Educ Psychol Rev (2015) 27:317–326
DOI 10.1007/s10648-015-9309-3

COMMENTARY

The Testing Effect Is Alive and Well with Complex Materials

Jeffrey D. Karpicke[1] **· William R. Aue**[1]

Published online: 14 May 2015

Abstract Van Gog and Sweller (2015) claim that there is no testing effect—no benefit of practicing retrieval—for complex materials. We show that this claim is incorrect on several grounds. First, Van Gog and Sweller's idea of "element interactivity" is not defined in a quantitative, measurable way. As a consequence, the idea is applied inconsistently in their literature review. Second, none of the experiments on retrieval practice with worked-example materials manipulated element interactivity. Third, Van Gog and Sweller's literature review omitted several studies that have shown retrieval practice effects with complex materials, including studies that directly manipulated the complexity of the materials. Fourth, the experiments that did not show retrieval practice effects, which were emphasized by Van Gog and Sweller, either involved retrieval of isolated words in individual sentences or required immediate, massed retrieval practice. The experiments failed to observe retrieval practice effects because of the retrieval tasks, not because of the complexity of the materials. Finally, even though the worked-example experiments emphasized by Van Gog and Sweller have methodological problems, they do not show strong evidence favoring the null. Instead, the data provide evidence that there is indeed a small positive effect of retrieval practice with worked examples. Retrieval practice remains an effective way to improve meaningful learning of complex materials.

Keywords Testing effect · Retrieval practice · Worked examples · Complex learning · Element interactivity

Over the past decade, there has been robust interest in the effects of retrieval practice on learning, with a special emphasis on how best to apply the benefits of retrieval (or testing) to the complex tasks, materials, and assessments found in educational settings. The consistent

✉ Jeffrey D. Karpicke
karpicke@purdue.edu

[1] Department of Psychological Sciences, Purdue University, 703 3rd Street, West Lafayette, IN 47907-2081, USA

finding from recent research has been that retrieval practice promotes meaningful learning of complex materials (Carpenter 2012; Dunlosky et al. 2013; Karpicke 2012; Roediger and Pyc 2012). In this issue, Van Gog and Sweller claim that there is no testing effect for complex materials and that this represents a "boundary condition" on the effect. This is a dangerous claim because it may mislead educators to think that retrieval practice is not effective for learning complex educational materials when in fact a wealth of research has shown that it is.

The reasoning behind the claim is flawed. Whether educational materials are simple or complex is orthogonal to whether retrieval practice enhances learning. To be able to retrieve, use, and apply knowledge in the long term, it is highly effective to practice retrieving, using, and applying knowledge during learning. Consider a similar scenario: When a student wants to learn to play a piece of music on the piano, he or she practices playing the piano, rather than merely reading the sheet music or reading a book about how to play the piece. Van Gog and Sweller's claim is akin to saying that practicing the piano works only for simple pieces, but to learn to play a complex piece of music, practicing does not work and students should not bother doing it. The reasoning simply makes no sense.

Van Gog and Sweller's analysis is questionable as well, and the following sections describe specific problems with their analysis and emphasize research showing benefits of retrieval practice for learning complex materials.

"Complexity" and "Element Interactivity" Are Poorly Defined

Van Gog and Sweller's analysis is ambiguous in part because it confuses the complexity of the materials, complexity of the initial learning activity, and complexity of the criterial assessment. Van Gog and Sweller define complex materials as those that are high in element interactivity. Material that is "high" in element interactivity contains elements or ideas that are related, so that the learning of some ideas depends on learning other ideas in the material. Material that is "low" in element interactivity contains items that can be learned in isolation, without reference to other items or ideas in the materials (paraphrasing Van Gog and Sweller 2015). Element interactivity is certainly an important aspect of educational materials that should be examined in a thorough analysis of the literature and rigorous experiments (indeed, that sentiment is not new: see McDaniel and Einstein 1989). Unfortunately, that was not done in the present issue.

The central problem is that Van Gog and Sweller never offer a quantitative metric for measuring element interactivity. The analysis of previous research is completely subjective, and without an objective measure, the idea of element interactivity can be applied on the fly to suit the authors' needs. A wide variety of measures exist to assess several dimensions of educational materials including, for example, Latent Semantic Analysis[1] (Foltz et al. 1998) which was used by de Jonge et al. (2015), and Coh-Metrix[2] (Graesser et al. 2004; Graesser et al. 2011) which provides more advanced measures of the cohesiveness of materials. No measure was used to define and assess element interactivity; instead, the critical idea in Van Gog and Sweller's analysis of the retrieval practice literature was evaluated purely subjectively.

Several strange things happen without an objective measure of complexity or element interactivity. The only materials that Van Gog and Sweller deemed high in element

[1] http://lsa.colorado.edu
[2] http://cohmetrix.com/

interactivity were their own worked-example experiments (Leahy et al., 2015; Van Gog and Kester 2012; and Van Gog et al., 2015), a paper by Tran et al. (2015), and de Jonge et al. (2015).[3] de Jonge et al. had students study a 1000-word text on black holes. For a retrieval practice task, students filled in missing words in individual isolated sentences. Van Gog and Sweller rated this as high element interactivity; they did not specify whether this rating applied to the materials, the retrieval activity, or both. Like de Jonge et al., Tran et al. (2015) had students study a set of seven to nine sentences (e.g., "Students commute from off-campus housing to campus by any of 3 routes") and practice retrieval by filling in words missing from the sentences ("Students commute from off-campus housing to campus by any of ___ routes"). The retrieval task involved recall of isolated words within individual facts. Nevertheless, these materials and activities received a high rating from Van Gog and Sweller.

The remaining experiments in which students read text materials (or watched videos) and practiced retrieval in various ways (e.g., by answering conceptual questions) were deemed "medium" element interactivity by Van Gog and Sweller. For example, Butler and Roediger (2007) used a 30-min videotaped classroom lecture on art history and had students answer short-answer questions. Van Gog and Sweller rated their materials as "Medium/High?" (question mark in original) and their short-answer activity as low element interactivity. All the other studies that used educational texts were rated as medium/high, often with a question mark (e.g., Agarwal et al. 2008; Blunt and Karpicke 2014; Johnson and Mayer 2009; Roediger and Karpicke 2006; Weinstein et al. 2010).

Van Gog and Sweller wrote that "instructional texts on scientific phenomena or mechanical systems are typically high in element interactivity" (2015), yet their analysis discounted or excluded essentially all the previous research with such materials. We wondered whether there were measurable differences among the materials deemed high complexity and those deemed medium complexity by Van Gog and Sweller. Table 1 shows several measures of the materials used in experiments highlighted by Van Gog and Sweller (and some experiments excluded from their analysis), including word length, Flesch Reading Ease, Flesch-Kincaid Grade Level, and a measure of referential cohesion from Coh-Metrix. Referential cohesion is the degree to which ideas within a text overlap and are connected across sentences (see Graesser et al. 2011); it provides a possible measure that may capture element interactivity within a text. Table 1 shows that the text used by de Jonge et al. was relatively high in referential cohesion. Two of the brief scenarios used by Tran et al. (2015) exhibited very high referential cohesion and the other two were reasonably high. Nevertheless, several experiments demonstrating retrieval practice effects have employed materials with relatively high referential cohesion; for example, materials used by Hinze and Wiley (2011), Johnson and Mayer (2009), and Roediger and Karpicke (2006) all scored above the 70th percentile for referential cohesion. Some experiments used materials with referential cohesion scores that were as high or higher than the de Jonge et al. and Tran et al. materials (namely, Karpicke and Blunt 2011, and McDaniel et al. 2009). All of the experiments displayed in Table 1, except de Jonge et al. and Tran et al., showed retrieval practice effects. Clearly, retrieval practice enhances learning for both low- and high-complexity materials.

Van Gog and Sweller's analysis of the complexity of retrieval practice tasks is perhaps even more bizarre than the analysis of materials. Van Gog and Sweller consider freely recalling (e.g., Blunt and Karpicke 2014; Roediger and Karpicke 2006) or summarizing (e.g., Johnson

[3] Kang et al. (2007) used 3000-word articles from *Current Directions in Psychological Science*. These earned a "High?" rating, with the question mark in the original.

Table 1 Measures of text characteristics from retrieval practice experiments

Paper	Text	Word count	Flesch Reading Ease	Flesch Grade Level	Referential Coherence	
					z-score	Percentile
Roediger and Karpicke (2006)	The Sun	258	77.3	6.7	0.78	78
	Sea Otters	274	54.6	9.4	−0.99	16
Agarwal et al. (2008)	Average of 6 texts	1005	78.3	5.5	−0.67	27
Johnson and Mayer (2009)	Lightning Storms	522	60.9	9.5	0.58	72
McDaniel et al. (2009)	Brakes	982	68.4	7.4	2.13	98
	Pumps	873	63.9	8.2	1.52	93
Weinstein et al. (2010)	Salvador Dali	565	50.6	11.6	−0.66	26
	The KGB	574	25.7	15.2	0.31	62
	Venice	541	46.5	11.9	−0.55	29
	The Taj Majal	605	48.9	12.2	−0.06	48
Hinze and Wiley (2011)	Cell Division	427	57.0	8.8	0.64	74
Karpicke and Blunt (2011)	The Human Ear (sequential)	260	68.8	7.6	1.64	95
	The Digestive System (sequential)	268	60.7	8.6	1.67	95
	Make-up of Human Blood (enumeration)	235	63.7	7.6	1.05	85
	Kinds of Muscle Tissue (enumeration)	248	56.7	8.7	0.90	82
Blunt and Karpicke (2014)	Enzymes	283	37.6	11.7	0.10	54
	Domains of Life	282	47.6	10.1	−0.51	31
de Jonge et al. (2015)	Black Holes	1066	62.8	8.8	1.20	88
Tran et al. (2015)	Scenario 1	60	82.5	3.8	1.03	85
	Scenario 2	100	83.3	4.7	2.33	99
	Scenario 3	91	75.7	5.2	0.83	79
	Scenario 4	90	63.8	7.2	2.34	99

Experiments were selected based on Van Gog and Sweller's (2015) analysis. McDaniel et al. (2009), Hinze and Wiley (2011), and Karpicke and Blunt (2011) were not included in their original analysis. Referential cohesion *z*-scores and percentile ranks were obtained from Coh-Metrix (www.cohmetrix.com)

and Mayer 2009; Weinstein et al. 2010) to be low element interactivity retrieval tasks. To freely recall or create a summary, a learner must rely on a mental model of how material is organized and use this relational knowledge structure as a plan to guide retrieval. Free recall and summarization require high degrees of element interactivity by their very nature.

Van Gog and Sweller also rated experiments in which students answered conceptual short-answer questions as low element interactivity (Agarwal et al. 2008; Blunt and Karpicke 2014; Johnson and Mayer 2009; Kang et al. 2007; Weinstein et al. 2010). For example, Agarwal et al. (2008) had students read 1000-word texts (e.g., one was about the Voyager spacecraft) and answer short-answer questions requiring them to make inferences and explain concepts (e.g., "Why did the Voyager have instruments that would measure ultraviolet and infrared light?"). Agarwal et al.'s retrieval task was rated low by Van Gog and Sweller. Similarly,

Johnson and Mayer (2009) had students study a notoriously difficult set of materials on how lightning storms develop, write a summary explanation of the materials, and answer inferential questions (e.g., "What could you do to decrease the intensity of lightning?"). Van Gog and Sweller gave them a "Medium/High?" (question mark in original) and did not distinguish whether this referred to the materials, the retrieval tasks, or both.

Blunt and Karpicke (2014) had subjects freely recall or create concept maps as retrieval practice tasks. Concept mapping involves creating diagrams in which students identify the individual elements within a set of material, place those elements in nodes, draw links to connect related nodes in a network, and write descriptions along the links to specify how the elements are related. If anything, concept mapping would seem to be a quintessential method for promoting and assessing the processing of element interactivity within a set of materials. Van Gog and Sweller rated the task as low/medium element interactivity.

To recap, tasks in which students filled in individual words in isolated sentences were rated as high in element interactivity, while tasks where students freely recalled, produced summaries, answered inferential short-answer questions, or created concept maps were deemed low or, at best, medium in element interactivity. The rating scheme appears to be completely backwards. Even without a quantitative measure of element interactivity during retrieval, it is clear that filling in individual words in sentences requires little or no integration across ideas, while all the other retrieval activities described here require significantly more organizational and relational processing. We will return to this point because the nature of the retrieval tasks is ultimately the key reason why some of the present studies failed to observe retrieval practice effects.

None of the Experiments Manipulated Complexity or Element Interactivity

There is an even more serious limitation in the research summarized by Van Gog and Sweller: Element interactivity was not manipulated in any of the worked-example experiments reported by Leahy et al. (2015), Van Gog and Kester (2012), or Van Gog et al. (2015). None of the experiments showed that, holding everything else constant, the testing effect exists for simple (low element interactivity) materials but disappears for complex (high element interactivity) materials.

The case can be made that de Jong et al. (2015) manipulated element interactivity across experiments. In Experiment 1, de Jonge et al. presented the materials intact (deemed high element interactivity by Van Gog and Sweller), while in Experiment 2, they presented the materials as a series of randomly ordered—but still clearly interrelated—facts. This is exactly what Chan (2009) did in a series of experiments (see too Chan 2010, and Chan et al. 2006; none of these papers was mentioned in the present special issue). Chan (2009) had students read lengthy texts intact or with the sentences randomly ordered, which he referred to as high and low integration conditions, respectively. The students then answered short-answer questions that required them to relate multiple concepts within the texts. Chan observed robust benefits of retrieval practice on delayed tests 1 day after the initial learning phase for both the low and high integration conditions.

Karpicke and Blunt (2011) also directly manipulated the type of materials that learners studied and practiced retrieving; again, a discussion of this fact is absent from Van Gog and Sweller's analysis. Karpicke and Blunt had students read texts with enumeration structures, which listed a series of facts and concepts about a topic, and texts with sequential structures, which described a connected series of events and steps

in a process (see too Cook and Mayer 1988; Meyer 1975). Sequential texts are likely higher in element interactivity than are enumeration texts—the measures of referential coherence in Table 1 support this claim. Karpicke and Blunt showed large benefits of retrieval practice on long-term retention for both types of text (see their Figure 2). It is also worth noting that Karpicke and Blunt used concept mapping as a final assessment of learning; thus, robust benefits of retrieval practice were evident on final assessments that explicitly required students to specify the interactions among elements.

Existing Research Has Shown Retrieval Practice Effects with Complex Materials

In addition to experiments that directly manipulated the complexity of the materials, several studies have shown retrieval practice effects with complex materials, often carried out in authentic educational settings. Many of these studies were excluded from Van Gog and Sweller's analysis.

A surprising omission is McDaniel et al. (2009). They used complex materials (included in the analysis in Table 1) about the workings of mechanical systems (brakes and pumps) and showed benefits of retrieval practice on delayed assessments that measured recall and the ability to apply knowledge and solve new problems. Chan's research (Chan 2009, 2010; Chan et al. 2006), which was also not discussed by Van Gog and Sweller, showed that retrieval practice enhanced long-term retention in low and high text-integration conditions. He also showed that practicing retrieval of a portion of complex material can spread to and enhance long-term retention of portions that were not explicitly tested, a phenomenon called retrieval-induced facilitation. Butler's (2010) results, which showed that retrieval practice enhanced transfer of knowledge with questions that required learners to integrate multiple concepts, were discounted by Van Gog and Sweller for unclear reasons. Several studies have shown that retrieval practice enhances learning of spatial information, such as the locations and relationships among objects on maps or diagrams (e.g., Carpenter and Kelly 2012; Rohrer et al. 2010). The task of retrieving spatial relations seems high in element interactivity, yet again these findings were excluded from Van Gog and Sweller's analysis.

A wealth of recent research has extended the benefits of retrieval practice to classroom learning. Several studies, carried out in authentic classroom settings, have shown that retrieval practice improves student learning of the materials studied in school, using educationally relevant retrieval activities and assessments (e.g., Agarwal et al. 2012; Butler et al. 2014; Dobson and Linderholm (2015); Jensen et al. 2014; Lyle and Crawford, 2011; McDaniel et al. 2007a, b, 2013; McDermott et al. 2014; Roediger et al. 2011). In one striking example, Larsen et al. (2013) had medical students practice retrieval of clinical knowledge (e.g., the symptoms that would be diagnostic of particular disorders). Six months after initial learning, practicing retrieval improved the medical students' performance at forming diagnoses in a simulated patient scenario. To us, this unquestionably represents complex learning of complex materials.

Given the evidence, Van Gog and Sweller's claim that retrieval practice does not enhance learning of complex materials is jarring. Indeed, the fact that retrieval practice enhances learning in "educationally relevant tasks that are closer to the ultimate goal of education" (2015) has already been affirmed.

The Worked-Example Data Are Ambiguous but Tend to Show a Positive Effect of Retrieval Practice

Based on the evidence reviewed so far, Van Gog and Sweller's central claim that retrieval practice does not enhance learning of complex materials is incorrect. The overwhelming evidence shows that retrieval practice is effective for materials that are both simple and complex, and it benefits meaningful, long-term learning in authentic educational settings. Why then were few positive effects observed in the studies highlighted by Van Gog and Sweller?

There are two clear explanations. First, as we have emphasized, de Jonge et al. (2015) and Tran et al. (2015) had people practice retrieval by filling in individual words in isolated sentences. That retrieval activity does not require the kind of integrative, relational processing that occurs in free recall, summarization, or answering inferential short-answer questions. The idea of element interactivity is indeed important for retrieval practice, but it is element interactivity during the retrieval activity that matters, not the complexity or element interactivity within the set of materials. The effects of fill-in-the-blank retrieval activities, like those used by de Jonge et al. and Tran et al., relative to more integrative retrieval activities were examined directly by Hinze and Wiley (2011), another report not included in Van Gog and Sweller's analysis (see Table 1). Hinze and Wiley showed that initial fill-in-the-blank tests did not produce retrieval practice effects relative to restudying, whereas freely recalling the material in paragraph format produced reliable retrieval practice effects.

Second, the worked-example experiments (Leahy et al., 2015; Van Gog and Kester 2012; Van Gog et al., 2015) essentially involved massed retrieval practice immediately after the occurrence of each problem. That is, students were given a worked example and then immediately given a problem to solve as a "retrieval practice" event. It is not at all clear that students needed to retrieve anything about the prior learning episode to solve such problems, and episodic retrieval is an essential ingredient for retrieval practice effects (Karpicke et al. 2014; Karpicke and Zaromb 2010; Lehman et al. 2014). Nevertheless, the task afforded immediate, massed retrieval practice at best, which requires little or no episodic context reinstatement (Delaney et al. 2010) and is certain to produce very poor long-term retention (e.g., Carpenter and DeLosh 2005; Karpicke and Bauernschmidt 2011; Karpicke and Roediger 2007; Pyc and Rawson 2007). In sum, the experiments reported by de Jonge et al. (2015) and Tran et al. (2015), and the worked-example experiments all failed to show retrieval practice effects because of the way retrieval practice was implemented, not because of the complexity of the materials.

Finally, even with these limitations—specifically, that the worked-example procedures involved massed retrieval practice—the data from worked-example experiments still show a small but positive benefit of retrieval practice. The data summarized by Van Gog et al. (2015) are very noisy, and the individual studies reported in their small-scale meta-analysis (their Figure 1) are all underpowered, as evidenced in part by the very large error ranges. Nevertheless, the data reported in the meta-analysis show an overall effect of $d=0.19$ with an error range that barely includes zero. Notably, the effect size observed in Leahy et al.'s (2015) Experiment 3 on a delayed final test was also $d=0.19$. Yet Van Gog et al. interpret the existing data not as evidence for a small, positive effect, but as evidence that there is no effect at all.

The results of the worked-example experiments, despite their problems, certainly show a small but positive effect of retrieval practice. A nonsignificant p value does not provide

evidence against an effect, or rather, in favor of a null effect (see Rouder et al. 2009). To gain more insight on the data, we entered the *t* statistics and samples sizes from the contrasts in Van Gog et al.'s (2015) small-scale meta-analysis into a Bayesian meta-analysis[4] (Rouder and Morey 2011). This allowed us to evaluate the strength of evidence in favor of the hypothesis that there was no effect ($d = 0$) relative to a small, positive effect ($0 < d < 0.20$). The Bayesian meta-analysis showed positive evidence in favor of a small effect relative to a null effect ($BF = 4.13$). In other words, given these data, a small positive effect is 4 times more likely than a null effect.

Conclusions

The claim that there is no testing effect for complex materials is incorrect. A wealth of research, reviewed here only briefly, has shown that practicing retrieval enhances learning of complex materials in educational settings. Much of this literature, including experiments that directly manipulated the complexity of the materials, was not included in Van Gog and Sweller's analysis. The experiments emphasized by Van Gog and Sweller involved either recall of isolated words in individual sentences or immediate, massed retrieval practice with worked-example materials. Retrieval practice effects were not observed in those experiments because of methodological issues, not because of the complexity of the materials. Despite the limitations of the worked-example experiments, they provided good evidence that there is a small but positive effect of (massed) retrieval practice with worked-example materials, contrary to Van Gog et al.'s interpretation. Finally, if element interactivity is to be a useful construct in educational research, it needs to be defined in a quantitative, measurable way. We offered referential cohesion as a possible measure, but it is likely that better measures can be developed. Ultimately, the influence of material complexity must be assessed in experiments that directly manipulate it. Given the results of the large base of relevant research, the testing effect is alive and well with complex materials.

Acknowledgments We thank James Nairne, the Purdue Cognition and Learning Laboratory, and Roddy Roediger and the Washington University Memory Laboratory for the comments on the manuscript. We also thank Mario de Jonge for providing materials.

References

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.

Agarwal, P. K., Karpicke, J. D., Kang, S. H., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open and closed book tests. *Applied Cognitive Psychology, 22*(7), 861–876.

Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology, 106*(3), 849–858.

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.

[4] We used the "*BayesFactor*" package (Morey and Rouder 2015) in the R programing language (R core team, 2014).

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*(4–5), 514–527.
Butler, A. C., Marsh, E. J., Slavinsky, J. P., & Baraniuk, R. G. (2014). Integrating cognitive science and technology improves learning in a STEM classroom. *Educational Psychology Review, 26*, 331–340.
Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*(5), 279–283.
Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology, 19*(5), 619–636.
Carpenter, S. K., & Kelly, J. W. (2012). Tests enhance retention and transfer of spatial learning. *Psychonomic Bulletin & Review, 19*(3), 443–448.
Chan, J. C. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61*(2), 153–170.
Chan, J. C. (2010). Long-term effects of testing on the recall of nontested materials. *Memory, 18*(1), 49–57.
Chan, J. C., McDermott, K. B., & Roediger, H. L., III. (2006). Retrieval-induced facilitation: initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*(4), 553.
Cook, L. K., & Mayer, R. E. (1988). Teaching readers about the structure of scientific text. *Journal of Educational Psychology, 80*(4), 448–456.
de Jonge, M., Tabbers, H. K., & Rikers, M. J. P. (2015). The effect of testing on the retention of coherent and incoherent text materials. *Educational Psychology Review*. doi:10.1007/s10648-015-9300-z
Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: a deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation, 53*, 63–147.
Dobson, J. L., & Linderholm, T. (2015). Self-testing promotes superior retention of anatomy and physiology information. *Advances in Health Sciences Education, 20*, 149–161.
Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58.
Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25*(2–3), 285–307.
Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-metrix providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223–234.
Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193–202.
Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory, 19*(3), 290–304.
Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test… or testing to teach: exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review, 26*(2), 307–329.
Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101*(3), 621–629.
Kang, S. H., 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*(4–5), 528–558.
Karpicke, J. D. (2012). Retrieval-based learning active retrieval promotes meaningful learning. *Current Directions in Psychological Science, 21*(3), 157–163.
Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(5), 1250–1257.
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., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: an episodic context account. *Psychology of Learning and Motivation, 61*, 237–284.
Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*(2), 151–162.
Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62*(3), 227–239.
Larsen, D. P., Butler, A. C., Lawson, A. L., & Roediger, H. L., III. (2013). The importance of seeing the patient: test-enhanced learning with standardized patients and written tests improves clinical application of knowledge. *Advances in Health Sciences Education, 18*(3), 409–425.

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

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

Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology, 38*(2), 94–97.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007a). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4–5), 494–513.

McDaniel, M. A., & Einstein, G. O. (1989). Material-appropriate processing: a contextualist approach to reading and studying strategies. *Educational Psychology Review, 1*(2), 113–145.

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy effective and portable. *Psychological Science, 20*(4), 516–522.

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007b). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*(2), 200–206.

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

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger III, 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*(1), 3.

Meyer, B. J. F. (1975). *The organization of prose and its effects on memory*. North-Holland Pub. Co

Morey, R. D. & Rouder, J. N. (2015). *BayesFactor: computation of Bayes factors for common designs*. R package version 0.9.10-2. http://CRAN.R-project.org/package=BayesFactor

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition, 35*(8), 1917–1927.

R Core Team (2014). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 233–239.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.

Roediger, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition, 1*(4), 242–248.

Roediger III, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17*(4), 382.

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review, 18*(4), 682–689.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237.

Tran, R., Rohrer, D., & Pashler, H. (2015). Retrieval practice: the lack of transfer to deductive inferences. *Psychonomic Bulletin & Review, 22*, 135–140.

Van Gog, T., & Kester, L. (2012). A test of the testing effect: acquiring problem-solving skills from worked examples. *Cognitive Science, 36*(8), 1532–1541.

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

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

Weinstein, Y., McDermott, K. B., & Roediger, H. L., III. (2010). A comparison of study strategies for passages: rereading, answering questions, and generating questions. *Journal of Experimental Psychology: Applied, 16*(3), 308–316.