

# The Effect of Testing on the Retention of Coherent and Incoherent Text Material

Mario de Jonge · Huib K. Tabbers · Remy M. J. P. Rikers

Published online: 17 March 2015

© The Author(s) 2015. This article is published with open access at Springerlink.com

**Abstract** Research has shown that testing during learning can enhance the long-term retention of text material. In two experiments, we investigated the testing effect with a fill-in-the-blank test on the retention of text material. In Experiment 1, using a coherent text, we found no retention benefit of testing compared to a restudy (control) condition. In Experiment 2, text coherence was disrupted by scrambling the order of the sentences from the text. The material was subsequently presented as a list of facts as opposed to connected discourse. For the incoherent version of the text, testing slowed down the rate of forgetting compared to a restudy (control) condition. The results suggest that the connectedness of materials can play an important role in determining the magnitude of testing benefits for long-term retention. Testing with a completion test seems most beneficial for unconnected materials and less so for highly structured materials.

**Keywords** The testing effect · Retrieval practice · Text coherence · Long-term retention

In recent years, there has been a resurgence of interest for the potential benefits of testing on long-term retention. Research has shown that taking tests during learning can have profound effects on later recall compared to less demanding learning strategies like repeated study (Roediger and Karpicke 2006a, 2006b). The general findings are especially surprising, since repeated study will most often result in superior performance on a recall test given shortly after learning. However, this short-term benefit is not long lasting. Repeated study will generally result in a relatively fast rate of forgetting, while successful retrieval of information during learning slows down the rate of forgetting (Carpenter et al. 2008; Wheeler et al. 2003). Consequently, testing generally results in superior recall performance after a relatively long retention interval. This so-called *testing effect* (also known as *the retrieval practice effect*) has been found with different types of materials and different types of tests and using a variety of retention interval conditions (Roediger and Karpicke 2006a). Claims have been made that the testing effect is of critical importance for education, and these claims have been corroborated by studies replicating

---

M. de Jonge (✉) · H. K. Tabbers · R. M. J. P. Rikers  
Department of Psychology, Institute of Psychology, Erasmus University Rotterdam, Woudestein, T12-45,  
P.O. Box 1738, 3000 Rotterdam, Netherlands  
e-mail: dejonge@fsw.eur.nl

the general findings in actual classroom settings (e.g., Carpenter et al. 2009; McDaniel et al. 2007).

The powerful effect of testing for simple verbal material has been consistently found using different types of tests (e.g., Carpenter et al. 2006; Carpenter et al. 2008; Karpicke and Roediger 2007, 2008; Kuo and Hirshman 1996; Pyc and Rawson 2009; Toppino and Cohen 2009; Wheeler et al. 2003). However, the positive effect of testing on retention appears to be less robust in studies using texts as to-be-learned materials. In particular, studies using test formats that are commonly used in education (i.e., short-answer questions) have come up with somewhat inconsistent results. That is, some studies have found benefits of testing only when feedback was provided after taking a test, but not when feedback was withheld (e.g., Kang et al. 2007; LaPorte and Voss 1975). One reason why these studies might have failed to find a benefit of testing without feedback could be due to low initial retrieval on the practice test (Kang et al. 2007; Wenger et al. 1980). For instance, in the Kang et al. (2007) study, recall on an initial practice test was only 54 % correct. The authors hypothesized that giving corrective feedback could restore the effectiveness of the test. Accordingly, in Experiment 2, they found that testing with feedback enhanced 3-day recall performance relative to a restudy (control) condition.

Other studies have found benefits of testing over restudy even when no feedback was given to participants. For instance, Nungester and Duchastel (1982) found that taking a short-answer test enhanced long-term recall performance for a factually oriented history passage. Also, in a more recent study, Hinze and Wiley (2011) found similar results using expository science texts and a fill-in-the-blank test. Interestingly, in their study, positive effects of taking a test were found even though performance on the initial practice test was well below 50 % correct. In Experiment 1 of their study, taking a fill-in-the-blank test enhanced recall performance on a similar test given 2 days later, compared to a restudy (control) condition. In Experiment 2, they found that taking a fill-in-the-blank test enhanced recall on a test given after a 1-week delay. However, in Experiment 3 of their study, taking a fill-in-the-blank test did not enhance recall on a subsequent multiple-choice test given 2 days later. This finding is especially surprising, since initial practice test performance was considerably higher in Experiment 3 (62 % correct) compared to performance in the first two experiments (44 and 45 % correct, respectively). This indicates that the failure to obtain a benefit testing in Experiment 3 was not due to insufficient recall on the practice test. The authors suggest that the failure to obtain a retention benefit of testing in Experiment 3 of their study might be due to the change in test format on the final test. However, as they also note, other researchers have generally found evidence suggesting that taking a short answer test can facilitate later multiple choice test performance (Kang et al. 2007; McDaniel et al. 2007; Nungester and Duchastel 1982). In other words, the absence of a testing effect in Experiment 3 of the Hinze and Wiley (2011) study cannot be readily explained.

To sum up, the testing effect is a well-established phenomenon in the literature. However, the effect appears to be less consistent and less robust in studies using cued recall like tests (e.g., short-answer questions) for learning text material compared to studies using cued recall tests for learning simple verbal materials. Interestingly, other researchers have made similar observations across different types of materials in the past. In fact, in some of the very first studies on the effects of testing, it was already noted that the benefits of testing varied considerably across different kinds of to-be-learned materials (e.g., Gates 1917; Kühn 1914). For instance, Kühn (1914) found that the benefit for nonsense syllables was quite large. However, for learning series of

words, the benefit was smaller, and for learning short verses, testing appeared to be least beneficial. Kühn concluded that the relative advantage of testing appeared to increase as the to-be-learned materials became less meaningful. Gates (1917) obtained similar results for unconnected material (nonsense syllables) and connected material (biographies). He concluded that testing appeared to be most beneficial for unconnected material and less so for connected material.

In the present study, we aimed to investigate two possible explanations for the inconsistencies in testing effect studies using text materials. A first possibility is that some of the inconsistencies reported in the literature are simply the result of the way recall was assessed. That is, in most testing studies using texts and short-answer tests, recall was assessed only after a single (long) retention interval. In the present study, we assessed recall at two retention intervals which enabled us to investigate the rate of forgetting. Second, in Experiment 2, we investigated the possible role of the connectedness of the to-be-learned materials on the effect of testing.

## Experiment 1

One possible explanation for the inconsistent results in testing effect studies using short-answer questions could be the way recall performance was assessed. Testing effect studies using short answer questions have almost exclusively assessed recall after relatively long retention intervals of days or weeks (e.g., Butler 2010; Duchastel 1981; Hinze and Wiley 2011; Kang et al. 2007; LaPorte and Voss 1975; Nungester and Duchastel 1982). Assessing recall performance at a single point in time makes it impossible to directly investigate the course of forgetting. As noted earlier, one of the unique advantages of taking a recall test is that it slows down the rate of forgetting (Wheeler et al. 2003). Since testing effect studies using short-answer questions have assessed recall only after relatively long intervals, we do not know how short-answer tests might affect the course of forgetting. For instance, a benefit of testing found after a relatively long interval could also be the result of an initial difference between conditions which has simply persisted over the course of the retention interval. This possibility pertains especially to those studies using tests with corrective feedback during initial learning, because testing with feedback can also improve recall performance after a relatively short retention interval (e.g., Butler et al. 2008). On the other hand, it could also be the case that the absence of a testing effect found after a certain interval reflects the point in time where the respective forgetting functions following different conditions of practice crossover (e.g., Wheeler et al. 2003). In that case, there can be no apparent difference in recall performance after a relatively long interval even though the preceding courses of forgetting were different.

In sum, the conflicting results in testing effect studies using text materials and short-answer tests could simply stem from the fact that recall was assessed solely after a single long retention interval. Perhaps the results from these studies would have been more consistent if the course of forgetting had been the subject of investigation. In Experiment 1 of the present study, we investigated this possibility. Instead of looking at recall performance after a single long (1-week) retention interval, we also included a short (5-min) retention interval. If taking a short-answer test improves the retention of text material, then the rate of forgetting over the course of the retention interval should be slower following a short-answer test compared to a restudy (control) condition.

## Method

*Participants* Sixty-nine psychology students from the Erasmus University Rotterdam participated in partial fulfillment of course requirements. Five participants were excluded for failing to show up for the 1-week session of the experiment. Of the remaining 64 participants, one half ( $N=32$ ) were assigned to the study condition, and the other half ( $N=32$ ) to the testing condition. Half of each group was tested after 5 min, and the other half after a week.

*Materials* For the purposes of the present experiment, a Dutch text about black holes was created. The text was 1070 words in length and consisted of 60 sentences. The information presented in the text was taken from several online sources (see [Appendix](#)). To obtain a rough estimate of readability for the black hole text, we used the sentence-to-sentence comparison feature on the Latent Semantic Analysis website (<http://lsa.colorado.edu/>). The average sentence-to-sentence cosine for a translated version of the black hole text was .39, indicating that the text was highly coherent (Foltz et al. 1998).

For testing purposes, a short-answer (fill-in-the-blank) test was created similar to the test used by Hinze and Wiley (2011). The test was created in such a fashion that it closely matched the restudy (control) condition. The test contained the exact same 60 sentences presented in the black hole text, but with information selectively omitted from it. Every single sentence contained one omission covering between one and three words in length. To get an estimate of prior knowledge, we asked 10 additional participants to answer the questions without having read the text prior to taking the test. Naturally, these participants did not participate in any of the other experiments using the black hole text. On average, participants were able to correctly answer 12 % ( $SD=4$  %) of the questions. Table 1 shows a translated excerpt from the black hole text with corresponding fill-in-the-blank questions. E-prime (Psychology Software Tools, Pittsburgh, PA) was used to create and run the experiment.

*Design and Procedure* A  $2 \times 2$  between-subjects design was used with learning condition (restudy vs. testing) and retention interval (5 min vs. 1 week) as independent variables, and test score as dependent variable.<sup>1</sup> In the first session of the experiment, all participants first studied the text during a 15-min learning trial. The text was presented one sentence at a time in the middle of the computer screen, and participants could proceed to the next sentence in the text by pressing the ENTER key. This kind of sentence-by-sentence reading procedure is a commonly used procedure in research on text coherence and comprehension (see also Lorch and O'Brien 1995). Note that, because the study was self-paced, it was possible to read the sentences more than once. After the last sentence of the text had been studied, the text was presented again one sentence at a time. Participants continued to study the text in this manner until the total of 15-min study time had expired. At the bottom left of the screen, participants received feedback about their progress (e.g., 3/60 indicated a participant was currently reading the third sentence out of 60 sentences), and at the bottom right of the screen, the remaining time was displayed. Upon completion of the first 15-min block, instructions diverged. During

<sup>1</sup> Of main interest in the present study was recall performance. However, all participants also received some evaluation questions about the materials used in the experiment and a transfer test given after completion of the final recall test. The evaluation and transfer data were collected for exploratory purposes. For the sake of brevity, the results are not reported in the present version of the manuscript. Interested readers can obtain the results from the first author upon request.

**Table 1** Translated (from Dutch) excerpt from the black hole text with corresponding fill-in-the-blank questions

Excerpt from the text	<p>Most black holes rotate, because the stars from which they are formed also rotate.</p> <p>Space outside a rotating black hole is dragged along with the black hole.</p> <p>The result is a sort of cosmic whirlpool where it is impossible for objects to remain stationary.</p> <p>This area, where everything is forced to move with the black hole, is called the ergosphere.</p>
Fill-in-the-blank questions	<p>Most black holes rotate, because the ..... from which they are formed also rotate.</p> <p>Space outside a rotating black hole is ..... with the black hole.</p> <p>The result is a sort of cosmic whirlpool where it is impossible for objects to .....</p> <p>This area, where everything is forced to move with the black hole, is called the .....</p>

the subsequent 15-min study block, one group of participants continued to study the text material, whereas the other group of participants received a 15-min fill-in-the-blank test. Participants in the testing condition were told that the text would again be presented to them but that each sentence would now have some information omitted from it. They were told that they should try and complete the sentences by typing in the missing information using the keyboard. No corrective feedback was given during testing. As in the initial block, both restudy and testing were self-paced, so participants could go through the text or test more than once. Time on task was equated between the two learning conditions.

Following the learning phase, all participants worked on Sudoku puzzles for 5 min as a distractor task. Afterward, half of the participants received a final fill-in-the-blank test identical to the one used in the learning phase of the experiment. The other half of participants received the final test 1 week later.

## Results and Discussion

*Scoring* The responses on the cued recall test were scored by awarding 1 point for every correct response, 0.5 points for partially correct responses, and 0 points for completely incorrect responses. For a small number of items, paraphrases were possible. Paraphrased responses that contained the same meaning conveyed by the original text were scored as correct.

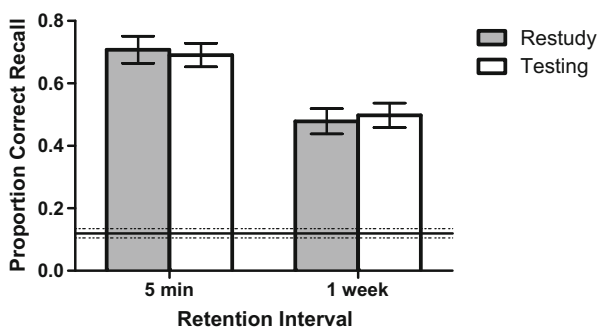
*Learning Phase* For both conditions, we calculated the average number of study or test cycles during the initial learning phase (i.e., the mean number of sentences processed divided by the total number of sentences in the text). During the first block, participants in the restudy condition studied the text 2.74 times ( $SD=1.01$ ), and participants in the testing condition studied the text 2.45 times ( $SD=.79$ ). The difference in number of study cycles did not reach the level of significance,  $F(1, 62)=1.69, p=.20$ . During the second block, participants in the restudy condition studied the text 2.85 times ( $SD=.85$ ), while participants in the testing condition went through the test 1.60 times ( $SD=.73$ ). Analysis showed that, for the second block, the difference in number of cycles was significant,  $F(1, 62)=39.52, p<.001, \eta_p^2=.39$ , indicating that the fill-in-the-blank test was more time consuming compared to simply restudying the information. This finding is not surprising, and in line with the general idea

that overt retrieval practice during a test requires more time and effort compared to restudying (see also Roediger and Karpicke 2006a). On average, participants in the testing condition scored 67 % correct on the test.

**Recall Performance** Figure 1 shows the mean proportion of correct recall for both learning conditions as a function of retention interval. Participants in the 5-min group outperformed the participants in the 1-week group (70 vs. 50 %),  $F(1, 60)=27.48, p<.001, \eta_p^2=.31$ , suggesting that forgetting occurred during the 1-week interval. However, there was hardly any difference between the restudy and the test conditions at both intervals. On the 5-min test, participants in the restudy condition correctly recalled 71 %, and participants in the testing condition correctly recalled 69 %. On the 1-week test, participants in the restudy condition correctly recalled 48 %, and participants in the testing condition correctly recalled 50 %. The main effect of learning condition and the learning condition  $\times$  retention interval interaction did not reach the level of statistical significance (both  $F<1$ ). Thus, we did not find a difference in the rate of forgetting between the restudy and the testing condition.

## Experiment 2

The results from Experiment 1 extend those from previous studies. By looking at recall performance after two retention intervals rather than using a single long interval, we investigated the effect of taking a fill-in-the-blank test on the rate of forgetting. Importantly, however, we found no evidence for the idea that taking a fill-in-the-blank test can slow down the rate of forgetting. The results from the present study and those from previous studies (e.g., Kang et al. 2007; Hinze and Wiley 2011) seem to suggest that the benefits of testing might be less robust for text material. This could be related to some critical aspect of the materials used. One distinctive feature of text material, relative to simple verbal materials like foreign language vocabulary word pairs, is the highly structured and organized fashion by which information is presented. A text is not simply a list of facts that has been randomly put together, but rather it is a coherent set of ideas presented in a very particular and logical order. Studies have shown that text coherence can have profound effects on the retention of text material (Britton and Gülgöz 1991; Kintsch 1994). Especially when readers have



**Fig. 1** Proportion correct on the final recall test as a function of learning condition and retention interval in Experiment 1. The horizontal line represents baseline recall test performance for the coherent fill-in-the-blank test used in Experiment 1. Error bars represent standard errors of the means

little prior knowledge, text coherence is a very important factor determining learning from text (McNamara 2001; McNamara and Kintsch 1996).

The issue of text coherence has received very little consideration in research on the testing effect. Still, as already noted, the idea that the organization or connectedness of materials might attenuate the effect of testing is not at all new. In some of the pioneering work (e.g., Gates 1917; Kühn 1914) on this topic, it was already noted that the benefits of testing can vary considerably across different kinds of materials, and it has been suggested that the connectedness of to-be-learned materials might play an important role determining the magnitude of the benefits of testing. To investigate the possible role of connectedness, we conducted a second experiment. In Experiment 2, we disrupted the coherence of the text material used in Experiment 1 by presenting the information contained in the text as a list of randomly ordered facts (low element interactivity) rather than connected discourse.

## Method

*Participants* Seventy psychology students from the Erasmus University Rotterdam participated in partial fulfillment of course requirements. None of the participants had participated in Experiment 1. Data from five participants were excluded from analysis, because they failed to show up for the 1-week session of the experiment. Data from one participant were excluded for failing to follow basic instructions. Of the remaining 64 participants, one half ( $N=32$ ) were assigned to the study condition, and the other half ( $N=32$ ) to the testing condition. Half of each group was tested after 5 min, and the other half after a week.

*Materials* The coherence of the text used in Experiment 1 was disrupted by presenting the sentences in a scrambled order. In order to be comprehensible out of context, it was necessary to make some minor changes to the sentences taken from the black hole text. For instance, in some sentences, an adverb was deleted (e.g., “So, black holes are...” was changed to “Black holes are...”). Also, in some sentences, anaphoric references were replaced by their corresponding nouns (e.g., “they” was replaced with “black holes”). The average sentence-to-sentence cosine (<http://lsa.colorado.edu/>) of the scrambled text in Experiment 2 was significantly lower ( $M=0.23$ ,  $SD=0.19$ ) than the cosine of the text used in Experiment 1 ( $M=0.39$ ,  $SD=0.22$ ),  $t(116)=4.16$ ,  $p<.001$ ,  $d=0.77$ , indicating that the disruption of the text coherence had been successful. A fill-in-the-blank test was subsequently devised containing the exact same omissions as the test used in Experiment 1. The presentation order of items on the scrambled fill-in-the-blank test was kept constant throughout the experiment. As in Experiment 1, we asked 10 additional participants to answer the questions without having studied the materials prior to taking the test. Baseline test performance for the scrambled version of the test was similar to performance in Experiment 1. On average, participants were able to correctly answer 11 % ( $SD=7$  %) of the questions.

*Design and Procedure* As in Experiment 1, we used a two (learning condition) $\times$ 2 (retention interval) between-subjects design. The procedure was virtually identical to the one used in Experiment 1. The only important difference was the way we referred to the to-be-learned materials in the instructions. In Experiment 2, participants were told that they would learn a list of facts about black holes.



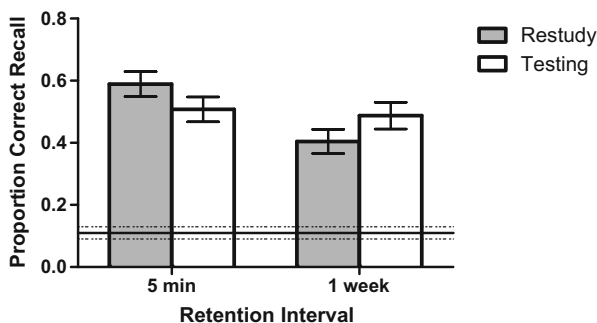
## Results and Discussion

**Learning Phase** During the first block, participants in the restudy condition studied the list of facts 1.97 times ( $SD=.69$ ) and participants in the testing condition studied the text 1.86 times ( $SD=.55$ ). The difference in number of study cycles did not reach the level of significance,  $F < 1$ . During the second block, participants in the restudy condition studied the list of facts 2.29 times ( $SD=.94$ ), while participants in the testing condition went through the test 1.45 times ( $SD=.44$ ). As in Experiment 1, taking the test was more time consuming compared to simply restudying the list of facts,  $F(1, 62)=21.0$ ,  $p < .001$ ,  $\eta_p^2=.25$ . Participants in the testing condition scored 54 % correct on the test.

**Recall Performance** Figure 2 shows the mean proportion of correct recall for both learning conditions as a function of retention interval. There was a significant main effect of retention interval,  $F(1, 60)=6.42$ ,  $p < .05$ ,  $\eta_p^2=.10$ . Participants in the 5-min group recalled more on the final test (54 %) compared to participants in the 1-week group (45 %). The main effect for learning condition did not reach the level of significance,  $F < 1$ . Importantly, however, there was a significant learning condition  $\times$  retention interval interaction,  $F(1, 64)=4.13$ ,  $p < .05$ ,  $\eta_p^2=.06$ . As can be seen in Fig. 2, the restudy group showed a substantial amount of forgetting (31 %). However, for the testing group, there was no apparent decline in recall performance across the 1-week interval. Accordingly, follow-up analysis revealed that the effect of retention interval was significant for the restudy condition,  $t(30)=3.33$ ,  $p < .001$ ,  $d=1.18$ , but not for the testing condition ( $t < 1$ ). Thus, for the incoherent materials used in Experiment 2, we observed a difference in rate of forgetting between the restudy (control) condition and the testing condition.

## General Discussion

In the present study, we aimed to investigate two possible explanations for the inconsistencies in testing effect studies using text materials and completion tests. One possibility was related to the way recall was assessed in most previous studies. As noted, in most previous studies on the effect of short-answer testing, recall performance was assessed after a single long-term retention interval. In our study, we assessed recall at two retention intervals which enabled



**Fig. 2** Proportion correct on the final recall test as a function of learning condition and retention interval in Experiment 2. The horizontal line represents baseline recall test performance for the scrambled fill-in-the-blank test used in Experiment 2. Error bars represent standard errors of the means



us to investigate the rate of forgetting. In Experiment 1, using a highly coherent text, we found no apparent retention benefit of testing compared to a restudy (control) condition. The testing group and the restudy group showed comparable rates of forgetting over the course of the 1-week interval. However, in Experiment 2, when text coherence was disrupted, we found that testing slowed down the rate of forgetting compared to a restudy (control) condition. Taken together, these results suggest that the benefits of testing can be dependent on the connectedness of the to-be-learned materials.

Past research on text coherence has shown that the connectedness of material can have a powerful effect on later recall of text material (Britton and Gülgöz 1991; Kintsch 1994). Since we disrupted the coherence of the black hole text and presented the material as a list of facts in Experiment 2, one would expect that test scores would be lower in Experiment 2 compared to those in Experiment 1. Inspection of the practice test and retention test scores in Experiments 1 and 2 suggests that this was indeed the case. Averaged across conditions, participants in Experiment 2 performed worse compared to the participants in Experiment 1 on the retention test (50 vs. 59 % correct), and also on the practice test (54 vs. 67 % correct). As already noted, researchers have argued that testing can sometimes be ineffective when recall is relatively low on an initial practice test (e.g., Kang et al. 2007). Interestingly, in Experiment 2 of the present study, we found that testing slowed down the rate of forgetting even though recall performance on the initial practice test was considerably lower compared to performance in Experiment 1. Thus, the discrepancy in results between the two experiments in the present study cannot be explained by the amount recalled on the practice tests. In fact, given the level of recall performance on the practice tests in the respective experiments, one would have expected a more pronounced testing effect in Experiment 1 rather than in Experiment 2.

In Experiment 2, using a list of facts, we found evidence suggesting that testing can slow down the rate of forgetting. It has been argued that tests appear to slow down the rate of forgetting because taking a practice test can result in stronger memory traces for successfully retrieved items compared to non-recalled items or restudied items (Halamish and Bjork 2011; Kornell et al. 2011). One reason why tests might result in stronger memory traces is offered by the *elaborative retrieval hypothesis* (e.g., Carpenter 2009; Carpenter and DeLosh 2006). This hypothesis suggests that testing will result in more elaborate memory traces compared to passive restudy of information. Support for this hypothesis has been provided by studies showing that the effect of testing can get more pronounced as the amount of cue support on the practice tests diminishes. For instance, in a study by Carpenter and DeLosh (2006), it was found that retrieving items with fewer letter cues was associated with better final recall test performance. One way to explain the results from the present study could be in light of the elaborative retrieval hypothesis. As already noted, in a coherent text, ideas are presented in a very particular logical order. It has been argued that the organizational structure of text materials can also serve as a retrieval cue to enhance later recall (Shimmerlik 1978). In the present study, the coherent context of the materials used in Experiment 1 might also have functioned as a retrieval cue. If this was the case, then one could argue that the test used in Experiment 1 might not have resulted in more elaborate processing relative to the processing already invited by the cue support provided by the context of the text. However, for the isolated statements in Experiment 2, the absence of the supporting context of the text might have resulted in more elaborative processing on the test. Investigating this possible explanation could be a fruitful avenue to pursue in future research.

In the present study, we investigated the testing effect using a short-answer test. Clearly, our conclusions are limited to the test format used and, importantly, we would not want to imply that testing as a general learning activity might not be a useful tool for learning texts. In fact, some studies using free recall tests have found substantial memorial benefits of testing for

learning texts (e.g., Hinze and Wiley 2011; Karpicke and Blunt 2011; Roediger and Karpicke 2006b). Interestingly, research suggests that taking a free recall test can also facilitate organizational processing (Congleton and Rajaram 2011, 2012; Zaromb and Roediger 2010). In the case of learning from text, organizational processing seems especially important. Perhaps a free recall test is a more potent device for improving the retention of text material compared to a short-answer test such as the one used in the present study, because a free recall test results in more organizational processing.

To conclude, the results of the present study indicate that the benefits of testing can be dependent on the connectedness of the materials. These results are in line with observations from earlier research across different kinds of to-be-learned materials (e.g., Gates 1917; Kühn 1914). Also, in a similar vein, other researchers have recently argued that testing might not be beneficial for the learning of high element interactivity material in a problem solving task (Leahy et al. 2015). The present study represents a first step toward explaining the discrepancy between different kinds of materials by addressing the issue in two experiments using material of differential coherence, but equal content. Future research is necessary to establish to which extent coherence plays a role in the testing effect. However, on the basis of our results, we have identified coherence as one possible factor determining the relative benefits of testing.

## Appendix

### Online Sources Used to Create the Materials for Experiment 1 and 2

[http://nl.wikipedia.org/wiki/Zwart\\_gat](http://nl.wikipedia.org/wiki/Zwart_gat)

<http://www.rug.nl/sciencelinx/blackholegame/home.html>

<http://www.kennislink.nl/publicaties/zwarte-gaten>

**Open Access** This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

## References

- Britton, B. K., & Gülgöz, S. (1991). Using Kintsch's computational model to improve instructional text: effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology*, *83*, 329–345.
- 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.
- Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2008). Correcting a metacognitive error: feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 918–928.
- 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.
- Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: support for the elaborative retrieval explanation of the testing effect. *Memory and Cognition*, *34*, 268–276.
- 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.
- 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.
- Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory and Cognition*, *36*, 438–448.

- Congleton, A. R., & Rajaram, S. (2011). The influence of learning methods on collaboration: prior repeated retrieval enhances retrieval organization, abolishes collaborative inhibition, and promotes post-collaborative memory. *Journal of Experimental Psychology: General*, *140*, 535–551.
- Congleton, A. R., & Rajaram, S. (2012). The origin of the interaction between learning method and delay in the testing effect: the roles of processing and conceptual retrieval organization. *Memory and Cognition*, *40*, 528–539.
- Duchastel, P. C. (1981). Retention of prose following testing with different types of test. *Contemporary Educational Psychology*, *6*, 217–226.
- Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *25*, 285–307.
- Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *6*, 1–104.
- 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.
- Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using short answer tests. *Memory*, *19*, 290–304.
- Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 528–558.
- Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*, 772–775.
- Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*, 151–162.
- Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968.
- Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, *49*, 294–303.
- 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.
- Kühn, A. (1914). Über Einprägung durch Lesen und durch Rezitieren. *Zeitschrift für Psychologie*, *68*, 396–481.
- Kuo, T. M., & Hirshman, E. (1996). Investigations of the testing effect. *American Journal of Psychology*, *109*, 451–464.
- LaPorte, R. E., & Voss, J. F. (1975). Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology*, *67*, 259–266.
- Leahy, W., Hanham, J., & Sweller, J. (2015). High element interactivity information during problem solving may lead to failure to obtain the testing effect. *Educational Psychology Review*.
- Lorch, R. F., & O'Brien, E. J. (Eds.). (1995). *Sources of coherence in reading*. Hillsdale: Erlbaum.
- 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.
- McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology*, *55*, 51–62.
- McNamara, D. S., & Kintsch, W. (1996). Learning from texts: effects of prior knowledge and text coherence. *Discourse Processes*, *22*, 247–288.
- Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: effects on retention. *Journal of Educational Psychology*, *74*, 18–22.
- Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*, 437–447.
- Roediger, H. L., III, & Karpicke, J. D. (2006a). The power of testing memory: basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.
- Roediger, H. L., III, & Karpicke, J. D. (2006b). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255.
- Shimmerlik, S. M. (1978). Organization theory and memory for prose: a review of the literature. *Review of Educational Research*, *48*, 103–120.
- Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: questions and answers. *Experimental Psychology*, *56*, 252–257.
- Wenger, S. K., Thompson, C. P., & Bartling, C. A. (1980). Recall facilitates subsequent recognition. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 135–144.
- Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*, 571–580.
- Zaromb, F. M., & Roediger, H. L., III. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory and Cognition*, *38*, 995–1008.