to return. The size of the object is calculated by the intensity of the echo: a smaller object will reflect less of the sound wave, and thus produce a less intense echo." The related transfer question was the following: "Submarines use SONAR to navigate underwater much like bats use echolocation to navigate at night. Using SONAR, how does a submarine determine that an object is moving towards it (i.e. rather than away from it)?" (Answer: The submarine can tell the direction that an object is moving by calculating whether the time it takes for the sound waves to return changes over time. If the object is moving towards the submarine, the

time it takes the sound wave to return will get steadily shorter. Also, the intensity of the sound wave will increase because the object will reflect more of the sound wave as it gets closer.").

This example demonstrates that the retrieval practice questions and the final tests questions are conceptually related; the same principles that are learned during retrieval practice needed to be applied to the final test questions in the different knowledge domains. This means that participants in the retrieval practice condition may have had an advantage compared to the reread condition because they had already seen the relevant principles in the form of the key information that was provided as feedback during retrieval practice. Although participants in the reread condition had also seen these principles when they reread the whole text, they had only seen them as a part of the full text that contained additional information, not as answers to specific questions. Hence, in the retrieval practice condition, participants only needed to retrieve the key information from the feedback in the learning phase and apply it to the final test questions. By contrast, in the reread condition participants had to retrieve and select the part of the text relevant to the problem, and apply it to the final test items. In this condition, it might have been difficult to determine which part of the text was relevant for a final test question. As a result, final test performance in the reread condition might have suffered compared to the retrieval practice condition.

We therefore carried out another experiment that was identical to our first experiment, but with an extra reread-plus-statements condition, besides the reread and the retrieval practice conditions. In the reread-plus-statements condition, participants reread a text, followed by focused exposure to key information. This information consisted of statements that contained the same information as the feedback in the retrieval practice condition. In this way, we tested whether the focused exposure to key information could – partly – account for the testing effect found in Butler (2010).

Experiment 1

The first experiment was a direct replication of Butler's (2010) third experiment, but with an extra five-minutes retention interval. Participants first read six texts and then repeatedly reread three of the texts, and repeatedly took conceptual cued recall tests with immediate feedback on the other three texts. After five minutes or after one week, participants completed a final transfer test.

Method

Participants

This experiment was carried out in accordance with the recommendations of the Ethical Committee of the Department DPECS at the Erasmus University Rotterdam, with written informed consent from all participants. Fifty-six people participated in the study and were rewarded with course credits. Their mean age was 19.71 ($SD = 3.43$). Sixteen of them were males, forty were females. All of the participants were Psychology undergraduates, with eighteen of them having the Dutch nationality, while the others had (twenty-five) other nationalities. The non-Dutch participants were students of the English-taught international bachelor in Psychology and had been selected on the basis of their scores on internationally accepted English language tests. The Dutch participants had been taught English for eight years during primary and secondary education, and can be considered as highly proficient in written English.

Materials and design

We used Butler's (2010) original materials: six prose passages in English about different topics of between 550 and 600 words in length. Each passage included four concepts. For each concept, Butler (2010) created a question to assess transfer to a different knowledge domain (see the introduction for an example). Each transfer question required the application of a concept from the initial learning session. The correct response was between one and three sentences long. The experiment had a 2 Study Method (reread vs. retrieval practice) * 2 Retention Interval (five minutes vs. one week) mixed design with repeated measures on the first factor. Note that in our study, we added an additional five-minutes retention interval to Butler's (2010) original design, in order to observe whether the testing effect would increase over time, which is sometimes regarded as a defining feature of the testing effect (e.g., Delaney, Verkoeijen, & Spirgel, 2010; Kornell, Bjork, & Garcia, 2011; Roediger & Karpicke, 2006b). Participants were randomly assigned to the levels of the between-subjects factor. Like in Butler (2010), we used four counterbalanced versions of the experiment by combining two orders of initial learning condition with two orders of the passages. Also as in Butler's experiment, the dependent variable was the proportion of correct answers to the twenty-four final test transfer questions. Following Simmons, Nelson, and Simonsohn (2011), we have reported all conditions and all measures in this experiment.

Procedure

The procedure was identical to the procedure of Butler's (2010) third experiment. Our study was conducted on a computer using E-Prime software, and Butler had provided

us with the original E-Prime files that he had used for his original study (2010). In the first session, participants read the six English prose texts for two minutes each. Afterwards they repeatedly (i.e., three times) reread three of the passages for two minutes each, and repeatedly (i.e., three times) took identical cued recall tests with immediate feedback (retrieval practice) on the three other passages. In the retrieval practice condition, participants answered four conceptual cued recall questions per text, and received feedback in the form of the correct response after each question. There was no time limit to answer the questions or review the feedback. Participants were encouraged to think a while, and to generate a response to every question (Butler, personal communication, 6 October, 2014). Half of the participants took the final test after five minutes, the other half after one week. The final test was self-paced and consisted of twenty-four transfer questions about different knowledge domains.

Results

Following Kline (2004, Chapter 3) and Cumming (2014), we use the term 'statistical' instead of 'significant' for all statistical analyses, because the latter is often erroneously understood as meaning 'important'. All data from this study can be retrieved from the Open Science Framework[1].

Scoring

Two research assistants and the first author independently scored 27% of the answers to the final test questions. Each answer was scored as either correct or incorrect based on the correct answers provided in Butler's (2010) supplemental material. Cohen's kappa was used as the interrater reliability measure and was .82 for the five-minutes condition and .74 for the one-week condition. These coefficients indicate a substantial level of agreement (Landis & Koch, 1977). The remaining responses were scored by the first author. Note that Cohens's kappa is based on the absolute agreement between raters. Such an agreement is unnecessarily strict when the aim is – like in this experiment – to evaluate the mean difference between groups on a dependent variable, rather than to obtain a reliable estimate of the absolute level of performance within each group. In the former case, it is sufficient that raters are *consistent* regarding their final test total scores, without absolute agreement. We therefore also calculated the Pearson correlation coefficient between the total final test scores given by the two raters, which were $r = .97$, $p < .001$, for the five-minutes condition and $r = .93$, $p = .002$, for the one-week condition.

1 https://osf.io/2jdxp/

Retrieval practice tests

The proportion of correct responses to the initial cued recall tests increased in a curvilinear fashion from Test 1 (M = .39, SD = .20) to Test 2 (M = .77, SD = .18) to Test 3 (M = .84, SD = .14). A repeated measures ANOVA yielded a statistical main effect of Test Session, $F(1.82, 98.10) = 231.20$, MSE = .02, $p < .001$, *Partial* η^2 = .81, for which there was a linear trend, $F(1, 54) = 329.52$, MSE = .02, $p < .001$, *Partial* η^2 = .86, as well as a quadratic trend $F(1, 54) = 75.00$, MSE = .01, $p < .001$, *Partial* η^2 = .58. The mean proportion of correct responses during the learning phase did not differ between the five-minutes condition (M = .69, SD = .13) and the one-week condition (M = .64, SD = .15), $t(53) = 1.16$, p = .25, *Partial* η^2 = .03, 95% CI of the difference [-.03, .12]. Note that due to an error, the retrieval practice data of one participant were not included in these calculations.

Time on task

During retrieval practice, there were three texts with four questions each. All questions were repeated three times, resulting in a total number of thirty-six questions. The distribution of the number of seconds that participants spent on answering a question during retrieval practice was skewed to the right, so we report both the mean and the median. The mean was 76.88 s (SD = 38.55) and the median 70.73 s. The number of seconds that participants spent on reading the feedback was also skewed to the right. The mean was 14.42 s (SD = 8.96) and the median 12.14 s. As a result, the mean number of seconds participants spent on each question (responding and reading feedback) was 91.29 (median = 81.76). Because there were four questions per passage, it took participants on average 365.16 s to complete a test on each passage (median = 327.04 s). In the reread condition, all participants had 120 s to reread a passage. Note that due to an error, the time-on-task data of three participants were not included.

Final tests

A 2 Study Method (reread vs. retrieval practice) * 2 Retention Interval (five minutes vs. one week) Repeated Measures ANOVA on the proportion of correct answers did not yield a statistical interaction effect between study method and retention interval, $F(1, 54) = 2.41$, MSE = .02, p = .126, *Partial* η^2 = .04 (see Table 1). In accordance, the difference between retrieval practice and rereading was large after five minutes, $t(27) = 6.98$, $p < .001$, *Cohen's d* = 1.32, 95% CI of the difference [.18, .33], as well as after one week, $t(27) = 5.11$, $p < .001$, *Cohen's d* = 0.97, 95% CI of the difference [.11, .25]. In sum, after both retention intervals there was large advantage of retrieval practice over rereading. In addition, there was a statistical main effect of study method, $F(1, 54) = 73.66$, MSE = .02, $p < .001$, *Partial* η^2 = .58, 95% CI of the difference [.17, .27], but not of retention interval, $F(1, 54) = 1.50$, MSE = .06, p = .226, *Partial* η^2 = .03, 95% CI of the difference [-.04, .15].

We also performed a Bayesian Repeated Measures ANOVA (Rouder, Morey, Speckman, & Province, 2012; Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2017) in the software program JASP (Love et al., 2015; Wagenmakers et al., 2016) with a default Cauchy prior width of $r = 0.50$ for effect size on the alternative hypothesis (for arguments for the Bayesian approach, see, e.g., Dienes, 2011). The Bayes factor for the interaction effect between study method and retention interval was BF01 = 1.38, showing that the likelihood of the data under the null hypothesis was 1.38 times the likelihood of the data under the alternative hypothesis. Following Wetzels and Wagenmakers (2012), this is anecdotal evidence for the null hypothesis that postulates the absence of the interaction effect. The Bayes factor for the factor study method was larger than 100, indicating that the observed data were more than 100 times more likely under the alternative hypothesis than under the null hypothesis. Bayes Factors larger than 100 are decisive evidence in favour of the alternative hypothesis. In this experiment, this meant decisive evidence for the advantage of retrieval practice over rereading on the final transfer test. The Bayes Factor for the factor retention interval was BF01 = 2.02, which can be interpreted as anecdotal evidence in favour of the null hypothesis of no difference between the two intervals on the proportion of correct answers on the final transfer test.

Table 1. Proportion of Answers Correct in Experiment 1 by Retention Interval and Study Method. Standard Errors are between Brackets

	Retention Interval	
Study Method	**Five minutes**	**One week**
Retrieval practice	.57 (0.04)	.59 (0.04)
Reread	.31 (0.03)	.41 (0.03)

The time-on-task differed considerably between the two conditions in our experiment, which might have confounded the final test results. To assess whether this was the case, we calculated the time-on-task difference between the retrieval practice condition and the reread condition for each participant. Because time-on-task in the reread condition was a constant, the variance of these time-on-task difference scores amounted to the variance of the time-on-task scores in the retrieval practice condition. Furthermore, for each participant we calculated a final test difference score by subtracting the transfer score in the reread condition from that in the retrieval practice condition. Subsequently, we correlated these two difference scores and found no trace of a statistical correlation between the time-on-task difference scores and the difference scores on the transfer test, $r = .04$, $p = .765$. This low and non-statistical correlation indicates that an increased time-on-task in the retrieval practice condition was not associated with a larger advantage of retrieval practice over reread, suggesting that the retrieval practice effect was not confounded by time-on-task differences between conditions.

Discussion

Experiment 1 showed a large benefit of retrieval practice compared to rereading on the final far transfer test administered after one week, that is, a far transfer testing effect. Furthermore, the effect size associated with this testing effect was large (*Cohen's d* = 0.97), which is comparable to Butler's (2010) *Cohen's d* of 1.17[2]. Hence, the results of Experiment 1 convincingly replicated the results of Butler's (2010) third experiment. As such, our findings provide a crucial independent reinforcement of this important finding within the testing effect literature.

Note that the overall final test performance in Experiment 1 did not differ between the five-minutes condition and the one-week condition. The length and the nature of the experimental procedure might explain this result. It took on average almost 1,5 hours (85 minutes) to complete the first session. At the end of the experiment, the experiment leader always asked how everything went, and whether the participant had any comments or questions. More than 50% of the participants reported that the experiment was tiring or boring. Hence, participants in the five-minutes condition were probably not as motivated and concentrated to start the final test session as participants in the one-week condition. This, in turn, might have resulted in comparable performance for both retention intervals. Although there was no difference in the number of skipped final test questions between the one-week condition and the five-minutes condition, participants in the one-week condition ($M = 32.84$, $SD = 11.56$) took almost seven minutes longer to complete the final test than participants in the five-minutes condition ($M = 25.94$, $SD = 8.43$), $t(50) = -2.46$, $p = .018$, $r = .33$, again indicating that the latter might have been less motivated than the former.

The novel result of Experiment 1 was that the benefit of retrieval practice already emerged after a retention interval of five minutes. Although there is agreement in the literature that the testing effect usually arises after a long retention interval (i.e., longer than one day), mixed support has been found for short intervals (Rowland, 2014). A factor that can partly explain these mixed findings is feedback (Rowland, 2014). In several studies where feedback was provided after retrieval practice, testing effects did emerge after a short retention interval (e.g., Bishara & Jacoby, 2008; Carpenter, Pashler, Wixted, & Vul, 2008; Carrier & Pashler, 1992; Jacoby, Wahlheim, & Coane, 2010; Kang, 2010; Kornell, Hays, & Bjork, 2009; Wartenweiler, 2011). Now, in our first experiment, participants received feedback after retrieval practice in the form of the correct response. This might have been the factor underlying the short-term testing effect.

The feedback in Experiment 1 may also have prevented the interaction effect between study method and retention interval to occur, which is in line with the predictions of

2 Due to a miscalculation of the standard deviation, the reported *Cohen's d* in Butler's article (2010) was 0.99, but in reality it was 1.17 (personal correspondence, July 14, 2016).

the bifurcation framework (Kornell, Bjork, & Garcia, 2011). According to this framework, a test bifurcates the distribution of items' memory strength: non-retrieved items remain low in strength while retrieved items become high, resulting in a gap between the two sets of items. Furthermore, items that are retrieved during testing are strengthened more than items that are restudied. Because strong memories last, testing will result in better performance than restudying after a long interval. Feedback, however, also boosts the memory strength of non-retrieved items, thereby preventing bifurcation to occur. In that case, testing with feedback will strengthen all tested items, giving rise to a benefit of testing after both a short-term and a long-term interval.

Experiment 2

To answer the final test questions of Experiment 1 correctly, participants in the retrieval practice condition had to apply the key information that had been provided as feedback. This may have granted an advantage to participants in the retrieval practice condition compared to the reread condition. Hence, our results, as well as those of Butler's (2010) third experiment, might have been partly driven by this focused exposure to key information rather than by retrieval practice per se. To test this alternative account of our results, we conducted a second experiment. This experiment was similar to the first, but now with an extra reread-plus-statements condition, besides the retrieval practice condition and the reread condition. In the retrieval practice condition, participants answered four cued recall questions per text and received feedback after each question. Participants in the reread-plus-statements condition first reread a text for two minutes, and then read four isolated statements that contained the same information as the feedback in the retrieval practice condition (cf. Butler, 2010, Experiment 2 on near transfer). In the reread condition, participants reread a text for two minutes. So, in all three conditions, participants received the same key information that was necessary to answer the final test questions. However, in the reread condition, the key information was presented together with the additional information that was in the text. By contrast, in both the retrieval practice condition and the reread-plus-statements condition, participants received focused exposure to the key information. We therefore expected the difference between the retrieval practice condition and the reread-plus-statements condition on the final transfer test to be considerably smaller (or perhaps even absent) than the difference between retrieval practice and rereading.

Note that we decided to drop the five-minutes condition in the second experiment, because the short-term final test results in the first experiment might have suffered from participants' fatigue. A convenient side effect of this choice was that we had more participants – and hence more power – to detect an effect of our experimental manipulation on transfer after a retention interval of one week.

Method

Participants

This experiment was carried out in accordance with the recommendations of the Ethical Committee of the Department DPECS at the Erasmus University Rotterdam, with written informed consent from all participants. Fifty-five people participated in this study and were rewarded with course credits. One participant was removed because she said she had only paid attention to the texts with the questions (retrieval practice condition), leaving a total number of fifty-four participants. Their mean age was 20.00 (*SD* = 4.10). Twenty of them were males, thirty-four were females. All of the participants were Psychology undergraduates (see Experiment 1). Twenty-two of them had the Dutch nationality, while the others had (fifteen) other nationalities.

Materials and design

Butler's (2010) six prose texts and test questions were used again, see Experiment 1. The experiment had a 3 Study Method (reread vs. retrieval practice vs. reread-plus-statements) within-subjects design. We used a Latin Square to create nine counterbalance conditions, using three sets of two texts and six orders of initial study conditions. Following Simmons, Nelson, and Simonsohn (2011), we have reported all conditions and all measures in this experiment.

Procedure

The experiment was conducted on a computer using E-Prime software. As in Experiment 1 we used Butler's (2010) original files, but now with some adjustments to include the extra reread-plus-statements condition. The procedure was identical to that in Experiment 1, except for the elimination of the five-minutes retention interval and the new reread-plus-statements condition (and, as a result, two instead of three texts per study method). In the reread condition, two texts were reread three times. In the retrieval practice condition, participants took three identical four-item cued recall tests with immediate feedback on two other texts. This feedback was identical to that in Butler (2010) and in our Experiment 1. In the reread-plus-statements condition, participants repeatedly (i.e., three times) reread two of the texts, each followed by four statements that contained the same information as presented as feedback in the retrieval practice condition, except that it was rephrased in order to make sense as prose (cf. the isolated sentences in Butler, 2010, Experiment 2). For example, one of the questions in the retrieval practice condition was the following: "A bat has a very different wing structure from a bird. What is the wing structure of a bat like relative to that of a bird?" The answer to this question was as follows: "A bird's wing has fairly rigid bone structure that is efficient at providing lift, whereas a bat has a much more

flexible wing structure that allows for greater manoeuvrability." In the reread-plus-statements condition, the corresponding key statement was the following: "A bat has a very different wing structure from a bird. A bird's wing has fairly rigid bone structure that is efficient at providing lift, whereas a bat has a much more flexible wing structure that allows for greater manoeuvrability." Hence, the key statement contained part of the question in order to be comprehensible. The instruction for the key statements was as follows: "Next, you will see four short pieces of information about the subject of the text that you have just read. Please read them carefully." No time limit was given to read the key statements. One week later, participants returned to take the self-paced final test that consisted of twenty-four transfer questions about different knowledge domains.

Results

Scoring

One research assistant and the first author independently scored 15% of the cued recall questions from the initial learning session. Each answer was scored as either correct or incorrect. Cohen's kappa was used as the interrater reliability measure and was .67. This indicates a substantial level of agreement (Landis & Koch, 1977). The Pearson correlation coefficient between the total scores given by the two raters was $r = .79$, $p = .021$. All the remaining questions were scored by the first author.

Ten people had coincidentally pushed Enter when they wanted to answer the first question of the final test, thereby going directly to the second question on the next screen. These ten questions were treated as 'missing', and their values were estimated by taking the average of the scores on the other three questions corresponding to the same text.

Retrieval practice tests

The proportion of correct responses to the initial cued recall tests increased in a curvilinear fashion from Test 1 ($M = .48$, $SD = .23$) to Test 2 ($M = .79$, $SD = .19$) to Test 3 ($M = .90$, $SD = .13$). A repeated measures ANOVA yielded a statistical effect of Test, $F(1.82, 94.79) = 128.23$, $MSE = .02$, $p < .001$, *Partial* $\eta^2 = .71$, for which there was a linear trend, $F(1, 52) = 176.77$, $MSE = .03$, $p < .001$, *Partial* $\eta^2 = 77$, as well as a quadratic trend $F(1, 52) = 31.16$, $MSE = .01$, $p < .001$, *Partial* $\eta^2 = .38$. Note that due to an error, the retrieval practice and time-on-task data of one participant were not included.

Time on task

During retrieval practice, there were two texts with four questions each. The questions were repeated three times, resulting in a total number of twenty-four questions. The distributions of number of seconds spent reading and answering questions were all

skewed to the right, so we report both the mean and the median. The average number of seconds that participants spent on answering a question during cued recall was 63.40 s (*SD* = 26.26), median 57.86 s. The mean number of seconds that participants spent on reading the feedback was 12.24 (*SD* = 5.61), median 10.90. As a result, the mean number of seconds that participants spent on each question (responding and reading feedback) was 75.64 (median = 68.76). Because there were four questions per passage, it took participants 302.56 s on average to complete a test on each passage (median = 275.04 s). Furthermore, in the reread-plus-statements condition, there were two texts that were both followed by four statements. Participants read the texts and the statements three times, coming down to twenty-four statements in total. The mean number of seconds that participants spent on reading a statement was 20.44 (*SD* = 11.59), median 16.30, and they spent 120 s on rereading a passage. Because there were four statements per passage, this resulted in a total time per passage of 201.76 s (median 185.20 s). In the reread condition, all participants had 120 s to reread a passage.

Final tests

A 3 Study Method (reread vs. retrieval practice vs. reread-plus-statements) Repeated Measures ANOVA on the proportion of correct answers on the final transfer test revealed a statistical main effect of study method, $F(2, 106) = 22.62$, $MSE = .04$, $p < .001$, *Partial* $\eta^2 = .30$ (see Table 2). Bonferroni corrected pairwise comparisons revealed statistical differences between all three conditions. Retrieval practice differed from reread-plus-statements, $t(53) = 3.08$, $p = .009$, $r = .39$, 95% CI of the difference [.02, .20]. In addition, there was a difference between retrieval practice and rereading, $t(53) = 6.26$, $p < .001$, $r = .65$, 95% CI of the difference [.15, .34]. Also, reread-plus-statements differed from rereading, $t(53) = 3.80$, $p = .001$, $r = .46$, 95% CI of the difference [.05, .22].

Table 2. Proportion of Answers Correct in Experiment 2 by Study Method. Standard Errors are between Brackets

Study Method	
Retrieval practice	.64 (0.03)
Reread-plus-statements	.53 (0.03)
Reread	.40 (0.03)

In addition, we performed a Bayesian Repeated Measures ANOVA (Rouder, Morey, Speckman, & Province, 2012; Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2017) in the software program JASP (Love et al., 2015; Wagenmakers et al., 2016) with a default Cauchy prior width of $r = 0.50$ for effect size on the alternative hypothesis. The Bayes Factor for the factor study method (reread vs. retrieval practice vs. reread-plus-statements) was larger than 100, indicating that the observed data were more than

100 times more likely under the alternative hypothesis than under the null hypothesis. According to Wetzels and Wagenmakers (2012), this is decisive evidence in favour of the alternative hypothesis that postulates a difference in means between the three conditions (reread vs. retrieval practice vs. reread-plus-statements). As follow-up tests, we performed three new two-sided Bayesian paired samples T-Tests (Rouder, Speckman, Sun, Morey, & Iverson, 2009) with a default Cauchy prior width of $r = 0.71$ for effect size on the alternative hypothesis (which is equivalent to the prior width of $r = 0.50$ used for the Bayesian Repeated Measures ANOVA, see Wagenmakers et al., 2016). The comparison between retrieval practice and rereading delivered a Bayes factor of larger than 100, which is decisive evidence for the hypothesis that there is a difference between retrieval practice and rereading. In this experiment, there was a large benefit of retrieval practice compared to rereading. The comparison between reread-plus-statements and reread produced a Bayes factor of BF10 = 86.71, showing that the likelihood of the data under the alternative hypothesis was 86.71 times the likelihood of the data under the null hypothesis. This presents very strong evidence for the alternative hypothesis that there is a difference, in this case a large advantage of reread-plus-statements compared to reread. Finally, retrieval practice was compared to reread-plus-statements. This yielded a Bayes factor of BF10 = 10.77, meaning that the observed data were 10.77 times more likely under the alternative hypothesis than under the null hypothesis. This is strong evidence for the alternative hypothesis of a difference, in this case an advantage of retrieval practice. To inspect the robustness of the latter analysis, the Bayes factor is plotted as a function of the scale parameter *r* of the Cauchy prior in Figure 1. As the scale parameter *r* increases (i.e., the prior becomes wider), the evidence for the alternative hypothesis gets weaker. However, even under the prior settings that least favour the alternative hypothesis, the Bayes factor is still larger than 6.50, indicating substantial evidence for the alternative hypothesis (Wetzels & Wagenmakers, 2012).

Again, the time-on-task differed considerably between the three conditions in Experiment 2. To assess whether this variable confounded the final test results, we calculated for each participant the time-on-task differences for each of the three unique condition combinations (with time-on-task in the reread condition being a constant). Furthermore, for each participant we calculated the final test difference scores for each of the three condition combinations. Subsequently we correlated these time-on-task difference scores with the relevant final test difference scores. For 'retrieval practice / reread', there was no statistical correlation between time-on-task differences and final test differences, $r = .09$, $p = .525$. The same applied to 'reread-plus-statements / reread', $r = .01$, $p = .920$, and 'retrieval practice / reread-plus-statements, $r = .08$, $p = .569$. These non-statistical correlations indicate that the final test results were not confounded by time-on-task differences.

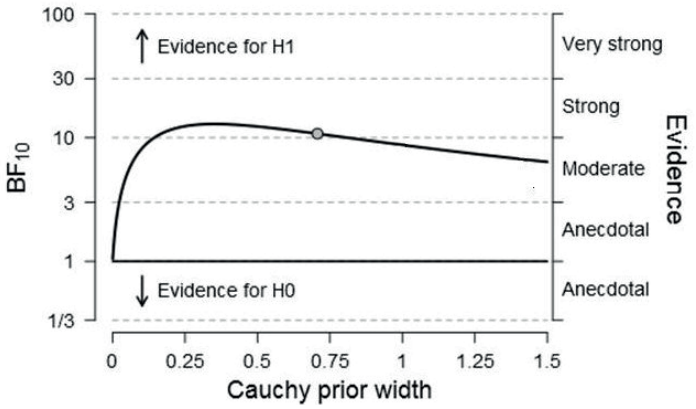


Figure 1. Bayes Factor for the comparison between retrieval practice and reread-plus-statements (Experiment 2) as a function of the scale parameter *r* of the Cauchy prior for effect size under the alternative hypothesis. The dot indicates the used prior width of $r = 0.71$. Figure adjusted from JASP, jasp-stats.org.

Discussion

The results of Experiment 2 showed that reread-plus-statements resulted in a higher score on the final transfer test than rereading. In addition, retrieval practice led to better performance on the final test than reread and better than reread-plus-statements. When retrieval practice was compared to reread-plus-statements, however, the effect size was much smaller than when it was compared to reread (resp., $r = .39$ versus $r = .65$); the proportion of explained variance fell by about 65% when retrieval practice was contrasted with reread-plus-statements (testing effect magnitude of $r^2 = .15$) instead of reread (testing effect magnitude of $r^2 = .42$). Taken together, the results of Experiment 2 suggest that the advantage of retrieval practice, found in Butler (2010) and in our first experiment, was partly due to the focused exposure to key information (i.e., the feedback). However, the advantage of retrieval practice over the reread-plus-statements condition indicates that practicing retrieval added something extra, above and beyond providing participants with focused exposure to key information.

General discussion

In Experiment 1, we replicated the third experiment of Butler (2010). Retrieval practice produced better performance than rereading on the final transfer test administered after one week, and also after five minutes. Experiment 2 was similar to Experiment 1, but with an extra reread-plus-statements condition. In this condition, the retrieval practice questions were replaced by rereading, followed by focused exposure to the key

information. These key statements contained the same information as the feedback that participants received in the retrieval practice condition. In this manner, we examined whether the focused exposure to key information could explain the large advantage of retrieval practice over reread in Experiment 1.

In Experiment 2, retrieval practice again outperformed rereading on a delayed far transfer test. Moreover, transfer performance in the reread-plus-statements condition was considerably better than in the reread condition. In addition, retrieval practice resulted in a higher final test score than reread-plus-statements. However, the testing effect was much smaller when retrieval practice was compared to reread-plus-statements (r = .39, testing effect magnitude of r^2 = .15, BF = 10.77) than when retrieval practice was compared to reread (r =.65, testing effect magnitude of r^2 =.42, BF > 100). Note that it is important to focus not only on the *p*-value but also on the size of the effect. Based on the *p*-values, one would simply conclude that retrieval practice leads to better far transfer than reread and reread-plus-statements. However, when the effect sizes are taken into account, a different picture emerges. That is, it becomes clear that the effect of retrieval practice might be partly attributed to the focused exposure to key information (i.e., the feedback). These findings are important from a theoretical as well as from a practical perspective. Both for theory development and for real-world applications (such as in educational practice), it is crucial to realize that the benefit of retrieval practice varies with the control condition to which it is compared.

Because time-on-task differed between conditions in both our experiments, we wanted to exclude the possibility that this variable was a confounder. We therefore inspected the correlations between time-on-task differences between conditions and final test differences. These correlations were all very small and statistically non-significant, indicating that there was no association between time-on-task and final test advantages of retrieval practice and reread-plus-statements over rereading. This seems in line with Butler's (2010) study. In his second experiment, the time-on-task spent per text on the conceptual questions in the retrieval practice condition (168.4 s) was considerably *lower* than the time-on-task in the reread condition (240 s). Conversely, in his third experiment, the time-on-task in the reread condition was 120 s per text, and although the time-on-task in the retrieval practice condition was not reported, it is reasonable to assume that this was comparable to the time-on-task in the second experiment (168.4 s). In that case, the time-on-task in the retrieval practice condition in the third experiment was *higher* than the time-on-task in the reread condition. Still, in both experiments retrieval practice outperformed rereading on the final test. Together these findings indicate that time-on-task did not matter much for performance on the final transfer test that Butler and we have used. This is in keeping with other studies where increased time-on-task was not related to retention performance (e.g., Amlund,

Kardash, & Kulhavy, 1986; Callender & McDaniel, 2009). Hence, we think that the time-on-task differences between conditions did not confound the final test results.

It might be argued that the benefit of retrieval practice in Experiment 2 was reduced compared to the reread-plus-statements condition because participants received less exposure to the key information in the retrieval practice condition than in the reread-plus-statements condition. In the reread-plus-statements condition, participants were exposed to the key information twice: first when rereading the text and second when reading the isolated statements. By contrast, one could assert that participants in the retrieval practice condition only were exposed to the key information twice when the material was successfully retrieved. However, some research (e.g., Kornell, Hays, & Bjork, 2009; Richland, Kornell, & Kao, 2009) suggests that retrieval attempts promote learning even when the attempts are unsuccessful. Moreover, in the vast majority of the testing effect studies, exposure to the to-be-learned information is lower after retrieval practice than after restudying because retrieval practice is not perfect. Nevertheless, large testing effects are observed in these studies on delayed final tests (e.g., Roediger & Karpicke, 2006b).

In addition, one could argue that the focused exposure to key information reduced the beneficial effect of retrieval practice compared to reread-plus-statements because the exposure to key information was more spaced (i.e., distributed over time) in the latter than in the former. This is because in the retrieval practice condition, the key information was provided as feedback immediately following the questions to which participants had to respond (massed repetition), whereas participants in the reread-plus-statements condition received the information after rereading the text (more spaced repetition). As a consequence, transfer performance in the reread-plus-statements condition might have benefitted from a spacing effect (see, e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006) and this – rather than focused exposure to key information per se – might have resulted in a smaller retrieval practice advantage.

Although the above line of reasoning is correct, it is only directed at repetitions *within* a relearning session, while it does not consider repetitions *across* the three relearning sessions. However, Karpicke and Bauerschmidt (2011) showed that when comparing repetition schedules on final test performance, it is pivotal to determine the absolute or total spacing per schedule. Consequently, in the present study, it is not appropriate to only focus on repetition *within* a relearning session (or only on repetitions *across* sessions, for that matter); instead, one should compare conditions on total spacing. This total spacing is obtained by combining spacing lags within and across learning sessions. In the present study, participants went through three relearning sessions in all conditions. Within a relearning session, repetition of key information was relatively massed in the retrieval practice condition and relatively spaced in the reread-plus-

statements condition. However, the total spacing of repetitions (i.e., the combination of spacing within and across sessions) reveals a somewhat different picture. Consider a participant in the retrieval practice condition who correctly answers the first question in all three relearning sessions. This participant will be exposed to the key information six times (i.e., answer and feedback in each of the three sessions), leading to five spacing intervals. Within a session, feedback is presented immediately after answering a question, resulting in a spacing interval of 0 seconds. However, given the time-on-task of one relearning session in the retrieval practice condition (i.e., answering the other three questions and reading the feedback), the next repetition in the second session appears after 226.92 s (¾ * 302.56). That is, the spacing from the first to the second session is 226.92 s. The same applies for the spacing between the second and the third session. Hence, in the retrieval practice condition, total spacing was 453.84 s. By contrast, in the reread-plus-statements condition, the total spacing between the six repetitions of the key information was about 392.64 s (2 * (¾ * 201.76) + ¾ * 120).

So, in both the retrieval practice condition and the reread-plus-statements condition the exposure to the key information was spaced, but the conditions differ in total spacing. However, because the function between total spacing and memory performance reaches approximately an asymptote at a total spacing considerably shorter than those in our Experiment 2 (cf, Glenberg, 1976; Raaijmakers, 2003), it is unlikely that the difference in transfer test performance between the retrieval practice condition and the reread-plus-statements condition can be attributed to the total spacing difference. Still, the conditions differ on total spacing, and we cannot completely rule out the possibility that this variability – rather than the experimental manipulation – has confounded the final test results. Hence, future research might include a key statements condition that is modeled after the retrieval practice condition, i.e., with two massed exposures of key information without rereading the total text. Furthermore, the time-on-task should be held fixed and equated between conditions. In this way, total spacing will be equal between conditions, and possible final test differences can be exclusively attributed to the experimental manipulation.

A remaining question is why retrieval practice led to better performance than reread-plus-statements. This finding could be due to both indirect and direct effects that retrieval practice has on learning (Roediger & Karpicke, 2006a). An indirect effect means that the influence of retrieval practice is mediated by another factor, such as motivation. In our study it is possible that during the retrieval attempt, participants became aware of what they did not yet know, causing them to pay more attention in the subsequent feedback. This process might have enhanced their final test scores. That would also explain why there was a short-term testing effect in our first study; in other studies where feedback was provided, a short-term benefit of retrieval practice occurred

as well (e.g., Bishara & Jacoby, 2008; Carpenter, Pashler, Wixted, & Vul, 2008; Carrier & Pashler, 1992; Jacoby, Wahlheim, & Coane, 2010; Kang, 2010; Wartenweiler, 2011). In addition to this indirect effect, retrieval practice might have exerted its influence in a direct way. By retrieving knowledge from memory, the knowledge itself is altered, thereby accommodating retrieval at a later point in time (Karpicke & Grimaldi, 2012).

Taken together, the current study shows that the testing effect in far transfer across different knowledge domains (Butler, 2010) is robust. We replicated the results of Butler's (2010) third experiment with a comparable effect size, indicating that retrieval practice can greatly enhance performance on a far transfer test. However, our results also show that the success of retrieval practice was partly a matter of providing focused exposure to key information. When retrieval practice was compared to a condition that involved rereading the texts and then reading the key information in the form of isolated statements, the benefit of retrieval practice decreased to a fair extent. Hence, the focused exposure to key information (i.e., the feedback) seems be of crucial importance in the retrieval practice condition. Upcoming research could investigate the precise role of focused exposure to key information in the far transfer testing effect.

Acknowledgements

We would like to thank Andy Butler for his helpful discussion of the results and for sharing his materials with us.

7

Summary and general discussion

Taking a test on previously studied material is a powerful way to promote long-term retention, a phenomenon called the retrieval practice (testing) effect. The retrieval practice effect has been demonstrated across a wide range of practice formats, final tests, and materials (e.g., Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). A significant part of the retrieval practice research has focused on relatively simple study materials, like words or word pairs (e.g., Coppens, Verkoeijen, Bouwmeester, & Rikers, 2016; Hogan & Kitsch, 1971; Wheeler, Evans, & Buonanno, 2003). However, there is an increasing number of retrieval practice studies using more educationally relevant material, like expository texts (e.g. Glover, 1989; Kang, Roediger, & McDermott, 2007; Nungester & Duchastel, 1982). In a non-exhaustive literature review, we found twenty-five studies that demonstrated a benefit of retrieval practice over restudy with text materials (see Chapter 5). Furthermore, most retrieval practice research involved final tests that only assessed retention (e.g., Roediger & Butler, 2011; Rowland, 2014), whereas less is known about the effect of retrieval practice on tests that measure *transfer* of knowledge (e.g., Blunt & Karpicke, 2014; Butler, 2010; Eglington & Kang, in press; Foos & Fisher, 1988; Hinze, Wiley, & Pellegrino, 2013; Johnson & Mayer, 2009; Karpicke & Blunt 2011; McDaniel, Howard, Einstein, 2009; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013), which is the ability to apply previously learned knowledge or skills in a novel context and to solve new problems (e.g., Barnett & Ceci, 2002; Mayer, 1996; Salomon & Perkins, 1989).

In this thesis, we have explored the retrieval practice effect with expository and argumentative text and with final tests that measure transfer. In Chapter 2, the fuzzy trace theory of the retrieval practice effect (Verkoeijen, Bouwmeester, & Camp, 2012) was assessed within a near transfer context. In Chapters 3, 4, and 6, we addressed the question whether a retrieval practice benefit would emerge with expository or argumentative text on a final transfer test. In Chapter 3, retrieval practice was compared to *self-explanation* on a final transfer test. In Chapter 5, we focused on the question whether a retrieval practice benefit would occur on a final retention text with the same expository texts as used in Chapter 4. In Chapters 5 and 6, we also explored the additional benefit of providing a re-exposure opportunity after retrieval practice (i.e. feedback).

Summary of main findings

The fuzzy trace theory holds that rereading mostly strengthens the verbatim memory traces of studied information, while retrieval practice mostly strengthens the gist memory traces. The theory therefore predicts that a retrieval practice effect will emerge when people cannot use verbatim/surface cues in a final test, and have to rely

exclusively on semantic/gist cues instead. In Chapter 2, this prediction was tested by gradually reducing the surface features overlap between cues in the learning phase and the final yes/no recognition test over five experiments. First, participants studied simple word lists either by restudy or by free recall retrieval practice. In all five experiments, participants in the control/word condition received as final test cues the same words that had been studied during the learning phase. In Experiment 1, the experimental final test cues were scrambled words, words in a new context, or scrambled words in a new context. In Experiment 2, the experimental final test cues consisted of synonyms, and in Experiments 3, 4a, and 4b, they consisted of images. Such final tests, with only semantic cues available, can be considered to measure (near) transfer, because the surface cues that were present during learning were absent in the final test. The results showed no retrieval practice benefits for any of the final control/word conditions. Moreover, in Experiments 1 and 2, the reduction of surface cue availability in the experimental final tests did not result in a benefit of testing over restudying, which is not in line with the fuzzy trace theory. The three image conditions in Experiments 3, 4a and 4b, on the other hand, did yield a mean recognition benefit of tested items over restudied items, although only Experiment 4a showed a statistical effect. However, the results in the image conditions were small and quite variable. Together these five experiments did not provide strong evidence in support of the fuzzy trace theory of the testing effect.

In Chapter 3, we compared retrieval practice to self-explanation of an argumentative text on an immediate final open-book transfer (comprehension) test. Participants first read an argumentative text that had been part of the Dutch national exams in 2009, and then completed a read-recite-review (RRR) condition, a self-explanation condition, or a baseline control condition. Participants in the RRR-condition first read a paragraph, then recited as much as possible, and afterwards read the paragraph again. Participants in the self-explanation condition clarified and explained the central ideas of each of the paragraphs. In the baseline control condition, participants only read the text for the first time like in the other conditions, and then immediately performed the final open-book multiple-choice transfer test. The results showed that the three learning strategies did not differ on this final test, which suggests that the benefit of retrieval practice might not generalize to this type of argumentative text and/or open-book final test. However, these findings should be interpreted with extreme caution due to the low Cronbach's alphas of the final test. Also, we did not include a long-term retention interval in this study, while the retrieval practice effect is often only observed after a longer interval.

The aim of Chapter 4 was to investigate whether free recall retrieval practice would benefit transfer performance relative to rereading the expository text material. Participants read four expository texts and then engaged in either verbatim free recall, generative recall, or rereading. In the verbatim free recall condition, participants were

asked to type in verbatim everything they could remember from the studied texts. In the generative recall condition, participants were instructed to type in their own words what they had comprehended from the text, allowing more room for elaboration and making inferences within and beyond the text (e.g., Hinze et al., 2013). In the reread condition, participants received five minutes extra study time. The final test consisted of sixteen short-answer transfer questions, administered immediately and after a 1-week delay. These transfer questions were aimed at measuring *inferences* going beyond what was stated in the text. This final test showed forgetting over time, but no differences between the three conditions appeared. These results again suggest that retrieval practice does not foster transfer performance with expository text.

Possibly, the results of Chapter 4 were due to the nature of the final test. That is, it might be that retrieval practice is not useful for transfer performance, especially when no feedback is provided. Only a small number of studies has shown retrieval practice to produce better performance on a final transfer test (e.g., Blunt & Karpicke, 2014; Butler, 2010; Johnson & Mayer, 2009; Karpicke & Blunt 2011; McDaniel et al., 2009), and to the best of our knowledge there is only one study with free recall without feedback as the retrieval practice format (Hinze et al., 2013). These studies suggest that the retrieval practice effect might be weaker with final transfer tests. However, a different explanation for the findings of Chapter 4 is that the text materials that they used were not suitable for producing a testing effect, not even on a pure retention final test. To explore the latter hypothesis, in Chapter 5 we performed an experiment with the same materials as in Chapter 4, where we compared a free recall condition to a reread condition. In contrast to Chapter 4, instead of a final transfer test, we used a final free recall retention test. With this kind of memory test, the beneficial effect of retrieval practice over restudying has been shown to be robust (e.g., Rowland, 2014). So, if the results of Chapter 4 were exclusively driven by the characteristics of the final transfer test, then replacing this kind of final test by a final free recall memory test would confirm the classic retrieval practice effect.

In the first Experiment of Chapter 5, participants read two of the four expository texts that were also used in Chapter 4. After first reading the texts, participants reread one of the texts and performed free recall retrieval practice on the other text. Immediately or after a one week delay, the final free recall final test was administered. On this final test we found a small benefit of rereading over free recall retrieval practice after five minutes, but after one week no difference was left between conditions. Moreover, the interaction effect was too small to be statistically significant. These results suggest that the outcomes as observed in Chapter 4 cannot fully be explained by the nature of the final transfer test, because the typical retrieval practice effect was not reinstated with this pure retention final test.

The second Experiment in Chapter 5 was similar to Experiment 1 but now with a re-exposure opportunity after retrieval practice (i.e., feedback), because feedback is known to enhance the retrieval practice effect (e.g., Rowland, 2014). In Experiment 2, only a small benefit of the free-recall-plus-feedback condition emerged relative to the reread condition, and no interaction effect between study method and retention interval occurred. Together, the results of Chapters 4 and 5 suggest that retrieval-based learning from text is more useful with a retention measure than with a transfer measure, but the effects observed in Chapter 5 were very small and occurred only when feedback was provided.

The additional benefit of providing feedback after retrieval practice was also examined in Chapter 6. The first experiment was a direct replication of the third experiment by Butler (2010), which had been the only study so far to show a retrieval practice testing effect on a final transfer test tapping onto a different knowledge domain. Participants studied expository texts and then either reread them three times or went through three cycles of short-answer questions (cued recall) with feedback (i.e., exposure to the right answer). The final test consisted of sixteen inference questions that required the application of a concept from the studied text to a different knowledge domain. As in Butler (2010), an advantage of retrieval practice emerged on the final transfer test after one week. Additionally, we observed an advantage of retrieval practice on the final transfer test administered after five minutes. Possibly, these advantages were due to the focused exposure to key information (i.e., feedback) that was needed to answer the final transfer questions. That is, although tapping from different knowledge domains, the retrieval practice questions and the final tests questions were conceptually related: the same principles that were learned during retrieval practice had to be applied to the final test questions in the different knowledge domains. We therefore conducted a second experiment with an extra reread-plus-statements condition, which involved rereading the text followed by statements containing the same key information as the feedback in the retrieval practice condition. In this way we investigated whether this exposure to key information could – partly – account for the retrieval practice effect found in the first experiment. The results showed that the far transfer retrieval practice effect was considerably reduced when retrieval practice was compared to the reread-plus-statements condition. Furthermore, retrieval practice and reread-plus-statements both led to better performances on the final test than the rereading condition. Taken together, the results as observed in Chapter 6 demonstrate that Butler's (2010) far transfer effect is robust. Moreover, focused exposure to key information (i.e., feedback) appears to be a significant factor in the retrieval practice effect in far transfer.

General discussion

This thesis has focused on the effects of retrieval practice with expository text on retention (Chapter 5) and on transfer (Chapter 3, 4, and 6). It also explored the additional benefit of providing feedback after retrieval practice (Chapters 5 and 6). Additionally, the fuzzy trace theory of the retrieval practice effect was assessed within a near transfer context (Chapter 2).

Chapter 6 clearly demonstrated a retrieval practice effect with expository text material. However, in the three other studies with text material, the retrieval practice effect was absent (Chapters 3, 4, and 5: Experiment 1) or very small (Chapter 5: Experiment 2). These findings suggest that the retrieval practice effect with expository text may be slightly limited. Indeed, already in 1917 Gates wrote that "the advantage of recitation over reading is greater in learning senseless, non-connected material than in learning senseful, connected material" (p. 23). More recently, Van Gog and Sweller (2015) also claimed that the retrieval practice effect decreases when the complexity or coherence of the study material increases. The underlying reason might be that retrieval practice helps to construct a gist-feature that structures/organizes the material, which then serves as an effective retrieval cue. However, if the material is already coherent, a learner can construct a gist-feature based on the material itself, thereby reducing the advantage of retrieval practice relative to restudy (Delaney, Verkoeijen, & Spirgel, 2010). Hence, when integrated/coherent material is concerned, the function of retrieval practice as organizer of the material might become at least partly redundant, and the advantage of free recall over restudy becomes smaller or disappears (e.g., Bouwmeester & Verkoeijen, 2011; Congleton & Rajaram, 2012; De Jonge, Tabbers, & Rikers, 2015; Karpicke & Zaromb, 2010; Zaromb & Roediger, 2010). However, one would expect that high text coherence would result in a high general recall performance, but Chapters 4 and 5 showed relatively *low* initial and final recall. Possibly as a result of the relatively high coherence of the texts, participants only remembered its gist, which forms only a small part of the whole text. Now, because the degree of text coherence was not manipulated in this thesis, it is impossible to draw any further conclusions on the effect of text coherence on the size of the retrieval practice effect (for a comparison between a coherent and an incoherent text over two experiments, see De Jonge et al., 2015). Still, we did not find a retrieval practice effect in three out of the six experiments that used expository text. Note, however, that in Rowland's meta-analysis (2014) the size of the retrieval practice effect did not differ between coherent (i.e., "semantically related") and non-coherent ("semantically unrelated") materials. Taken together, future research should shed more light on the relationship between text coherence and the retrieval practice effect.

Another possible explanation for the absent (Chapters 3, 4, and 5: Experiment 1) and small (Chapter 5: Experiment 2) retrieval practice effects with expository text are the relatively low percentages of retrieved idea units in the initial tests, which were 55% (Chapter 4), 44% (Chapter 5, Experiment 1) and 43% (Chapter 5, Experiment 2). In Chapter 3 the idea units were not scored. In Chapter 6, where there was a large retrieval practice effect, the percentages of correct responses to the last of the three successive initial tests were as high as 84% (Experiment 1) and 90% (Experiment 2). According to the bifurcation framework (Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011), a test bifurcates the distribution of items' memory strength: memory traces of non-retrieved items remain low in strength while the memory traces of retrieved items become high in strength, resulting in a gap between the two sets of items. Furthermore, items that are restudied are strengthened more in memory than non-retrieved items (but less than retrieved items). Because strong memories last, testing will result in better performance than restudying after an interval that is long enough for only the strongest memories (i.e., the memory representations of items that were retrieved during testing) to survive. Together this implies that when a small number of items is retrieved in the initial test, the benefit of testing will be limited. This theory might explain the results of Chapter 4 and the first experiment of Chapter 5. In the latter, the initial test scores were low (44%), and a small advantage of reread over free recall emerged after five minutes. In Chapter 4, the initial test scores were also relatively low (55%), and again no retrieval practice effect occurred. Furthermore, providing feedback after retrieval practice also increases the memory strength of non-retrieved items, thereby preventing bifurcation to occur. That is, retrieval practice with feedback strengthens the memory traces of all tested items, giving rise to a benefit of testing after both a short-term and a long-term interval. The second experiment of Chapter 5 demonstrated exactly such a pattern of results, although the effect size was small.

Both Chapters 5 and 6 showed that feedback enhances the retrieval practice effect with expository text. Butler and Winne (1995, p. 275) wrote that "feedback is information with which a learner can confirm, add to, overwrite, tune, or restructure information in memory, whether that information is domain knowledge, meta-cognitive knowledge, beliefs about self and tasks, or cognitive tactics and strategies." To further understand the additional effect of feedback with retrieval practice, it is useful to consider the distinction between its direct and indirect effects (Karpicke, & Grimaldi, 2012; Roediger & Karpicke, 2006a). An *indirect* effect of retrieval practice means that the influence of retrieval practice is mediated through another factor, like motivation or feedback. Providing feedback by having someone restudy the text can improve future study because it enables students to correct errors, maintain correct responses, and to improve metacognitive monitoring (e.g., Butler, Karpicke, & Roediger, 2008). Feedback

can also increase performance through affective processes like motivation, effort or engagement (Hattie & Timperley, 2007). A *direct* effect of retrieval practice means that the retrieved knowledge itself is altered, thereby accommodating retrieval at a later point in time (Karpicke & Grimaldi, 2012). Feedback may also influence performance such a direct way, by strengthening the memory traces of the practiced words through restudy.

In sum, we did not find a retrieval practice effect in three out of the six experiments with expository text. However, two factors that boost the effectiveness of retrieval practice with expository text are initial retrieval and feedback. First, a sufficient part of the material needs to be retrieved in order for a retrieval practice to occur. If a retrieval attempt is unsuccessful, there might be no subsequent gain in retrieval practice performance (Bjork & Bjork, 2006). Second, Chapters 5 and 6 showed that feedback increases the beneficial effect of retrieval practice. This finding is line with a number of other studies (e.g., Agarwal et al., 2008; Erdman, & Chan, 2013; Hays, Kornell, & Bjork, 2013; Kang et al., 2007; Pashler et al., 2005; Pashler, Kang, & Mozer, 2013).

The main research question in this thesis was whether retrieval practice is beneficial for transfer performance. Although Chapter 6 did result in a retrieval practice advantage on transfer, Chapters 3 and 4 did not. Now, it is not possible to completely separate the transfer component and the expository text component in these studies. That is, the reason we did not find a retrieval practice transfer effect could have been driven primarily by the fact that we used expository text, rather than by the transfer measures per se (and vice versa; maybe we did not find a retrieval practice effect with expository text because we mostly used transfer measures). To investigate this hypothesis, in Chapter 5 we performed an experiment with the same materials as in Chapter 4, but instead of a final transfer test, a final free recall retention test was used. No retrieval practice effect emerged in this first experiment of Chapter 5. This indicates that the use of expository text might also account for the absence of a retrieval practice transfer effect in Chapter 4. However, the findings in Chapter 5 do not imply that the results on the transfer measure in Chapter 4 were primarily due to the use of expository text. That is, the results on the transfer measure in Chapter 4 might have still been the same if instead of expository text, other material had been used. In other words, the results of Chapter 5 do not imply that the retrieval practice effect does in fact generalize to transfer measures. The present thesis does not tell us which of the two explanations for the limited retrieval practice effect on transfer (i.e., the nature of the final test or the use of text materials), or a combination of two, is correct. This is an example of the problem of *underdetermination* of the theory by the data (e.g., Stanford, 2016).

Still, in Chapters 3 and 4 no retrieval practice effects were found on the final transfer tests, while in Chapter 6, there was. Some differences between the studies, besides the

ones already discussed (i.e., initial retrieval and feedback), might explain these findings. First, in Chapter 3, the final test had an open-book character, and its Cronbach's alpha was very low. Also, we did not include a long-term retention interval in this third study. Possibly, these specific features of the design and the final test can explain why no differences between conditions emerged on the final transfer test. Second, Chapters 3 and 4 employed a free recall retrieval practice format, while Chapter 6 used short-answer questions. Possibly, short-answer questions, which are a more guided form of retrieval practice, are more efficient for transfer performance than free recall. The meta-analysis by Rowland (2014) indeed found that cued recall, the category that comprises short-answer questions, produced larger retrieval practice effects than free recall. However, this comparison should be taken with caution because cued recall was associated with a more frequent use of feedback and higher initial test scores, as is the case in the present thesis. When controlling for these differences, cued recall and free recall yielded comparable testing effects (Rowland, 2014).

The present studies do suggest that the retrieval practice effect does not simply generalize to transfer performance. Indeed, only a small number of studies has shown retrieval practice to produce better performance than rereading on a final transfer test (Blunt & Karpicke, 2014; Butler, 2010; Foos & Fisher, 1988; Hinze, Wiley, & Pellegrino, 2013; Karpicke & Blunt 2011; McDaniel et al., 2009; McDaniel et al., 2013). According to Barnett and Ceci (2002, see also Butler, 2010), three memory demands can be distinguished in the process of transfer: recognition, recall, and execution. First, the learner should *recognize* that prior knowledge can be used in this new situation. Second, the prior knowledge has to be successfully *recalled*. Third, the prior knowledge needs to be applied in order to *execute* the transfer task. However, the usefulness of retrieval practice has been well established with recall tests (e.g., Rowland, 2014). Also, every transfer question in Chapter 6 included a reference to the relevant concept in the initial retrieval practice phase, so there were no recognition demands for these transfer questions. The explanation for the limited transfer success of retrieval practice therefore seems to be that learners did not *execute* the transfer task successfully. However, this thesis shows that a transfer effect can arise when initial retrieval is high and when short answer questions with feedback are used as the retrieval practice format. In sum, only one of the three studies with expository text showed a retrieval practice benefit on a transfer measure. Together the studies indicate that the retrieval practice effect for transfer is limited. However, providing feedback after retrieval practice enhances the retrieval practice transfer effect.

A specific type of near transfer was measured in Chapter 2, where the fuzzy trace theory was assessed by reducing the surface features overlap between cues in the learning phase and the final recognition test. In none of the final control/word conditions,

a retrieval practice effect emerged. In Experiments 1 and 2, no retrieval practice effects arose on the experimental (near transfer) final tests either. The image final test cues, however, did produce a small retrieval practice effect, particularly in Experiment 4a. Still, together the set of experiments provided only weak evidence for the fuzzy trace theory, and did not convincingly show a retrieval practice effect for near transfer. From a broader perspective, this also implies that this type of elaborative retrieval accounts (e.g. Carpenter, 2009, 2011; Pyc & Rawson, 2010) was not substantiated.

The findings of Chapter 2 did not square well with the bifurcation framework. This theory predicts that the more difficult the final test, the larger the benefits of testing. However, in the condition with the lowest average performance (the synonym condition) of Experiment 2, no benefit of retrieval practice emerged. Moreover, in the two most difficult final test conditions of Experiment 1, there was no difference between retrieval practice and rereading. Furthermore, performance in the image condition was lower in Experiment 4b than in Experiment 4a, while the advantage of testing compared to restudy was somewhat larger in Experiment 4a than in Experiment 4b. Taken together, the bifurcation model could not account for the findings in Chapter 2.

Implications and directions for future research

The goal of instruction is to yield knowledge and skills that are durable and flexible, that is, "not only accessible within the instructional context, but ... also accessible in the various post-instructional real-world settings to which they are applicable" (Bjork & Bjork, 2006, p. 109). This thesis suggests that retrieval-based learning from text is more useful with a retention measure than with a transfer measure. Moreover, practicing retrieval is mainly valuable when initial retrieval is high or when feedback is provided. These findings provide important insights for educational practice.

When pure retention is the goal of learning, the educational value of retrieval practice is well established. However, if knowledge needs to be transferred to a different context, the educational usefulness of retrieval practice may be limited. As mentioned above, feedback can boost the effectiveness of retrieval practice for the transfer of knowledge. Providing feedback after retrieval practice is easy to implement, both in class or when studying at home. Also, initial retrieval should be relatively high (but not too high, because retrieval effort needs to be sufficiently high as well). This condition can also be met by providing feedback after a retrieval practice phase, and then to repeat retrieval practice.

The present thesis does not give a concluding answer to the question whether the limited transfer effects were the result of the nature of the final test, the use of expository text, or both. Future research could shed a light on this question by assessing the retrieval practice effect on transfer with other material than expository text (see, e.g.,

Johnson & Mayer, 2009). Also, additional research could explore whether the size of the retrieval practice effect decreases when the degree of text coherence increases. With integrated/coherent text, the function of retrieval practice as organizer of the material might become partly redundant, thereby reducing the advantage of retrieval practice over restudy (e.g., Bouwmeester & Verkoeijen, 2011; Congleton & Rajaram, 2012; De Jonge, Tabbers, & Rikers, 2015; Delaney, Verkoeijen, & Spirgel, 2010; Karpicke & Zaromb, 2010; Zaromb & Roediger, 2010). Taken together, future research should further explore the boundaries of the testing effect in terms of stimuli and types of learning.

Samenvatting
Summary in Dutch

Het *retrieval practice* effect houdt in dat informatie beter wordt onthouden op de lange termijn wanneer de informatie na bestudering wordt opgehaald uit het geheugen (i.e., *retrieval*) dan wanneer deze alleen herhaaldelijk wordt bestudeerd. Dit effect is aangetoond met verschillende typen van retrieval practice, eindtests en studiematerialen e.g., Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). Een groot deel van dit onderzoek heeft zich gericht op eenvoudige studiematerialen, zoals simpele woorden en woordparen (e.g., Coppens, Verkoeijen, Bouwmeester, & Rikers, 2016; Hogan & Kitsch, 1971; Wheeler, Evans, & Buonanno, 2003). Er bestaat echter een toenemend aantal studies naar retrieval practice met materiaal dat nog meer relevantie heeft voor het onderwijs, zoals verklarende teksten (e.g. Glover, 1989; Kang, Roediger, & McDermott, 2007; Nungester & Duchastel, 1982). In een literatuurreview hebben we vijfentwintig studies gevonden die een voordeel van retrieval practice ten opzichte van herbestuderen lieten zien met tekstmateriaal (zie hoofdstuk 5). Verder is er bij het grootste deel van het onderzoek naar retrieval practice gebruik gemaakt van eindtests waarin werd gemeten hoe goed deelnemers de tekst hadden onthouden (e.g., Roediger & Butler, 2011; Rowland, 2014), wat ook wel *retentie* wordt genoemd. Slechts een beperkt aantal studies heeft een voordeel van retrieval practice gevonden op de *transfer* van kennis (e.g., Blunt & Karpicke, 2014; Butler, 2010; Eglington & Kang, in press; Foos & Fisher, 1988; Hinze, Wiley, & Pellegrino, 2013; Johnson & Mayer, 2009; Karpicke & Blunt 2011; McDaniel, Howard, Einstein, 2009; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013). Transfer houdt in dat de eerder verworven kennis wordt gebruikt om binnen een nieuwe context problemen op te lossen (e.g., Barnett & Ceci, 2002; Mayer, 1996; Salomon & Perkins, 1989).

In dit proefschrift is het retrievalpractice-effect met tekstmateriaal en met transfereindtests verder onderzocht. In hoofdstuk 2 werd de *fuzzy trace* theorie van het retrievalpractice-effect (Verkoeijen, Bouwmeester, & Camp, 2012) getoetst in een transfercontext. In hoofdstuk 3, 4 en 6 is onderzocht of er een voordeel van retrieval practice was na het lezen van verklarende/betogende teksten op een transfereindtest, waarbij in hoofdstuk 3 retrieval practice werd vergeleken met de strategie van het aan jezelf uitleggen van de stof. In hoofdstuk 5 werden dezelfde verklarende teksten gebruikt als in hoofdstuk 4, maar nu met een retentie-eindtest. In hoofdstuk 5 en 6 is het bijkomende voordeel onderzocht van het nogmaals aanbieden van (een deel van) de tekst na retrieval practice (i.e., feedback).

Samenvatting van de belangrijkste bevindingen

Volgens de *fuzzy trace* theorie versterkt herlezen voornamelijk de orthografische geheugensporen (i.e., de fysieke en contextuele details) van de bestudeerde informatie, terwijl retrieval practice vooral de semantische geheugensporen activeert. Deze

theorie voorspelt dat een retrievalpractice-effect zal optreden wanneer mensen niet op orthografische cues kunnen terugvallen in de eindtest, maar zich slechts kunnen beroepen op semantische cues. Deze voorspelling is onderzocht in hoofdstuk 2 door in vijf experimenten geleidelijk de overlap in orthografische kenmerken tussen de cues in de leerfase en de latere herkenningstest te verkleinen. Deelnemers bestudeerden eerst simpele woordlijsten door middel van herstudie of door middel van het ophalen van de woorden uit het geheugen (i.e., retrieval practice). In de herkenningseindtest werd deelnemers gevraagd of ze het getoonde woord tijdens de leerfase wel of niet hadden geleerd. In de controleconditie van de vijf experimenten waren de cues in de eindtest gelijk aan de woorden die waren geleerd in de leerfase. In de experimentele conditie van Experiment 1 waren de eindtestcues gelijk aan de bestudeerde woorden, maar nu met de letters door elkaar gehusseld, de woorden gepresenteerd op een andere achtergrond, of een combinatie van die twee. In Experiment 2 waren de experimentele eindcues synoniemen van de bestudeerde woorden, en in Experimenten 3, 4a, en 4b waren het plaatjes van de bestudeerde woorden. In dergelijke eindtests zijn er alleen semantische cues beschikbaar, en ontbreken dus de orthografische cues zoals die beschikbaar waren in de leerfase. Deze experimentele eindtests kunnen daarom beschouwd worden als maten voor (*near*) transfer. De resultaten van de vijf experimenten lieten geen voordeel van retrieval practice zien in een van de controlecondities. Ook in de experimentele eindtestcondities van Experiment 1 en 2 trad er geen voordeel op van retrieval practice, wat niet strookt met de voorspellingen van de *fuzzy trace* theorie. In de drie experimentele plaatjescondities van Experiment 3, 4a en 4b was er echter wel een voordeel van retrieval practice, hoewel alleen Experiment 4a een statistisch ("significant") effect liet zien, en de effecten over het geheel genomen klein en variabel waren. Tezamen leveren deze vijf experimenten hooguit zwak bewijs ter ondersteuning van de *fuzzy trace* theorie van het retrievalpractice-effect. Ook laat deze studie zien dat het retrievalpractice-voordeel op de transfereindtests beperkt was.

In hoofdstuk 3 werd retrieval practice vergeleken met het aan zichzelf uitleggen van de bestudeerde tekst op een transfereindtest. Deze eindtest had een openboekvorm en werd direct na de studiefase werd afgenomen. Deelnemers lazen eerst de betogende tekst, die afkomstig was uit de Nederlandse eindexamens van 2009. Daarna doorliepen ze een retrievalpractice-conditie, een uitlegconditie of een controleconditie. In de retrievalpractice-conditie lazen ze eerst een alinea van de tekst, schreven dan alles op van wat ze zich herinnerden van die alinea en lazen daarna de alinea opnieuw. Deze procedure doorliepen ze voor alle alinea's. Aan deelnemers in de uitlegconditie werd gevraagd om de belangrijkste ideeën per paragraaf aan zichzelf uit te leggen op papier. In de controleconditie maakten deelnemers de eindtest direct na het lezen van de tekst. Op deze eindtest werden uiteindelijk geen verschillen gevonden tussen de

drie condities, wat suggereert dat het retrievalpractice-effect niet generaliseert naar dit type tekst en/of naar een dergelijke openboektransfertest. De Cronbachs alfa van deze eindtest was echter zeer laag, waardoor deze resultaten met terughoudendheid geïnterpreteerd dienen te worden. Ook werd de eindtest alleen direct na de studiefase afgenomen en niet na een langer retentie-interval, terwijl het retrievalpractice-effect juist pas vaak na een langer interval optreedt.

Het doel van hoofdstuk 4 was om te onderzoeken of retrieval practice van een verklarende tekst de prestatie op een transfermaat zou bevorderen ten opzichte van herlezen. Deelnemers lazen vier teksten en doorliepen daarna een van de volgende drie condities: letterlijk ophalen, constructief ophalen, of herlezen. Bij het letterlijk ophalen werd deelnemers gevraagd om alles letterlijk op te halen wat ze zich nog konden herinneren van de tekst. Bij het constructief ophalen werd deelnemers gevraagd om in eigen woorden uit te leggen wat ze hadden begrepen van de tekst. Met deze instructie werden ze meer aangemoedigd om te elaboreren en inferenties te maken (e.g., Hinze et al., 2013). In de herleesconditie kregen deelnemers vijf minuten extra studietijd. De eindtest bestond uit zestien korte open tranfervragen en werd zowel direct na de studiefase als na een week afgenomen. Deze eindtest was gericht op het maken van inferenties, dus het afleiden van nieuwe informatie die niet letterlijk in de tekst stond. De gemiddelde prestatie bleek beter op de eindtest die direct werd afgenomen dan op die afgenomen na een week, maar tussen de drie condities werden geen verschillen gevonden. Opnieuw duiden deze resultaten erop dat retrieval practice niet voordelig is voor de prestatie op een transfermaat.

Een alternatieve verklaring voor de resultaten in hoofdstuk 4 is echter dat de materialen sowieso niet geschikt waren voor het vinden van een retrievalpractice-effect, ook niet op een standaard retentiemaat. Op een dergelijke maat is het voordeel van retrieval practice immers al overtuigend is aangetoond (e.g., Rowland, 2014). Om deze hypothese te onderzoeken werd in hoofdstuk 5 retrieval practice vergeleken met herlezen op een retentie-eindtest, met dezelfde materialen als gebruikt in hoofdstuk 4. Als de resultaten van hoofdstuk 4 volledig toe te schrijven waren aan het specifieke type eindtest dat werd gebruikt, dan zou het vervangen van deze test door een standaard retentietest in hoofdstuk 5 het klassieke retrievalpractice-effect moeten herbevestigen.

In het eerste experiment van hoofdstuk 5 lazen deelnemers twee van de vier verklarende teksten die ook werden gebruikt in hoofdstuk 4. Hierna herlazen ze een van de teksten en ze haalden de andere tekst op uit hun geheugen. Onmiddellijk of na een week werd de retentie-eindtest afgenomen. Op deze eindtest werd na vijf minuten een klein voordeel gevonden van herlezen, maar na een week was er geen verschil meer tussen de condities. Ook werd er geen statistisch interactie-effect gevonden. Deze resultaten suggereren dat de uitkomsten in hoofdstuk 4 niet volledig verklaard kunnen

worden aan de hand van het type eindtest dat gebruikt werd, omdat het klassieke retrievalpractice-effect ook niet gevonden werd op de retentiemaat in hoofdstuk 5.

Het tweede experiment van hoofdstuk 5 was vergelijkbaar met Experiment 1, maar nu met feedback aangeboden na retrieval practice, omdat feedback het positieve effect van retrieval practice versterkt (e.g., Rowland, 2014). In dit experiment werd er geen interactie-effect gevonden en slechts een klein voordeel van retrieval practice in vergelijking met herlezen. Tezamen suggereren hoofdstuk 4 en 5 dat retrieval-gestuurd leren bij verklarende teksten nuttiger is wanneer de eindtest retentie meet dan wanneer er transfer wordt gemeten. Het voordeel van *retrieval practice* op de retentiemaat in hoofdstuk 5 was echter erg klein en trad bovendien alleen op wanneer er feedback werd aangeboden na retrieval practice.

Het voordeel van het aanbieden van feedback na retrieval practice is verder onderzocht in hoofdstuk 6. Het eerste experiment in hoofdstuk 6 was een directe replicatie van het derde experiment van de studie van Butler (2010). Dit experiment (Butler, 2010) is tot nu toe het enige waarin een voordeel van retrieval practice ten opzichte van herlezen is gevonden op een transfermaat die betrekking had op een geheel nieuw kennisdomein. Deelnemers lazen eerst zes verklarende teksten. Daarna herlazen ze drie van deze teksten drie keer en beantwoordden ze over de drie andere teksten drie keer dezelfde korte open vragen ("*cued recall* retrieval practice"), waarna ze het goede antwoord te zien kregen. De eindtest bestond uit zestien transfervragen waarin een concept uit de bestudeerde tekst moest worden toegepast binnen het nieuwe kennisdomein. Net als in de studie van Butler (2010) bleek er een voordeel te zijn van retrieval practice op de eindtest die werd afgenomen na een week. Ter aanvulling werd in ons experiment de eindtest ook afgenomen direct na de studiefase, waarop wederom een voordeel verscheen van retrieval practice ten opzichte van herlezen.

Een mogelijke verklaring voor de resultaten van het eerste experiment is echter dat het voordeel van retrieval practice kan worden toegeschreven aan de gerichte blootstelling aan de cruciale informatie tijdens de feedback. De informatie in de feedback was namelijk essentieel voor het correct beantwoorden van de transfervragen in de eindtest. Met andere woorden, de tussentijdse vragen en de uiteindelijke transfervragen waren conceptueel gerelateerd: dezelfde principes die waren geleerd tijdens retrieval practice met feedback dienden te worden toegepast voor het beantwoorden van de eindtestvragen in het nieuwe kennisdomein. Om deze verklaring te onderzoeken is er een tweede experiment uitgevoerd met dezelfde condities als in het eerste experiment maar met een extra herlees-plus-statements-conditie, waarin eerst de tekst werd herlezen en daarna een aantal statements met daarin dezelfde informatie als in de feedback van de retrievalpractice-conditie. Op deze manier is onderzocht of het succes van retrieval practice in het eerste experiment (deels) verklaard kon worden door

deze gerichte blootstelling aan de cruciale informatie (i.e., de feedback). De resultaten lieten zien dat het retrievalpractice-effect reduceerde wanneer retrieval practice werd afgezet tegen een herlees-plus-statements-conditie. Verder bleken de retrievalpractice-conditie en de herlees-plus-statements-conditie te resulteren in een betere prestatie op de transfereindtest dan de herleesconditie. Samengevat tonen de resultaten van hoofdstuk 6 aan dat het transfereffect zoals gevonden in Butler (2010) robuust is, maar dat de feedback hierin een belangrijke rol speelt.

Conclusie

Samenvattend is er in slechts drie van de zes experimenten met verklarende/betogende tekst een retrievalpractice-effect gevonden: in de twee experimenten van hoofdstuk 6 en een klein effect in het tweede experiment van hoofdstuk 5. Een mogelijke verklaring voor de afwezigheid van een retrievalpractice-voordeel in de hoofdstukken 3, 4 en 5 (experiment 1) zijn de relatief lage percentages opgehaalde informatie tijdens retrieval practice. De percentages opgehaalde informatie-eenheden waren 55% (hoofdstuk 4), 44% (hoofdstuk 5, experiment 1) en 43% (hoofdstuk 5, experiment 2), terwijl in hoofdstuk 6 de percentages 84% (experiment 1) en 90% (experiment 2) waren. In hoofdstuk 3 zijn deze percentages niet gescoord. Deze uitkomsten zijn consistent met de meta-analyse van Rowland (2014), die liet zien dat het retrievalpractice-effect groter wordt naarmate de prestatie op de tussentijdse retrieval practice toeneemt. Wanneer echter feedback wordt aangeboden na retrieval practice dan is het succes van retrieval practice maximaal, onafhankelijk van de hoeveelheid opgehaalde informatie tijdens retrieval practice (Rowland, 2014). Deze bevinding is in lijn met de uitkomsten van dit proefschrift, aangezien hoofdstuk 6 en ook hoofdstuk 5 laten zien dat feedback de positieve invloed van retrieval practice vergroot.

De belangrijkste onderzoeksvraag in dit proefschrift was of retrieval practice nuttig is voor de transfer van kennis. Alleen hoofdstuk 6 bleek te resulteren in een voordeel van retrieval practice ten opzichte van herlezen op een transfermaat. In hoofdstuk 2 is er slechts beperkt bewijs gevonden voor een voordeel van retrieval practice op transfer. In de hoofdstukken 3 en 4, waarin teksten werden gebruikt als studiemateriaal, werd geen retrievalpractice-voordeel op transfer gevonden. Het is echter niet mogelijk om in de tekststudies van hoofdstuk 3, 4 en 6 de transfercomponent en de tekstcomponent goed te onderscheiden. Dat wil zeggen dat het niet mogelijk is om te bepalen of het beperkte succes van retrieval practice op transfer nu voortkomt uit de gebruikte tekstmaterialen, de aard van de eindtest, of een combinatie van beide. Wel is duidelijk dat een voordeel van retrieval practice op een transfermaat deels kan worden verklaard op basis van de aangeboden feedback na retrieval practice.

R

References

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*, 861–876. doi: 10.1002/acp.1391

Ainsworth, S., & Burcham, S. (2007). The impact of text coherence on learning by self-explanation. *Learning and Instruction, 17*, 286–303. doi:10.1016/j.learninstruc.2007.02.004

Ainsworth, S. & Loizou, A. T. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science, 27*, 669–681. doi: 10.1207/s15516709cog2704_5

Amlund, J. T., Kardash, C. A. M, & Kulhavy, R. W. (1986). Repetitive reading and recall of expository text. Reading Research Quarterly, 21, 49–58. doi: 10.2307/747959

Anderson, J. R. (1974). Verbatim and propositional representations of sentences in immediate and long-term memory. *Journal of Verbal Learning and Verbal Behavior, 13*, 149–162. doi: 10.1016/S0022-5371(74)80039-3

Atkinson, R. K., Renkl, A. & Merrill, M. M. (2003). Transition from studying examples to solving problems: effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology, 95,* 774–783. Doi: 10.1037/0022-0663.95.4.774

Baddeley, A. D. (1976). *The Psychology of Memory*. New York, NY: Basic Books.

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

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge, England: Cambridge University Press.

Bielaczyc, K., Pirolli. P. L., & Brown, A. L. (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. *Cognition and Instruction, 13,* 221–252.

Bishara, A. J., & Jacoby, L. L. (2008). Aging, spaced retrieval, and inflexible memory performance. *Psychonomic Bulletin and Review*, *15*, 52–57. doi: 10.3758/PBR.15.1.52

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

Bjork, R. A., & Bjork, E. L. (2006). Optimizing treatment and instruction: Implications of a new theory of disuse. In L-G. Nilsson & N. Ohta (Eds.), *Memory and Society: Psychological Perspectives* (pp. 109–133). New York, NY: Psychology Press.

Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology, 3*, 849–858. doi: 10.1037/a0035934

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

Brainerd, C. J., & Reyna, V. F. (2004). Fuzzy-trace theory and memory development. *Developmental Review, 24,* 396–439. doi: 10.1016/j.dr.2004.08.005

Buhrmester, M., Kwang, T., & Gosling, S.D. (2011). Amazon's Mechanical Turk. A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3–5. doi: 10.1177/1745691610393980

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

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

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

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*, 245–281.

Callender, A. A., & McDaniel, M. (2009). The limited benefits of rereading educational texts. *Contemporary Education Psychology*, *34*, 30–41. doi:10.1016/j.cedpsych.2008.07.001

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

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

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

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name-learning. *Applied Cognitive Psychology, 19,* 619–636. doi: 10.1002/acp.1101

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

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin and Review, 14*, 474–478. doi: 10.3758/BF03194092.

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

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

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effect of tests on learning and forgetting. *Memory and Cognition*, *36*, 438–448. doi: 10.3758.MC.36.2.438

Carpenter, S. K., & Yeung, K. L. (2017). The role of mediator strength in learning from retrieval. *Journal of Memory and Language, 92,* 128–141. doi: 10.1016/j.jml.2016.06.0080749-596X

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory and Cognition*, *20*, 632–642. doi: 10.3758/BF03202713

Cartwright, C. (1991). Replicability, reproducibility, and robustness: Comments on Harry Collins. *History of Political Economy*, *23*, 143–155.

Casler, K., Bickel, L. & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*, 2156–2160. doi: 10.1016/j.chb.2013.05.009

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*, 354–380. doi: 10.1037/0033-2909.132.3.354

Chamberland, M, St-Onge, C., Setrakian, J., Lanthier, L., Bergeron, L., Bourget, A., Mamede, S., Schmidt, H., & Rikers, R. (2011). The influence of medical students' self-explanations on diagnostic performance. *Medical Education, 45*, 688–695. doi: 10.1111/j.1365-2923.2011.03933.x

Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory, 18*, 49–57. doi: 0.1080/09658210903405737

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology*, *4*, 553–571. doi: 10.1037/0096-3445.135.4.553

Chi, M. T. H. (2009). Active-constructive-interactive : A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, *1*, 73–105. doi: 10.1111/j.1756-8765.2008.01005.x

Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self–explanations improves understanding. *Cognitive Science, 18*, 439–477. doi: 10.1207/s15516709cog1803_3

Congleton, A. R., & Rajaram, S. (2012). The origin of the interaction between learning history and delay in the testing effect: The roles of processing and retrieval organization. *Memory and Cognition*, *40*, 528–539. doi: 10.3758/s13421-011-0168-y

Coppens, L. C., Verkoeijen, P. P. J. L., Bouwmeester, S., & Rikers, R. M. J. P. (2016). The testing effect for mediator final test cues and related final test cues in online and laboratory experiments. *BMC psychology, 4*, 1–14, doi: 10.1186/s40359-016-0127-2

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

Coursey, D., Hovis, J., & Schulze, W. (1987). The disparity between willingness to accept and willingness to pay measures of value. *Quarterly Journal of Economics, 102*, 679–690. doi: 10.2307/1884223

Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting, *European Journal of Cognitive Psychology, 21*, 919–940. doi: 10.1080/09541440802413505

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25,* 7–29.

De Bruin, A., Rikers, R. M. J. P., & Schmidt, H. G. (2007). The effect of self-explanation and prediction on the development of principled understanding of chess in novices. *Contemporary Educational Psychology, 32*, 1740–1761. doi: 10.1016/j.cedpsych.2006.01.001

De Groot, A. M. B. (2002). Lexical representation and lexical processing in the second language user. In Cook, V. (Ed.), *Portraits of the L2 User* (pp. 29–63). Clevedon: Multilingual Matters.

De Jonge, M. O., Tabbers, H. K., & Rikers, R. M. J. P. (2015). The effect of testing on the retention of coherent and incoherent text material. *Educational Psychology Review, 27*, 305–315. doi: 10.1007/s10648-015-9300-z

De Koning, B. B., Tabbers, H. K., Rikers, R. M. J. P., & Paas, F. (2011). Improved effectiveness of cueing by self-explanations when learning from a complex animation. *Applied Cognitive Psychology, 25*, 183–194. doi: 10.1002/acp.1661

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58,* 17–22. doi: 10.1037/h0046671

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. S. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol.53, pp. 63–147). Burlington: Academic Press. doi: 10.1016./S0079-7421(10)53003-2

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*, 274–290. doi: 10.1177/1745691611406920

Dirkx, K. J., Kester, L., & Kirschner, P. A. (2014). The testing effect for learning principles and procedures from texts. *The Journal of Educational Research, 107*, 357–364. doi: 10.1080/00220671.2013.823370

Dobson, J. L., & Linderholm, T. (2015). Self-testing promotes superior retention of anatomy and physiology information. *Advances in Health Science Education: Theory and Practice, 20*, 149–161. doi: 10.1007/s10459-014-9514-8.

Duchastel, P. C. (1981). Retention of prose following testing with different types of tests. *Contemporary Educational Psychology, 6*, 217–226.

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*, 4–58. doi: 10.1177/1529100612453266.

Eglington, L.G., & Kang, S.H.K. (in press). Retrieval practice benefits deductive inference. *Educational Psychology Review*. doi:10.1007/s10648-016-9386-y

Erdman, M. R., & Chan, J. C. K. (2013). Providing corrective feedback during retrieval practice does not increase retrieval-induced forgetting. *Journal of Cognitive Psychology, 25*, 692-703. doi:10.1080/20445911.2013.790389

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Fiorella, L. & Mayer, R. (2015). Eight ways to promote generative learning. *Educational Psychology Review, 28*, 717–741. doi: 10.1007/s10648-015-9348-9

Fonseca, B., & Chi, M. T. H. (2011). Instruction based on self-explanation. In R. Mayer & P. Alexander (Eds.), *The Handbook of Research on Learning and Instruction* (pp. 296–321). New York, NY: Routledge Taylor and Frances Group.

Furr, R. M., & Bacharach, V.R. (2014). *Psychometrics. An Introduction*. Thousand Oaks, CA: SAGE Publications.

Gámez, E., Diaz, J.M., & Marrero H. (2011). The uncertain universality of the Macbeth effect with a Spanish sample. *The Spanish Journal of Psychology, 14*, 156–162.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, 6 (40), 1–104.

Gerard, L. D., & Scarborough, D. L. (1989). Language-specific access of homographs by bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 305–315. doi: 10.1037/0278-7393.15.2.305

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior*, *15*, 1–16. doi: 10.1016/S0022-5371(76)90002-5

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

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37, 801–812. doi: 10.1037/a0023219

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. doi: 10.3102/003465430298487

Hays, M. J., Kornell, N, & Bjork, R. A. (2013) When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 290–296. doi: http://dx.doi.org/10.1037/a0028468

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory, 19*, 290–304. doi: 10.1080/09658211.2011.560121.

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language, 69*, 151–164. doi: 10.1016/j.jml.2013.03.00

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior, 10,* 562–567. doi: 10.1016/S0022-5371(71)80029-4

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economy, 14*, 399–425. doi: 10.1007/s10683-011-9273-9

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 0696–0701. doi: 10.1371/journal.pmed.0020124

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1441–1451. doi: 10.1037/a0020636

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446. doi:10.1016/j.jml.2007.11.007

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 3*, 621–629. doi: 10.1037/a0015183

Johnson, C.J., Paivio, A., & Clark, J. M. (1996). Cognitive components of picture naming. *Psychological Bulletin, 120,* 113–139. doi: 10.1037/0033-2909.120.1.113

Jones, H. E. (1923–1924). The effects of examination on the performance of learning. *Archives of Psychology*, 10, 1–70.

Kamalski, J., Sanders, T., Lentz, L., & Van Den Bergh, H. (2005). Hoe kun je het beste meten of een leerling een tekst begrijpt? Een vergelijkend onderzoek naar vier methoden. *Levende Talen Tijdschrift*, *4*, 3–9.

Kang, S. H. K. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory and Cognition, 38*, 1009–1017. doi: 10.3758/MC.38.8.1009

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

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, *21*, 157–163. doi: 10.1177/0963721412443552

Karpicke, J. D.**,** & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review, 27,* 317-326. doi: 10.1007/s10648-015-9309-3

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*, 1250–1257. doi: 10.1037/a0023436

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

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Meta-cognitive strategies in student learning: Do students practice in retrieval when they study on their own? *Memory, 17*, 471–479. doi: 10.1080/09658210802647009

Karpicke, J. D., & Grimaldi, P. J. (2012). Retrieval-based learning: A perspective for enhancing meaningful learning. *Educational Psychological Review, 24*, 401–418. doi: 10.1007/s10648-012-9202-2

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation, 61*, 237–284. San Diego, CA: Elsevier Academic Press. doi: 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*, 704–719. doi: 0.1037/0278-7393.33.4.704

Karpicke, J. D., & Roediger, H. L. (2010). Is expanding retrieval a superior method for learning text materials? Memory and Cognition 38, 116–124. doi:10.3758/MC.38.1.116

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

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49*, 249–303. doi: 10.1037/0003-066X.49.4.294

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. New York: Cambridge University Press.

Kintsch, W., & Rawson, K. A. (2005). Comprehension. In M. J. Snowling & C. Hulme (Eds.), The science of reading: A handbook (pp. 209–226). Malden, MA: Blackwell.

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language*, *29*, 133–159. doi:10.1016/0749-596X(90)90069-C

Kirsner, K., Brown, H. L., Abrol, S., Chadna, N. K., & Sharma, N. K. (1980). Bilingualism and lexical representation. *Quarterly Journal of Experimental Psychology, 32,* 585–594. doi: 10.1080/14640748008401847

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr., R. B., Bahník, Š., Bernstein, M. J., . . .

Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*, 142–152. doi: 10.5334/jopd.ad

Kline, R. B. (2004). *Beyond Significance Testing. Reforming Data Analysis Methods in Behavioral Research*. Washington D.C.: APA Books.

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

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

Kvanvig, J. (2003). *The Value of Knowledge and the Pursuit of Understanding*. Cambridge: Cambridge University Press.

Kvanvig, J. (2009). The Value of Understanding. In A. Haddock, A. Millar & D. H. Pritchard (Eds.), Epistemic Value (pp. 95–111). Oxford: Oxford University Press. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174. doi: 10.2307/2529310

Lakens, D., & Etz, A. (2017). Too true to be bad: When sets of studies with significant and non-significant findings are probably true. *Social Psychological and Personality Science*. doi: 10.1177/1948550617693058

Larsen, D. P., Butler, A. C., & Roediger, H. L. (2009). Repeated testing improves long-term retention relative to repeated study: A randomized, controlled trial. Medical Education, 43, 1174–1181. doi: 10.1111/j.1365-2923.2009.03518.x

Larsen, D. P., Butler, A. C., & Roediger, H. I. (2013). Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Medical Education, 47*, 674–682. doi: 10.1111/medu.12141

Lehman, M., Karpicke, J. D. (2016). Elaborative retrieval: Do semantic mediators improve memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42,* 1573–1591. doi: 10.1037/xlm0000267

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D. Verhagen, A. J., , Ly, A., Gronau, Q. F., Smira, M., Epskamp, S., Matzke, D., Wild, A., Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2015). JASP (*Version 0.8.0.0*) [Computer software].

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

Magliano, J. P., Trabasso, T., & Graesser, A. C. (1999). Strategic processing during comprehension. *Journal of Educational Psychology, 91,* 615–629.

Mayer, R.E. (1996). Learning strategies for making sense out of expository text: The SOI model for guiding three cognitive processes in knowledge construction. *Educational Psychology Review, 8*, 357–371.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-Enhanced Learning in a Middle School Science Classroom: The Effects of Quiz Frequency and Placement. *Journal of Educational Psychology*. 103, 399–414. doi: 10.1037/a0021782

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

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The Read-Recite-Review study strategy: Effective and portable. *Psychological Science, 20,* 516–522. doi: 10.11 11/j.1467-9280.2009.02325.x

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

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied, 20*, 3–21. doi: 10.1037/xap0000004

McEldoorn, K. L., Durkin, K. L., & Rittle-Johnson, B. (2012). Is self-explanation worth the time? A comparison to additional practice. *British Journal of Educational Psychology, 83,* 615–632. doi:10.1111/j.2044-8279.2012.02083.x

McNamara, D. S. (2004). SERT: Self-Explanation Reading Training. *Discourse Processes, 38*, 1–30.doi: 10.1207/s15326950dp3801_1

McNamara, D. S., & Magliano, J. (2009) Toward a comprehensive model of comprehension. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 51, pp. 297–384). Burlington: Academic Press. doi: 10.1016/S0079-7421(09)51009-2

Meuffels, B., & Van Den Bergh, H. (2006). De ene tekst is de andere niet: The language-as-a-fixed-effect fallacy revisisted: Statistische implicaties. *Tijdschrift voor Taalbeheersing, 28*, 323–345.

Morey, R. D., & Lakens, D. (2016). Why most of psychology is statistically unfalsifiable. Manuscript submitted for publication. Retrieved from https://github.com/richarddmorey/psychology_resolution/blob/master/paper/response.pdf

Neuman, Y., & Schwarz, B. (1998). Is self-explanation while solving problems helpful? The case of analogical problem-solving. *British Journal of Educational Psychology, 68*, 15–24. doi: 10.1111/j.2044-8279.1998.tb01271.x

Nokes, T. J., Hausmann, R. G. M., VanLehn, K., & Gershman, S. (2011). Testing the instructional fit hypothesis: The case of self-explanation prompts. *Instructional Science, 9,* 645 –666. doi: 10.1007/s11251-010-9151-4

Nokes-Malach, T. J., VanLehn, K., Belenky, D. M., Lichtenstein, M., & Cox, G. (2012). Coordinating principles and examples through analogy and self-explanation. *European Journal of Educational Psychology, 28*, 1237–1263. doi: 10.1007/s10212-012-0164-z

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

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology, 74*, 18–22. http://dx.doi.org/10.1037/0022-0663.74.1.18

O'Reilly, T., Symons, S., & MacLatchy-Gaudet, H. (1998). A comparison of self-explanation and elaborative interrogation. *Contemporary Educational Psychology, 23*, 434–445. doi: 10.1006/ceps.1997.0977

Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. Perspectives on Psychological Science, *7*, 657–660. doi: 10.1177/1745691612462588

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*, 63-71. doi: 10.1207/S15326985EP3801_8

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411–419.

Pashler, H., Kang, S. H. K., & Mozer, M. C. (2013). Reviewing erroneous information facilitates memory updating. *Cognition, 128*, 424–430. doi: 10.1016/j.cognition.2013.05.002

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7,* 528–530. doi:10.1177/1745691612465253

Perea, M., & Rosa, E. (2002). The effect of associative and semantic priming in the lexical decision task. *Psychological Research, 66*, 180–194. doi: 10.1007/s00426-002-0086-5

Ploetzner R., Dillenbourg P., Praier M. & Traum D. (1999). Learning by explaining to oneself and to others. In P. Dillenbourg (Ed.), *Collaborative-learning: Cognitive and Computational Approaches* (pp. 103–121). Oxford: Elsevier.

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

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: Application of the SAM model. *Cognitive Science*, *27*, 431–452. doi: 10.1016/S0364-0213(03)00007-7

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology, 299*, 172–179. doi:10.1016/j.jtbi.2011.03.004

Rawson, K. A. (2015). The status of the testing effect for complex materials: Still a winner. Educational Psychology Review, *27*, 327–331. doi: 10.1007/s10648-015-9308-4

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 and Cognition*, *43*, 619–633. doi: 10.3758/s13421-014-0477-z

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science*, 21, 1–29. doi: 10.1207/s15516709cog2101_1

Richey, J. E., & Nokes-Malach, T. J. (2015). Comparing four instructional techniques for promoting robust knowledge. *Education Psychology Review, 27*, 181–218. doi: 10.1007/s10648-014-9268-0

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

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development, 77*, 1–15. doi: 10.1111/j.1467-8624.2006.00852.x

Robinson, F. P. (1941). *Effective study*. New York: Harper and Brothers.

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. M. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*, 382–395. doi: 10.1037/a0026252

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

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications of educational practice. *Perspectives on Psychological Science, 17*, 181–208. doi: 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,* 249–255. doi: 10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 803–814. doi: 10.1037/0278-7393.21.4.803

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

Ross, J., Irani, L., Six Silberman, M., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers? Shifting demographics in Mechanical Turk. In *Proceedings of CHI 2010* (pp. 2863–2872). Atlanta, GA: ACM.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374. doi: 10.1016/j.jmp.2012.08.001

Rouder, J. N., Morey, R. D., Verhagen, A.J., Swagman, A.R., & Wagenmakers, E.-J. (2017).

Bayesian analysis of factorial designs. *Psychological Methods, 22,* 304–321. doi: 10.1037/met0000057

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*, 225–237. doi: 10.3758/PBR.16.2.225

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

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

Roy, M., & Chi, M. T. H. (2005). Self-explanation in a multi-media context. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 271–286). New York, NY: Cambridge University Press.

Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics, 2,* 437–442. doi: 10.3758/BF03208784

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. *Educational Psychologist, 24,* 113–142. doi: 10.1207/s15326985ep2402_1

Scarborough, D. L., Gerard, L., & Cortese, C. (1984). Independence of lexical access in bilingual word recognition. *Journal of Verbal Learning and Verbal Behavior, 23,* 84–89. doi:10.1016/S0022-5371(84)90519-X

Schworm, S., & Renkl, A. (2006). Computer-supported example-based learning: When instructional explanations reduce self-explanations. *Computers and Education, 46*, 426–445. doi: 10.1016/j.compedu.2004.08.011

Sensenig, A. E., Littrell-Baez, M. K., & DeLosh, E. L. (2011). Testing effects for common versus proper names, *Memory*, *19*, 664–673. doi: 10.1080/09658211.2011.599935

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*, 90–100. doi: 10.1037/a0015108

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. doi:10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). *A 21 Word Solution*. Retrieved from: https://ssrn.com/abstract=2160588. doi: 10.2139/ssrn.2160588

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*, 559–569. doi: 10.1177/0956797614567341

Smith, M. A., Blunt, J. R., Whiffen, J. W., & Karpicke, J. D. (2016). Does providing prompts during retrieval practice improve learning? Applied Cognitive Psychology, 30, 544–553. doi: 10.1002/acp.3227

Smith, M., Roediger, H. L. & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 460–471. doi: 10.1037/a0033569

Snow, C. (2002). *Reading for Understanding: Toward a R & D Program in Reading Comprehension*. Santa Monica, CA: RAND.

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

Stanford, P. K. (2016, March 21). Underdetermination of scientific theory. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy. Retrieved from https://plato.stanford.edu/archives/spr2016/entries/scientific-underdetermination/.

Tran, R., Rohrer, D., & Pashler, H. (2014). Retrieval practice: the lack of transfer to deductive inferences. *Psychonomic Bulletin Review, 22,* 135–140. doi: 10.3758/s13423-014-0646-x.

Van Eersel, G.G., Verkoeijen, P. P. J. L., Povilenaite, M., & Rikers, R. (2016). The testing effect and far transfer: The role of exposure to key information. *Frontiers in Psychology, 7*, 1977. doi: 10.3389/fpsyg.2016.01977.

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*, 247–264. doi: 10.1007/s10648-015-9310-x

Verkoeijen, P. P. J. L. , Bouwmeester, S., & Camp, G. (2012). A short term testing effect in cross-language recognition. *Psychological Science, 23*, 567–571. doi: 10.1177/0956797611435132

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, A.J., Selker, R., Gronau, Q.F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., Van Kesteren, E.-J., Van Doorn, J., Smira, M., Epskamp, S., Etz, A., Matzke, D., Rouder, J. N., & Morey, R. D. (2016, November 2). Bayesian Inference for Psychology. Part II: Example Applications with JASP. Retrieved from https://osf.io/ahhdr/

Wartenweiler, D. (2011). Testing effect for visual-symbolic material: Enhancing the learning of Filipino children of low socio-economic status in the public school system. *International Journal of Research and Review, 6,* 74–93.

Weinstein, Y., McDermott, K. B., & Roediger, H. L. (2010). A comparison of study strategies for passages: rereading, answering questions, and generating questions. *Journal of Experimental Psychology: Applied, 16*, 308–316. doi: 10.1037/a0020992

Wetzels, R. Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J. & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6*, 291–298. doi: 10.1177/1745691611406923

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

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

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

Wiley, J., Griffin, T. D., & Thiede, K. W. (2005) Putting the comprehension on metacomprehension. *The Journal of General Psychology, 132*, 408–428. doi: 10.3200/GENP.132.4.408-428

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology, 91,* 301–311. 10.1037/0022-0663.91.2.301

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition, 3*, 214–221. doi:10.1016/j.jarmac.2014.07.001

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory and Cognition, 38*, 995–1008. doi: 10.3758/MC.38.8.995

Zeelenberg, R., & Pecher, D. (2003). Evidence for long-term cross-language repetition priming in conceptual implicit memory tasks. *Journal of Memory and Language, 49*, 80–94. doi: 10.1016/S0749-596X(03)00020-2

C

Curriculum vitae

Curriculum vitae

Gerdien van Eersel was born in Rotterdam, The Netherlands, in 1981. She received a Bachelor's degree in Psychology from Maastricht University in 2007 and a Master's degree in Clinical Psychology in 2008. For her master's thesis she did a clinical internship at Psychology Practice Rekkers in Amsterdam, which specializes in the treatment of eating disorders. In 2008, Gerdien also obtained a Bachelor's degree in Philosophy of Science at Erasmus University Rotterdam, and in 2012 she obtained a Master's degree in Philosophy of Science. Her master's thesis was on the extrapolation of experimental results in Psychology with the help of the statistical technique Latent Class Analysis. In 2008, she worked as a teaching assistant at the Faculty of Philosophy (EUR) to set up a minor in Philosophy. Later in 2008, she started working as a consultant and teacher in statistics at the *Methodology Shop* of the Erasmus School of Social and Behavioural Sciences (ESSB, former: Faculty of Social Sciences). Gerdien enrolled in a Ph.D. program at ESSB in 2011, where she investigated the retrieval practice effect with text materials and transfer measures. She also taught courses in statistics, and was a member of the Faculty Council from 2013 to 2015. In 2015, she was representative of ESSB in a project group to improve participation at Erasmus University Rotterdam. Currently, Gerdien is working as a researcher at the municipality of Zaanstad.

Publications

Van Eersel, G. G., Bouwmeester, S., Verkoeijen, P. P. J. L., Tabbers, H. K. & Rikers, R. M. J. P. (2017). Does retrieval practice depend on semantic cues? Assessing the fuzzy trace account of the testing effect. *Journal of Cognitive Psychology, 29,* 583–598. doi: http://dx.doi.org/10.1080/20445911.2017.1300156

Van Eersel, G. G., Verkoeijen, P. P. J. L., Povilenaite, M., & Rikers, R. M. J. P. (2016). The testing effect and far transfer: The role of focused exposure to key information. *Frontiers in Psychology*, *7*, 1977. doi: https://doi.org/10.3389/fpsyg.2016.01977

Ruiter, R. A. C., Verplanken, B., & **Van Eersel, G. G.** (2003). Strengthening the persuasive impact of fear appeals: The role of action framing. *Journal of Social Psychology*, 143, 397–400. doi: 10.1080/00224540309598452

Submitted manuscripts

Van Eersel, G. G., Koppenol-Gonzalez, G. V., & Reiss, J. (*under revision for Philosophy of Science*). The average: Still in the running to be human sciences' top model? Extrapolation of experimental results on the basis of latent classes.

Van Eersel, G. G., Verkoeijen, P. P. J. L., Tabbers, H. K., Van Mierlo, S. A. A., Paas, F., & Rikers, R. M. J. P. (*under review*). A comparison of study strategies for inference learning: Reread, verbatim recall, and generative recall.

Presentations

Van Eersel, G. G., Verkoeijen, P. P. J. L., Povilenaite, M., & Rikers, R. M. J. P. (2016). *Het effect van toetsen op transfer: Grotendeels een kwestie van feedback?* Paper presented at the Onderwijs Research Dagen, Rotterdam, NL (May 25–27).

Van Eersel, G. G., Verkoeijen, P. P. J. L. & Rikers, R. M. J. P. (2015). *The testing effect and far transfer: Largely a matter of feedback*. Poster presented at the Annual Meeting of the Psychonomic Society, Chicago, Illinois, USA (Nov 19–22).

Van Eersel, G. G., Verkoeijen, P. P. J. L. & Rikers, R. M. J. P. (2014). *The effect of self-explanation and testing on reading comprehension.* Poster presented at the Annual Meeting of the Psychonomic Society, Long Beach, California, USA (Nov 20–23).

D

Dankwoord

Een aantal mensen ben ik veel dank verschuldigd. Allereerst Peter Verkoeijen, rots in de branding. Jij stond *altijd* voor me klaar met je briljante ideeën en waardevolle feedback. Je was een van de redenen dat ik voor dit project solliciteerde; ik wist dat ik veel van je zou leren. Dan mijn promotor Remy Rikers, dank voor je inzichten, je vertrouwen en voor de spiegel die je me voorhield. Op de cruciale momenten zei jij precies de juiste dingen. Via jouw reflecties heb ik mezelf beter leren kennen. Ook dank aan Huib Tabbers voor je aanstekelijke energie, motiverende ideeën en originele perspectieven.

Graag wil ik ook de statistieksupervrouwen bedanken die mij de eerste jaren aan de ESSB zo goed hebben begeleid en wijs hebben gemaakt in de statistiek: Samantha Bouwmeester, Marike Polak en Lidia Arends. Jullie zijn nog steeds voorbeelden voor mij. Prof. dr. Lidia Arends wil ik tevens danken voor het plaatsnemen in de leescommissie om dit proefschrift te beoordelen, samen met prof. dr. Fred Paas en prof. dr. Liesbeth Kesters. Ook wil ik de leden van de grote commissie bedanken: prof. dr. Tamara van Gog, dr. Gino Camp, prof. dr. Rolf Zwaan en prof. dr. Fred A. Muller. Een speciaal woord van dank voor prof. dr. Muller, die als niet-psycholoog in deze commissie wilde plaatsnemen. Ik ben vereerd dat zulke goede en integere wetenschappers met mij van gedachten willen wisselen tijdens de verdediging van mijn proefschrift.

Ik dank mijn paranimf en partner-in-testing Mario de Jonge voor de mooie reizen en de niet-aflatende stroom aan grappen, en mijn andere paranimf Gabriela Koppenol-Gonzalez voor de bijzondere vriendschap. Ik hoop dat jullie nog lang mijn paranimfen blijven.
Graag noem ik hier ook mijn kamergenoten Wim Pouw en Charly Eielts: jullie maakten dat ik me thuis voelde. Wim, dank voor je geestverwantschap en je enthousiasme. Charly, dank voor je interesse, humor en wijze adviezen.

De afdeling Psychologie/Pedagogiek heeft altijd aangevoeld als een warm bad. Hierbij dank aan de gehele afdeling, en speciaal aan Martine, Mirella, Jesper, Rob, Marloes, Migle, Christiaan, Gerrit Jan, Gertjan, René, Joran, Sabine, Lysanne, Margina, Margot, Jacqueline, Iris, Sofie, Jan, Marise, Noortje, Lisette, Kim, Nicole, Sander, Tim, Steven, Marit, Bonnie, Stijn, en tevens aan Diane, die als wetenschapper voor mij een voorbeeld is. *Last but not least* dank aan Vincent, voor de fijne gesprekken.

Mijn tijd aan de ESSB zit erop. Als filosofische wijsneus kwam ik binnen; op de hoogte van de wetenschappelijke theorie maar van de praktijk nog geen kaas gegeten. Het was een leerzame periode, niet alleen als wetenschapper, maar ook als lid van de Faculteitsraad, waar ik werd geconfronteerd met het contrast tussen mijn idealen en de micropolitieke werkelijkheid. Daarom hierbij ook mijn dank aan alle leden van de Faculteitsraad 2013-

2015 en aan het managementteam, in het bijzonder aan prof. dr. Henk van der Molen voor de mogelijkheid om twee maanden extra aan dit proefschrift te besteden.

Als laatste dank ik mijn moeder Arja, een kwarteeuw tussen de sterren, mijn vader Jaap, mijn zusje Soo Ja, mijn vriend Michiel, en mijn leuke en slimme neefje El Charoy, voor alles.