Educ Psychol Rev (2015) 27:247–264
DOI 10.1007/s10648-015-9310-x

REVIEW ARTICLE

Not New, but Nearly Forgotten: the Testing Effect Decreases or even Disappears as the Complexity of Learning Materials Increases

Tamara van Gog[1,2] **· John Sweller**[3]

Published online: 16 May 2015

Abstract The testing effect is a finding from cognitive psychology with relevance for education. It shows that after an initial study period, taking a practice test improves long-term retention compared to not taking a test and—more interestingly—compared to restudying the learning material. Boundary conditions of the effect that have received attention include the test format, retrieval success on the initial test, the retention interval, or the spacing of tests. Another potential boundary condition concerns the complexity of learning materials, that is, the number of interacting information elements a learning task contains. This insight is not new, as research from a century ago already had indicated that the testing effect decreases as the complexity of learning materials increases, but that finding seems to have been nearly forgotten. Studies presented in this special issue suggest that the effect may even disappear when the complexity of learning material is very high. Since many learning tasks in schools are high in element interactivity, a failure to find the effect under these conditions is relevant for education. Therefore, this special issue hopes to put this potential boundary condition back on the radar and provide a starting point for discussion and future research on this topic.

Keywords Testing effect · Retrieval practice · Complex learning tasks · Problem solving · Worked examples

In 1989, Glover published an article entitled *The "testing" phenomenon: Not gone but nearly forgotten*. The testing phenomenon, nowadays known as the testing effect or retrieval practice effect, "…refers to the finding that students who take a test on material between the time they first study and the time they take a final test remember more of the material than students who

✉ Tamara van Gog
t.vangog@uu.nl

[1] Department of Pedagogical and Educational Sciences—Education, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands

[2] Institute of Psychology, Erasmus University Rotterdam, Rotterdam, The Netherlands

[3] School of Education, University of New South Wales, Sydney, Australia

do not take an intervening test." (Glover 1989, p. 392). Glover argued that although multiple laboratory studies had been conducted on the testing effect in the late 1970s and the 1980s, there had been surprisingly little educationally relevant research since Spitzer's (1939) study (the first studies advocating the relevance of testing for education are even older—see Gates 1917). However, as Roediger and Karpicke (2006a) note, this neglect of educationally relevant research on the testing effect continued after Glover's article. The study by Roediger and Karpicke (2006a) as well as their review that appeared in the same year (2006b) marked a change in this situation, rekindling interest in what the testing effect could mean for improving student learning. A myriad of testing effect studies has been published in the last decade (see Rowland 2014, for a meta-analysis). Many of those studies used "educationally relevant materials," and the relevance of the testing effect for classroom learning is increasingly being stressed in the literature (e.g., Agarwal et al. 2012; Fiorella and Mayer 2015; Karpicke and Grimaldi 2012; Roediger et al. 2011b; Roediger and Pyc 2012a).

The reason for the quotes around "educationally relevant materials" is that the vast majority of studies in which such materials are used do not concern the complex learning that is required in many curriculum areas and that is the main goal of most curricula. We define complex learning tasks as tasks that are high in element interactivity (Sweller 2010; Sweller et al. 2011). When learning new information that is *low* in element interactivity, *each individual element can be learned without reference to the other elements in the task*. For example, when learning a list of new Spanish words with their English translation, one item, such as "gato–cat," can be memorized without reference to another item, such as "perro–dog." Or in history, the fact that "World War I began in 1914 and ended in 1918" can be learned without reference to another fact, like "The Netherlands was neutral in World War I." Therefore, such learning materials are not complex, although this does not mean they cannot be difficult for students.

In contrast, complex learning materials such as worked examples that demonstrate how to solve a problem, or instructional texts on scientific phenomena or mechanical systems, are typically *high* in element interactivity, containing *various information elements that are related and must therefore be processed simultaneously in working memory* (Sweller et al. 2011). For instance, when learning about the mechanics of a hydraulic car brake system in engineering, it is not only necessary to learn the individual components in the system (e.g., pistons, cylinders), but also how these components interact with each other (e.g., principles of hydraulic multiplication and friction). Moreover, the aim is usually not just to learn how the system works, but also to be able to apply that knowledge to real-world tasks (e.g., being able to diagnose and repair faults in a system).

The advantage of using element interactivity as a measure of complexity is that it takes into account essential characteristics of human cognitive architecture. Specifically, element interactivity not only considers the complexity of information, but simultaneously considers the knowledge base of individuals. Information that is complex with many interacting elements for a novice will be simple for someone with relevant expertise. For an algebra novice, the problem $a/b=c$, solve for a, along with its solution is likely to consist of about a dozen elements. Solving the problem may impose a heavy working memory load because the elements must be considered simultaneously due to high element interactivity. In contrast, for most readers of this paper, the problem and its solution may constitute no more than a single element. Because of information held in long-term memory, we instantly recognize the problem and know its solution. We are likely to be able to automatically process the problem and its solution in working memory without an excessive working memory load. Thus, the

element interactivity of a particular group of elements will change depending on the learner's level of expertise. It cannot be specified just by reference to the material.

Surprisingly, very little (published) research has investigated whether the testing effect would also apply to learning complex, high element interactivity materials. Responding to critiques and commentaries in which this issue was raised, Roediger and Karpicke (2006b) and Roediger and Pyc (2012b) confirmed the paucity of studies using complex information, but suggested that testing should also be beneficial for these materials. To foreshadow, we will present findings both very recent and a century old, which suggest that the testing effect decreases or even disappears with more complex materials. Table 1 provides an exemplary but a nonexhaustive review of studies that investigated the effects of testing as compared to a restudy control condition, on a delayed final test, that have appeared around or after Roediger and Karpicke's study (2006a) and review (2006b) that sparked renewed interest in the effect. We included only studies that used a restudy control condition, because even though the term testing effect is also used to refer to beneficial effects of testing as an additional activity (e.g., McDaniel et al. 2011; Smith and Karpicke 2014), any beneficial effects of testing as an additional activity could be due to increased time on task rather than the retrieval activity itself. Moreover, we only included studies that reported effects on a delayed final test, because even though some studies have shown testing effects on immediate final tests taking place shortly after the study/restudy or study/test phase (e.g., Carpenter 2009; Carpenter and DeLosh 2006; Carpenter and Kelly 2012; Kang et al. 2011; Verkoeijen et al. 2012), it has been shown that the testing effect often only becomes apparent at a delay of a couple of days or more, with restudy initially, being as, or sometimes even more, effective (e.g., Roediger and Karpicke 2006a; Wheeler et al. 2003).

As can be seen in Table 1, most research was conducted using materials that we estimate are low in element interactivity. On other occasions, when the learning materials seem to have been higher in element interactivity, the tests often did not tap high element interactivity knowledge because they only assessed facts or required recall of terms/ideas. Although we by no means wish to deny that learning low element interactivity material is required in education and as such are educationally relevant (e.g., learning new foreign language vocabulary words in combination with their translation in language class, or learning facts in history, art, or science classes), most educators are likely to place their emphasis on learning higher element interactivity information. Low element interactivity information is used as a prerequisite for learning the more meaningful and complex tasks that we want our students to acquire. For instance, learning lists of vocabulary words in a second language is a necessity for accomplishing the more meaningful and more complex tasks of being able to communicate in that language, understand literature in that language, or read and write business correspondence in that language, which is the ultimate goal of language education. So the question is, would the testing effect also apply to educationally relevant tasks that are closer to the ultimate goals of education, that is, to learning more complex tasks or learning from more complex materials? Table 1 also includes studies presented in this special issue in which this question was addressed, and as can be seen, they suggest that the answer is in the negative.

Studies on the Testing Effect with Complex Learning Materials

In the contribution by van Gog et al. (2015), four experiments are presented on the effects of testing when acquiring problem-solving skills through worked example study. They

Table 1 A representative but nonexhaustive overview of learning materials used in testing effect studies that appeared around/after Roediger and Karpicke's (2006b) review, with our estimate of their element interactivity. "Testing effect" in the last column is defined here as enhanced performance on a delayed test compared to restudy (i.e., excludes experiments comparing testing to no testing or only using an immediate test). Studies in italics are conducted with school-aged children/adolescents at their schools

Author(s)	Learning materials	EI	Testing effect?
Carpenter et al. (2006; exp. 1)	Word pairs + feedback on ini. test (restudy also included feedback time)	L	Yes
Carpenter et al. (2006; exp. 2)	Word pairs + feedback on ini. test (restudy also included feedback time)	L	Yes
Roediger and Karpicke (2006a; exp. 1)	Short text passages incl. many facts (fin. test = idea units)	Text: M-H? Test: L	Yes
Roediger and Karpicke (2006a; exp. 2)	Short text passages incl. many facts (fin. test = idea units)	Text: M-H? Test: L	Yes
Butler and Roediger (2007)	Video lectures (fin. test = facts)	Lect.: M-H? Test: L	Yes for SA ini. test No for MC ini. test
Kang et al. (2007; exp. 1)	Short papers (fin. test = facts)	Papers: H? Test: L	No for SA ini. test No for MC ini. test
Kang et al. (2007; exp. 2)	Short papers + feedback on ini. test (fin. test = facts)	Papers: H? Test: L	Yes for SA ini. test No for MC ini. test
Karpicke and Roediger (2007; exp. 1)	Word lists	L	Yes
Karpicke and Roediger (2007; exp. 2)[a]	Word lists	L	Yes
McDaniel et al. (2007)	Psychology facts	L	Yes
Agarwal et al. (2008; exp. 2)	Text passages (1000 words; fin. test = facts)	Text: M-H? Test: L	Yes[b]
Carpenter et al. (2008; exp. 1)	Obscure facts + feedback on ini. test (restudy also included feedback time)	L	Yes
Carpenter et al. (2008; exp. 2)	Obscure facts + feedback on ini. test (restudy also included feedback time)	L	Yes
Carpenter et al. (2008; exp. 3)	Swahili-English word pairs + feedback on ini. test (restudy also included feedback time)	L	Yes
Karpicke and Roediger (2008)[a]	Foreign vocabulary word pairs	L	Yes

Table 1 (continued)

Author(s)	Learning materials	EI	Testing effect?
Carpenter et al. (2009)	*History facts* + feedback on ini. test (restudy items also presented again)	*L*	*Yes*
Cranney et al. (2009; exp. 2)	Biological psychology video (tests = facts)	Video M/H? Test: L	Yes
Johnson and Mayer (2009)	Short animation (140 s)	M/H?	Yes
Larsen et al. (2009)	Clinical treatments	?[c]	Yes
Toppino and Cohen (2009; exp. 1)	Swahili-English word pairs	L	Yes
Toppino and Cohen (2009; exp. 2)	Swahili-English word pairs + ini. recall boosted: 8 study trials before restudy/testing	L	Yes
Kang (2010; exp. 2)	Chinese character—word pairs	L	Yes
Pyc and Rawson (2010)	Swahili-English word pairs (+ keyword generation) + feedback on ini. test	L	Yes
Tse et al. (2010; exp. 1)[a]	Face-name pairs	L	Adults: yes Older adults: no (rev.)
Tse et al. (2010; exp. 2)[a]	Face-name pairs + feedback on ini. test	L	Adults: yes Older adults: yes
Weinstein et al. (2010; exp. 2)	Text (575 words; fin. test = facts)	Text: M-H? Test: L	Yes
Zaromb and Roediger (2010; exp. 1)	Categorized word lists—randomly presented	L/M	Yes
Zaromb and Roediger (2010; exp. 2)	Categorized word lists—randomly presented	L/M	Yes
Coppens et al. (2011)	Symbol-word pairs	L	Yes
Roediger et al. (2011a; exp. 2)	*Social studies facts (+ feedback on ini. test)*	*L*	*Yes*
Carpenter et al. (2012)	Vocabulary in context of text	M/H?	No (though less forgetting)
McDaniel et al. (2012)	Psychology facts	L	Yes
Pyc and Rawson (2012; exp. 1a and b)	Swahili-English word pairs (+ keyword generation) + feedback on ini. test	L	Yes
Pyc and Rawson (2012; exp. 2)	Swahili-English word pairs (+ keyword generation) + feedback on ini. test	L	Yes

Table 1 (continued)

Author(s)	Learning materials	EI	Testing effect?
van Gog and Kester (2012)	Worked examples (electrical circuits)	H	No
Coane (2013)	Word pairs + feedback on ini. test	L	Yes (older adults)
Putnam and Roediger (2013; exp. 1)	Word pairs	L	No
Putnam and Roediger (2013; exp. 2)	Same word pairs + feedback on ini. test	L	Yes
Smith et al. (2013; exp. 3)	Categorized word lists	M?	No
Smith et al. (2013; exp. 4)	Same word lists but randomly presented (+ prompts to boost ini. test retrieval)	L	Yes
Blunt and Karpicke (2014; exp. 2)	Short texts	Text: M/H? Test: L/M	Yes
Goossens et al. (2014)	*Word pairs (new vocabulary with synonyms)*	*L*	*Yes for cued recall, no for MC final test*
Grimaldi and Karpicke (2014; exp. 1)[a]	Anatomy facts	L	Yes
McDermott et al. (2014; exp. 2)	*Science facts (+ feedback on ini. test)*	*L*	*Yes*
McDermott et al. (2014; exp. 3)	*Science facts (+ feedback on ini. test)*	*L*	*Yes for short answer, no for MC initial test*
Tran et al. (2015; exp. 2)	Scenarios consisting of premises, fin. test: deductive inferences	H	No
Tran et al. (2015; exp. 3)	Scenarios consisting of premises, fin. test: deductive inferences	H	No
Tran et al. (2015; exp. 4)	Scenarios consisting of premises, fin. test: literal premise recall and deductive inferences	H	Yes for literal recall No for inferences
Nestojko and Roediger (2014) cited in Roediger and Nestojko (2015)	Categorized word lists, randomly presented	L/M	No
De Jonge et al. (2015; exp. 1)	Coherent text (1070 words, 60 sentences)	H	No
De Jonge et al. (2015; exp. 2)	60 sentences from exp. 1 scrambled	L/M	Yes ("reduced forgetting")
Leahy et al. (2015; exp. 3)	*Worked examples (electrical circuits)*	*H*	*No*
van Gog et al. (2015; exp. 1)	Worked examples (electrical circuits)	H	No

Table 1 (continued)

Author(s)	Learning materials	EI	Testing effect?
van Gog et al. (2015; exp. 2)	Worked examples (electrical circuits)	H	No
van Gog et al. (2015; exp. 3)	Worked examples (electrical circuits)	H	No
van Gog et al. (2015; exp. 4)	*Worked examples (probability calculation)*	*H*	*No*

ini. test initial test, *fin. test* final test, *EI* element interactivity with L/M/H being low/medium/high, *SA* short answer, *MC* multiple choice

[a] Investigated testing/restudy after item had been learned (i.e., successfully recalled once)

[b] This study involved several types of initial tests and not all were more effective than restudy; more specifically the findings were as follows: closed book + feedback = open book > closed book with no feedback = restudy 2× > restudy 1× > study once

[c] Examples of materials or test questions were not provided in this manuscript

investigated the effects of different kinds of testing after worked example study, such as free recall of the example, solving a problem identical to the example, or solving an isomorphic problem (i.e., different surface features but the same solution procedure) on immediate and delayed performance on a final problem-solving test compared to further example study. The first three experiments were conducted with university students and used the same problems on diagnosing faults in electrical circuits, while the fourth experiment was conducted with vocational education students who learned to solve probability calculation problems. Whereas practice testing in experiments 1 and 2 immediately followed the study of a single example (i.e., relatively low levels of initial learning), in experiments 3 and 4, the practice test was taken after studying four examples (i.e., relatively high levels of initial learning). None of the experiments showed a testing effect.

In three experiments, Leahy et al. (2015) also investigated the testing effect in learning from worked examples. Their participants were elementary school children (an under-investigated population in testing effect research; for exceptions, see, e.g., Bouwmeester and Verkoeijen 2011; Goossens et al. 2014; Karpicke et al. 2014a), who either studied worked examples on reading bus timetables or alternately studied an example and solved a practice test problem. Experiments 1 and 2 showed no testing effect on an immediate test with the group that only studied examples outperforming the example-problem group. However, it is not uncommon to find no differences or an advantage for the restudy condition on an immediate test, but when taking a final test after a delay of several days, the testing condition usually outperforms the restudy condition (see, e.g., Roediger and Karpicke 2006a; Wheeler et al. 2003). Therefore, Leahy et al. conducted a third experiment in which the final test took place 1 week after the learning phase. Although the reversed effect disappeared, experiment 3 showed no testing effect either.

Caution is always warranted when interpreting null effects. But if we include an earlier experiment by van Gog and Kester (2012), then we can conclude that across eight experiments with three different types of problem-solving tasks, three different types of student populations, various types of practice tests, and even when success on the initial test is increased by means of a longer preceding study phase, no evidence was found that the testing effect would also apply when acquiring problem-solving skills through worked example study.[1] One might be tempted to argue that this lack of evidence in favor of a testing effect is not a consequence of the meaningfulness and complexity of the task, but somehow stems from the procedural nature of problem-solving tasks. Whereas word pairs or facts rely on declarative memory, acquiring a problem-solving skill requires moving from initial, declarative encoding via processes of schema construction and automation to procedural memory (e.g., Anderson 1982). However, another study by De Jonge et al. (2015) reported in this special issue, using materials

[1] Note that this is interesting and relevant for research on example-based learning. The studies by Van Gog et al. (2015) and Leahy et al. (2015) were not only inspired by research on the testing effect but also set out to make a relevant contribution to the research on example-based learning. Studies on learning from worked examples often only include immediate final tests, so it is necessary to investigate how well knowledge is retained over time and whether delayed retention could be improved. Moreover, there are only a few prior studies that made direct comparisons between effects on learning of examples only vs. example-problem pairs (Leppink et al. 2014; Van Gog and Kester 2012; Van Gog et al. 2011). Such direct comparisons have been called for in light of the "assistance dilemma," that is, the question of when to provide and when to withhold instructional guidance, or assistance (Koedinger and Aleven 2007). The finding that problem solving following example study, which is more active, does not lead to long-term retention benefits compared to the more passive continued example study is very interesting with regard to this question.

other than worked examples, seems to support the idea that the element interactivity of the materials is key.

In their first experiment, De Jonge et al. (2015) had university students learn a complex and relatively long science text (1070 words and 60 sentences) on black holes. Participants studied this text, which was presented one sentence at a time in the middle of a computer screen, for 15 min (self-paced). Then, half of the participants continued to study the text for another 15 min, while the other half took a fill-in-the-gap test, in which the text was again presented to them but now part of each sentence was omitted and learners had to fill it in. Following the restudy or testing phase, half of the participants in each condition completed the final test (the same fill-in-the-gap test as in the learning phase) after 5 min, whereas the other half completed the final test after 1 week. The results of experiment 1 showed no testing effect. De Jonge et al. decided to conduct a second experiment, in which they used the same material, but now presented the sentences in a scrambled order, which reduced the element interactivity of the material. Interestingly, with the sentences being learned in isolation, without reference to other elements of the text, a benefit of testing was found in experiment 2, albeit only in the sense that testing slowed down the rate of forgetting compared to restudy.

Note that the element interactivity in the learning materials used by De Jonge et al. (2015) resulted largely from the context in which the sentences were presented. Also interesting in this respect is the study by Smith et al. (2013; see Table 1) on learning words in categorized lists. In their third experiment, Smith et al. did not find a testing effect on a delayed test compared to restudy, but in their fourth experiment, they did. They ascribe this to their boosting of the initial retrieval success in experiment 4, by instructing subjects "…to think back to the original study list in order to complete the word stem" (p. 1719). However, almost as an aside, they mention another difference, which is interesting in light of our element interactivity perspective: "Whereas the study list was blocked by category in the three previous experiments, in experiment 4 the order of words within the list was randomized." (p. 1719). In other words, an alternative explanation for the occurrence of a testing effect in their fourth experiment might have been the reduction in element interactivity by presenting words randomly rather than in categorized blocks.

In sum, these studies seem to suggest that the complexity of learning materials may reduce or even eliminate (as several studies in this special issue suggest) the testing effect. Interestingly, this insight is not new, although, as was once true of the testing effect itself (Glover 1989), it seems to have been nearly forgotten.

Not New, but Nearly Forgotten

In his study "Über Einprägung durch Lesen und durch Rezitieren" ("On imprinting through reading and reciting"), that appeared over a century ago, Kühn (1914) reported evidence that the beneficial effects on memory of reciting (i.e., retrieval practice or testing) compared to rereading decreased as the complexity of the learning materials increased. According to a Google Scholar search (December 2014), his work has only been cited in nine other publications. Only two of those are from the present century (the rest is mostly from the 1930s) and neither of the references from the present century mentions the issue of task complexity effects (McDermott and Naaz 2014; Zaromb 2010; but see also De Jonge et al. 2015). While the fact that it was published in German may explain why Kühn's findings have not been widely cited, there must be other reasons for his work being largely ignored, as his study is elaborately

described in Gates (1917). Gates summarized Kühn's findings regarding the different learning materials very clearly: "It appears that the advantage of recitation differs considerably according to the kind of material being studied; the more senseless and less connected the material, the greater the advantage of recitation over reading. Thus, Table IV shows the superiority of recitation to be rather small in the learning of verses, about twice as great for learning series of words, and larger still for learning nonsense syllables" (p. 7).

Moreover, Gates (1917) went on to replicate Kühn's (1914) findings. He also used two different kinds of material in his own studies with schoolchildren and adult learners: "senseless, non-connected material" consisting of nonsense syllables and "connected, sense material in the form of biographies" (p. 24). The biographies were very brief, consisting of a few sentences stating mainly facts, so although higher in element interactivity than nonsense material, element interactivity was not very high. As Gates notes: "While this material is senseful and connected, the organization of different parts of the whole is not so complete and systematic as would be generally found in poetry or prose, in which ideas are more closely related…" (p. 25).

Both for children and adults, his findings showed that for the biographies, the effect of combining reading with recitation was almost twice as large at the retention test that was administered after 3 or 4 h than it was on the test immediately after learning, but the effect was much smaller than for the nonsense syllables. For nonsense material, he also found "the more recitation the better," whereas for sense material, a ratio of 40 % reading and 60 % reciting was best; for older children and adults, a larger proportion of reciting at the expense of reading the biographies produced no further benefits, and for children in the lower grades, it even hindered learning. In his work with adults, Gates (1917) additionally investigated learning of four-letter words (nonconnected sense material) and 20-line stanzas of poetry (connected sense material), but involving just two participants and one participant, respectively, so the results should be interpreted with caution. Nevertheless, the findings did suggest that "the advantage of recitations over reading is greater the more senseless and unconnected the material. Advantage is greatest for nonsense syllables, less great for lists of words, and still less great for connected prose or poetry" (p. 62).

Given that Gates' (1917) work is often referred to in contemporary scientific literature as a classic study on the testing effect, one would expect the fact that the complexity of learning materials substantially reduces the testing effect to have received more attention. Yet, many studies in which Gates is cited do not mention his findings regarding the role of the learning materials in the strength of the testing effect (though there are some exceptions, see for example, the review by Roediger and Karpicke 2006b), and in contemporary empirical research, this issue is largely ignored. When boundary conditions of the testing effect are being discussed, these concern issues such as the test formats, the retrieval success on the initial test or the provision of feedback when retrieval is unsuccessful, or test spacing (Roediger et al. 2010), but not the nature of the learning materials.

With the effect being advocated as educationally relevant, the nature of the learning materials may be an important boundary condition, as it would help teachers and instructional designers to know for which learning materials or learning tasks benefits of testing can and cannot be expected. That is, whereas Kühn (1914) and Gates (1917) showed that the effect decreases as the element interactivity of the material increases, several studies presented in this special issue that used even more complex materials suggest that it may disappear altogether. But why would the testing effect decrease or even disappear as the amount of element interactivity of the learning materials increases?

Why Does the Testing Effect Decrease with Increasing Complexity of Materials?

In order to start answering this question, we have to look at existing explanations for the testing effect with low element interactivity materials. Unfortunately, "although testing effects are empirically well established, *why* testing is beneficial for memory is not well understood. That is, very few studies have systematically evaluated theoretical explanations for the memorial benefits of testing." (Pyc and Rawson 2012, p. 737). Moreover, it seems from Table 1 that the testing effect is not always as clear-cut with low element interactivity learning or test materials either. There is agreement that testing can affect retention both directly through the act of engaging in retrieval itself and indirectly by affecting future study (afforded by feedback or otherwise). Although direct effects of testing have been established (see Roediger and Karpicke 2006b for a review), several studies in Table 1 only found a testing effect when additional measures were taken to enhance retrieval on the initial practice test (e.g., Putnam and Roediger 2013; Tse et al. 2010, for the older adults), and many studies in the table used feedback after testing as a default. Thus, it is possible that feedback is often a prerequisite for finding a testing effect with low element interactivity materials, suggesting that the indirect effect of testing via restudy opportunities may be stronger than direct effects.

Let us begin by looking at two explanations for why testing might be effective. The first focuses mainly on direct effects of testing through free recall, the second mainly on learning of paired associates that could explain both direct and indirect effects of testing. Both emphasize the establishment of associations between elements.

First, free recall tests have been demonstrated to improve retention because of the organization of information that occurs during the recall (e.g., Gates 1917; Tulving 1962; Zaromb 2010; Zaromb and Roediger 2010). This function may be obsolete when learning complex information in which the elements are already highly interrelated. Gates (1917) notes that several of his subjects reported that they grouped the nonsense syllables, and suggests that the fact that this was more easily done in recitation than in (re-)reading accounts for the large beneficial effects of recitation on learning nonsense material. Moreover, he also explains the difference in the effect of recitation between nonsense and sense materials as a function of the inherent level of organization, when he concludes: "In short, recitation rendered great service in creating usable associations within the material where there was none, or in more adequately noticing and exercising those that were already present. In nonsense material these bonds between items are absent, and this process of organization and creation of associations is difficult and essential; learning of such material exists in accomplishing just this organization. In the connected sense material such as that used in the present experiment, most of these associations are already present; the material is already organized… The function of recitation for the formation of these bonds is not required. What is needed is that the ready-formed associations be noticed and exercised, although, in most cases, bonds in addition to those found in the material will be required." (p. 97). This conclusion seems in line with the findings by Masson and McDaniel (1981) showing that awareness of relationships between items during studying also fosters performance on a delayed retention test, presumably because of the relations that are formed during encoding, as a consequence of which "the retrieval of a particular item is facilitated by retrievability of other items to which it is linked." (p. 109). Moreover, this conclusion seems to align with the findings by Bouwmeester and Verkoeijen (2011) who showed that children who formed strong gist traces during study or restudy benefited less from retrieval practice.

Second, when learning paired associates, it has been suggested that testing may strengthen memory traces (although there is debate about the mechanisms by which this process occurs; see, e.g., Carpenter 2009; Karpicke et al. 2014b; Kornell et al. 2015; Pyc and Rawson 2009, 2010, 2012). Memory traces may be strengthened either directly when retrieval is successful or indirectly through feedback. Without feedback, initial (ir)retrievability has been shown to determine whether a testing effect occurs at a delayed test compared to restudy, with a testing effect appearing only for retrievable but not for irretrievable items (Jang et al. 2012). Feedback has been shown to primarily improve delayed recall of initially irretrievable items but has little effect on initially retrieved items (Pashler et al. 2005). This result is in line with other studies showing that success of the retrieval attempt is not important, as long as correct answer feedback (e.g., Kornell et al. 2009) or a subsequent study opportunity (e.g., Richland et al. 2009) is provided.

Kornell et al. (2015) recently proposed a two-stage framework to explain these findings. Stage 1 constitutes the retrieval attempt during which related information (mediators) are activated. Stage 2 is the period in which the answer is available either through successful retrieval or feedback, at which point links between the cue, the mediators, and the target can be strengthened. Kornell et al. tested whether it matters if stage 2 is initiated via successful retrieval or feedback and (although somewhat mixed) their findings suggest that it does not. Thus, it seems that this model can account for both direct and indirect effects of testing on paired associate learning.

We have to admit that it is possible that the results of the studies presented in this issue may have been different if feedback or restudy opportunities after testing had been provided (although van Gog et al. 2015 did increase the level of success on the initial practice test through a longer study phase in experiments 3 and 4, yet still failed to find a testing effect). However, even though the effects of feedback or restudy opportunities on learning more complex tasks should definitely be addressed in future studies, it should be noted that this second explanation provided by the two-stage framework of Kornell et al. (2015) again hinges on the associations between elements established through testing, which is less necessary when such associations are inherent to the material or if the nature of the material is such that learners are likely to search for associations when reading the material in order to understand it. If the benefit of retrieval practice for low element interactivity materials lies mainly in the elaborative processes that establish relations between information elements, as both the "organization" and the "mediator strengthening" explanations seem to imply, this might explain why retrieval practice is less effective when such relations are inherent to the material or encourage search for associations when reading.

The other side of the coin is that for complex materials that contain many interacting information elements, restudy may be more effective than for low element interactivity materials. Most studies on the testing effect have investigated the benefits of taking practice tests with materials for which almost literal recall is required. Although it has been suggested that testing can also boost performance on transfer questions that require inferences (for a review, see Carpenter 2012), it seems that these studies did not always include a restudy control condition, so it is unclear whether that transfer effect would be maintained compared to restudy. Moreover, it is convincingly argued by Tran et al. (2015) that in many studies, the transfer questions either required very limited transfer (e.g., recalling a word in a sentence that differed from the word that had to be recalled from that same sentence on an initial test) or contained cues that were also present in the original text (i.e., corresponding words or terms) that would invite the correct answer (e.g., Butler 2010). Tran et al. set out to investigate what

would happen on transfer items that truly required deductive inferences, having students study scenarios consisting of premises and then engage in restudy or retrieval practice, after which, on the final test, they were required to answer multiple choice transfer questions that required deductive inferences from the premises. Across four experiments (of which three included a delayed test), they found no evidence that having engaged in retrieval practice improved the ability to make inferences, although their fourth experiment did show a testing effect with regard to the premises.

Thus, learners who had received practice tests were able to recall more premises (i.e., literal recall of items, low element interactivity) than students who had restudied, but this did not help them in making inferences (i.e., processing premises in relation to each other, high element interactivity). The explanation of Tran et al. (2015) for these findings is in line with this element interactivity perspective: "For the retrieval practice condition, participants must actively recall multiple premises and check whether the premises recalled are relevant to the presented inference question (i.e., item-specific processing). By contrast, for participants in the rereading condition, the lack of such demands may have allowed them time and resources to attend "online" to the relationships between premises (i.e., relational processing). This cognitive work in the learning phase may have paid special dividends in the inference task, but not in the explicit memory task." (p. 140).

While restudying complex materials may pay off on test tasks that require elements to have been processed in relation to each other, there may also be differences with regard to students' motivation and metacognitive feelings about their own learning between low and high element interactivity materials. It has been shown that feelings of difficulty influence estimates of effort and time to be spent on learning tasks and that feelings of familiarity are negatively related to feelings of difficulty (Efklides 2006). With low element interactivity material, when students are confronted with the exact items that they have just studied, they might have the feeling that these materials are already known from the initial study session, and may not be motivated to spend time and effort on further study (see, e.g., Metcalfe and Kornell 2003, for evidence that students will drop items they feel they have learned from further study when given a chance).

The idea that students might feel high familiarity and low difficulty when engaging in restudy of low element interactivity materials seems in line with the findings on students' metacognitive judgments. For instance, Roediger and Karpicke (2006a) found that students in the restudy condition rated the material as less interesting and were more confident in their ability to remember the materials than students in the testing conditions (the latter was also found by Agarwal et al. 2008). Although these findings might also simply reflect students' confidence in restudy as an effective learning method (see, e.g., Karpicke et al. 2009; Kornell and Son 2009), it is also likely that they underscore students' overconfidence after restudy due to feelings of high familiarity or low difficulty. Taking a practice test on the other hand will soon make it clear to students that they have not sufficiently learned the material, so when given another study opportunity after the practice test, they may pay more attention and learn new information they were previously unable to retrieve (cf. findings on "test-potentiated learning," Arnold and McDermott 2013a, b; see also the two-stage framework of Kornell et al. 2015).

With high element interactivity material, which requires making connections among various elements in order to be fully understood, it is much less likely that students would feel the material is familiar after an initial study phase, and they might spend more time (if that is under their control) and effort on restudy. In other words, perhaps the testing effect is not so much a consequence of beneficial processes occurring during testing, as it is of suboptimal

processes during restudy. Kang and Pashler (2014) recently attempted to address this issue. Instead of varying the complexity of the learning materials, however, they used common low element interactivity material (Swahili-English word pairs) and experimentally manipulated the incentives (monetary bonuses or time savings), which in turn was assumed to affect motivation. According to the authors, the fact that they still found a testing effect provides "some reassurance that lab findings from the testing effects literature likely generalize to real-world situations in which motivation to learn may be greater" (p. 183).

That may be true, but the authors used monetary or time incentives as a proxy for (extrinsic) motivation, with a low element interactivity task. They did not actually assess feelings of familiarity and difficulty, motivation to invest effort, or actual effort investment as a function of complexity of the materials. More direct measures of such affective variables would seem necessary in order to fully exclude the possibility that motivation and/or metacognitive monitoring may play a role in the reduction or disappearance of the testing effect with high element interactivity materials. The findings by van Gog et al. (2015) with regard to invested mental effort are interesting in light of the motivation discussion. The same materials were used in experiments 1, 2, and 3, and participants in these experiments were university students. In experiments 1 and 2, mental effort investment seemed somewhat lower in the restudy (examples only) condition, but did not differ significantly between the restudy and testing (example-problem pair) conditions (this finding replicates results from the same conditions in van Gog et al. 2011). So after having studied just one example, participants spent only a little (and not significantly) less effort on studying another (experiment 1) or the same (experiment 2) example than on solving a practice problem. However, in experiment 3, when participants had to engage in restudy or testing after having studied *four* examples, much less effort was invested in restudy than in testing. This suggests that indeed feelings of familiarity or perceived/actual task difficulty (i.e., the task becomes easier as learning progresses) may play a role in the effort participants have to or want to invest in restudy. Future research should attempt to replicate this finding, however, as the results from experiment 4, with a different participant population (i.e., substantially lower level of education) than the other three conditions, did not show the same result as experiment 3.

In sum, it seems that the complexity of learning materials reduces or eliminates the testing effect either by reducing the effectiveness of testing relative to restudy because the "organizational/relational processing" afforded by testing no longer presents a benefit or by increasing the effectiveness of restudy relative to testing, by affording processing of relations between elements and maintaining student motivation for engaging in restudy. Perhaps it is time to develop a "material appropriate processing" (McDaniel and Einstein 1989) framework for the testing effect. Such a framework could help specify for which learning materials and test tasks testing or restudy can be expected to be most effective, where "Significant enhancement in recall is anticipated only to the extent that the … study strategy encourages processing that is complementary to the processing invited by the material itself." (McDaniel and Einstein 1989, p. 113).

Conclusion

The studies collected in this special issue suggest that the complexity of learning materials might constitute another boundary condition of the testing effect, by showing that the effect decreases or even disappears with complex learning tasks that are high in element interactivity, which are plentiful in education. Of course, each of the studies presented here has its own

limitations, and in some of these studies, other known boundary conditions of the testing effect, such as test format, retrieval success on the initial test or feedback when retrieval is unsuccessful, or test spacing (Roediger et al. 2010), cannot be ruled out entirely as an alternative explanation. Nevertheless, taken together and in light of the nearly forgotten findings by Kühn (1914) and Gates (1917), the studies presented in this special issue suggest that it would be worthwhile to conduct further research on the complexity of learning and test tasks as a potential boundary condition of the testing effect. It would help teachers and instructional designers to know for which learning tasks they can and cannot expect benefits of having their students take practice tests instead of engage in further study. Therefore, we hope that this special issue encourages debate about and future research on the question of how the complexity of learning materials affects the testing effect.

Acknowledgments Tamara van Gog was supported by a Vidi grant (# 452-11-006) from the Netherlands Organisation for Scientific Research. The guest editors would like to thank all reviewers for their thoughtful comments on the manuscripts that were submitted for publication in this special issue, Peter Verkoeijen for commenting on a draft of this introductory review, and Jeffrey Karpicke, William Aue, and Katherine Rawson for their commentaries on the contributions to this special issue.

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.

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The value of applied research: retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review, 24*, 437–448. doi:10.1007/s10648-012-9210-2.

Anderson, J. R. (1982). Acquisition of a cognitive skill. *Psychological Review, 89*, 369–406. doi:10.1037/0033-295X.89.4.369.

Arnold, K. M., & McDermott, K. B. (2013a). Free recall enhances subsequent learning. *Psychonomic Bulletin & Review, 20*, 507–513. doi:10.3758/s13423-012-0370-3.

Arnold, K. M., & McDermott, K. B. (2013b). Test-potentiated learning: distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 940–945. doi:10.1037/a0029199.

Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology, 106*, 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.

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

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.

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. (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. (2006). Impoverished cue support enhances subsequent retention: support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*, 268–276. doi:10.3758/BF03193405.

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

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

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

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

Coane, J. H. (2013). Retrieval practice and elaborative encoding benefit memory in younger and older adults. *Journal of Applied Research in Memory and Cognition, 2*, 95–100. doi:10.1016/j.jarmac.2013.04.001.

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

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.

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.* doi:10.1007/s10648-015-9300-z.

Efklides, A. (2006). Metacognition and affect: what can metacognitive experiences tell us about the learning process? *Educational Research Review, 1*, 3–14. doi:10.1016/j.edurev.2005.11.001.

Fiorella, L., & Mayer, R. E. (2015). *Learning as a generative activity: eight learning strategies that promote understanding*. New York: Cambridge University Press.

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

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

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., & Tabbers, H. K. (2014). The effect of retrieval practice in primary school vocabulary learning. *Applied Cognitive Psychology, 28*, 135–142. doi:10.1002/acp.2956.

Grimaldi, P. J., & Karpicke, J. D. (2014). Guided retrieval practice of educational materials using automated scoring. *Journal of Educational Psychology, 106*, 58–68. doi:10.1037/a0033208.

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

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101*, 621–629. doi:10.1037/a0015183.

Kang, S. H. K. (2010). Enhancing visuo-spatial learning: the benefit of retrieval practice. *Memory & Cognition, 38*, 1009–1017. doi:10.3758/MC.38.8.1009.

Kang, S. H. K., & Pashler, H. (2014). Is the benefit of retrieval practice modulated by motivation? *Journal of Applied Research in Memory and Cognition, 3*, 183–188. doi:10.1016/j.jarmac.2014.05.006.

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

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

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

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*, 151–162. doi:10.1016/j.jml.2006.09.004.

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

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

Karpicke, J. D., Blunt, J. R., Smith, M. A., & Karpicke, S. S. (2014a). Retrieval-based learning: the need for guided retrieval in elementary school children. *Journal of Applied Research in Memory and Cognition, 3*, 198–206. doi:10.1016/j.jarmac.2014.07.008.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014b). Retrieval-based learning: an episodic context account. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 61, pp. 237–284). San Diego: Elsevier Academic. doi:10.1016/B978-0-12-800283-4.00007-1.

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review, 19*, 239–264. doi:10.1007/s10648-007-9049-0.

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

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.

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

Kühn, A. (1914). Über Einprägung durch Lesen und durch Rezitieren [On imprinting through reading and reciting]. *Zeitschrift für Psychologie, 68*, 396–481. Available from: http://babel.hathitrust.org/cgi/pt?id=njp.32101074935824;view=1up;seq=422.

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.

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.* doi:10.1007/s10648-015-9296-4.

Leppink, J., Paas, F., van Gog, T., van der Vleuten, C., & van Merriënboer, J. J. G. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction, 30*, 32–42. doi:10.1016/j.learninstruc.2013.12.001.

Masson, M. E., & McDaniel, M. A. (1981). The role of organizational processes in long-term retention. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 100–110. doi:10.1037/0278-7393.7.2.100.

McDaniel, M. A., & Einstein, G. O. (1989). Material-appropriate processing: a contextualist approach to reading and studying strategies. *Educational Psychology Review, 1*, 112–145. doi:10.1007/BF01326639.

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., 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., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: an experimental study. *Journal of Applied Research in Memory and Cognition, 1*, 18–26. doi:10.1016/j.jarmac.2011.10.001.

McDermott, K. B., & Naaz, F. (2014). Is recitation an effective tool for adult learners? *Journal of Applied Research in Memory and Cognition, 3*, 207–213. doi:10.1016/j.jarmac.2014.06.006.

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.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General, 132*, 530–542. doi:10.1037/0096-3445.132.4.530.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 3–8. doi:10.1037/0278-7393.31.1.3.

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

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*, 437–447. doi:10.1016/j.jml.2009.01.004.

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

Pyc, M. A., & Rawson, K. A. (2012). Why is retrieval practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 737–746. doi:10.1037/a0026166.

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.

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

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

Roediger, H. L., & Nestojko, J. F. (2015). The relative benefits of studying and testing on long-term retention. In J. G. W. Raaijmakers, A. H. Criss, R. L. Goldstone, R. M. Nosofsky, & M. Styvers (Eds.), *Cognitive modeling in perception and memory: a festschrift for Richard M. Shiffrin* (pp. 99–111). New York: Psychology.

Roediger, H. L., & Pyc, M. A. (2012a). Inexpensive techniques to improve education: applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition, 1*, 242–248. doi:10.1016/j.jarmac.2012.10.006.

Roediger, H. L., & Pyc, M. A. (2012b). Applying cognitive psychology to education: complexities and prospects. *Journal of Applied Research in Memory and Cognition, 1*, 263–265. doi:10.1016/j.jarmac.2012.10.006.

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

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011a). 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., Putnam, A. L., & Smith, M. A. (2011b). Ten benefits of testing and their applications to educational practice. In J. Mestre & B. Ross (Eds.), *Psychology of learning and motivation: cognition in education* (pp. 1–36). Oxford: Elsevier.

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.

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory, 22*, 784–802. doi:10.1080/09658211.2013.831454.

Smith, M. A., Roediger, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1712–1725. doi:10.1037/a0033569.

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

Sweller, J. (2010). Element interactivity and intrinsic, extraneous and germane cognitive load. *Educational Psychology Review, 22*, 123–138. doi:10.1007/s10648-010-9128-5.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York: Springer.

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

Tran, R., Rohrer, D., & Pashler, H. (2015). Retrieval practice: the lack of transfer to deductive inferences. *Psychonomic Bulletin and Review, 22*, 135–140. doi:10.3758/s13423-014-0646-x.

Tse, C. S., Balota, D. A., & Roediger, H. L. (2010). The benefits and costs of repeated testing: on the learning of face-name pairs in older adults. *Psychology and Aging, 25*, 833–845. doi:10.1037/a0019933.

Tulving, E. (1962). Subjective organization and effects of repetition in multi-trial free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 5*, 193–197. doi:10.1016/S0022-5371(66)80016-6.

van Gog, T., & Kester, L. (2012). A test of the testing effect: acquiring problem-solving skills from worked examples. *Cognitive Science, 36*, 1532–1541. doi:10.1111/cogs.12002.

van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices' learning. *Contemporary Educational Psychology, 36*, 212–218. doi:10.1016/j.cedpsych.2010.10.004.

van Gog, T., Kester, L., Dirkx, K., Hoogerheide, V., Boerboom, J., & Verkoeijen, P. P. J. L. (2015). Testing after worked example study does not enhance delayed problem-solving performance compared to restudy. *Educational Psychology Review*. doi:10.1007/s10648-015-9297-3

Verkoeijen, P. P. J. L., Bouwmeester, S., & Camp, G. (2012). A short term testing effect in cross-language recognition. *Psychological Science, 23*(6), 567–571. doi:10.1177/0956797611435132.

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

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

Zaromb, F. M. (2010). *Organizational processes contribute to the testing effect in free recall*. Doctoral dissertation, Washington University in St. Louis.

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