Educ Psychol Rev (2015) 27:327–331
DOI 10.1007/s10648-015-9308-4

COMMENTARY

The Status of the Testing Effect for Complex Materials: Still a Winner

Katherine A. Rawson[1]

Published online: 28 April 2015

Abstract The target articles in the special issue address a timely and important question concerning whether practice tests enhance learning of complex materials. The consensus conclusion from these articles is that the testing effect does not obtain for complex materials. In this commentary, I discuss why this conclusion is not warranted either by the outcomes reported in the target articles or by the available evidence from prior research. Importantly, the weight of the available evidence does not alter the prescription for teachers and students to use practice testing to enhance learning of complex materials. However, the special issue highlights the need for more empirical and theoretical work on test-enhanced learning for complex materials, to further examine when and why these effects may be limited and to inform efforts to optimize test-enhanced learning for educationally relevant materials and tasks.

Keywords Testing effect · Retrieval practice · Test-enhanced learning · Complex materials · Material complexity · Worked examples · Problem solving

Practice testing is arguably one of the most potent learning techniques documented to date. Several hundred experiments from more than 100 years of research have established that taking a test does more than just assess learning—it actually enhances learning (for recent reviews, see Dunlosky et al. 2013; Roediger et al. 2011; Rowland 2015). Although practice testing has been shown to enhance learning under a wide range of conditions, it is unreasonable to assume that testing (or any other learning strategy) will work equally well for all learners, materials, and tasks. Thus, identifying factors that moderate the testing effect is important for both theoretical and applied purposes.

Whereas prior research has identified some moderators (as described in the target articles and discussed further below), this special issue highlights material complexity as a plausible moderator that has not received sufficient attention in the literature. Furthermore, although testing effects have been demonstrated across a wide range of learning materials, the target

✉ Katherine A. Rawson
krawson1@kent.edu

[1] Department of Psychology, Kent State University, P.O. Box 5190, Kent, OH 44242-0001, USA

articles correctly note that the extant literature is heavily populated by research involving relatively simple materials (e.g., word pairs or word lists). Because practice testing is increasingly being prescribed to students and teachers as an effective learning technique (e.g., Dunlosky et al. 2013; Pashler et al. 2007), one would want high confidence that testing effects also hold for the kinds of complex materials that are commonly the object of educational learning goals. Thus, this special issue addressing the extent to which testing effects depend on material complexity is timely and important. If the testing effect is consistently absent—or worse yet, reversed—for complex materials, this finding would certainly have important implications for prescriptions for teachers and students.

Is the Testing Effect Absent for Complex Materials?

Prior research has shown that testing effects may be absent and sometimes even reversed when criterion performance is assessed immediately or shortly after initial learning (as briefly discussed by van Gog and Sweller 2015). However, this known moderator is arguably not troublesome for educational purposes, given that the goal of education is long-term maintenance of knowledge. Thus, the question of greatest practical interest here is whether the testing effect is absent for complex materials when criterion performance is assessed after a delay.

First, here is the good news: In the target articles, none of the experiments involving a delayed criterion test found a reversed testing effect. But here is the potentially bad news: Only one experiment reported a statistically significant positive testing effect (de Jonge et al. 2015, experiment 2) and then only for an incoherent text (d=0.58 vs. d=0.13 for the coherent text in experiment 1).[1] Based on the outcomes reported in their target article, van Gog et al. (2015) concluded that "In none of these experiments, nor in an overall analysis, did we find evidence of a testing effect (current p. 19)." Similarly, Leahy et al. (2015) conclude that "The testing effect may not be obtainable using high element interactivity materials (current p. 11)."

However, these all-or-none conclusions are arguably too strong and too heavily weighted by failures to reach conventional levels of statistical significance in underpowered experiments. For example, the observed effect in Leahy et al. (2015) experiment 3 was d=0.20, which was not significant given the small sample size (achieved level of power was only 0.08).[2] Similarly, the mean effect size estimate in the mini-meta-analysis reported by van Gog et al. (2015) was d=0.19, but despite the large combined sample, achieved power for the reported two-tailed test was 0.43. Arguably, a one-tailed test would be warranted for this comparison (given that the test is of an a priori directional prediction that testing outperforms restudy), which would have yielded a significant p=0.038. With that said, the effects demonstrated across the target articles are undeniably small, but they are consistently positive. Thus, the weight of the evidence across the target articles does not easily support the all-or-none conclusion that the testing effect is absent for complex materials. Rather, a more nuanced but arguably more appropriate conclusion about outcomes from the target articles would be that the testing effect generalizes in the expected direction but not in effect size.

Broader conclusions concerning whether testing effects obtain for complex materials can also be informed by available evidence from prior research. To this end, what is the evidence from prior research involving the kinds of complex materials used in the target articles (either

[1] I computed Cohen's d here and below using pooled standard deviation (as per Cortina and Nouri 2000).
[2] I computed achieved power here and below using G*Power 3.1.9.2 (Faul et al. 2007).

problem-solving tasks or text materials)? Concerning prior research involving problem-solving tasks, few prior studies in the worked-example literature have administered criterion tests after a delay, and few of these studies included the practice conditions of interest here (example-problem pairs vs. examples only, as in the target articles by Leahy et al. and van Gog et al., or the closely related comparison of examples only vs. problems only).[3] van Gog and Kester (2012) compared example-problem pairs vs. examples only for novice undergraduates learning how to troubleshoot electrical circuits and found a reversed testing effect on a criterion test administered 1 week later (d=−0.66), which is potentially worrisome. However, Darabi et al. (2007) showed strong positive effects of practice tests in a troubleshooting task. Undergraduates in engineering courses worked with software simulating a water-alcohol distillation plant to diagnose and repair malfunctions. After initial basic instruction,[4] students either studied four descriptive worked examples (similar in kind to those used in the target articles) or completed four problem-solving trials. On a transfer test several days later, the problem-solving group significantly outperformed the worked-example group (d=0.98), which is quite promising. Given the mixed outcomes reported by van Gog and Kester (2012) and Darabi et al. (2007) and the relative paucity of research on problem-solving tasks involving delayed tests, more research including these conditions is essential before any strong conclusions can be drawn.

Fortunately, the extant literature concerning testing effects with text material is on firmer footing. It is true that the majority of prior research on testing effects involved simpler materials, particularly older research (with notable exceptions such as Gates 1917, as described by van Gog and Sweller 2015). But the recent surge of research on test-enhanced learning has increasingly involved text materials. Rawson and Dunlosky (2011) summarized methods from 168 experiments reported from 2000 to 2010, and 36 of these involved text or lecture materials (32 of the 36 experiments included a delayed criterion test, although not all of them directly compared a testing condition to a restudy condition). Since 2010, many more papers have been published on test-enhanced learning for text materials with delayed criterion tests (some of these are summarized in Table 1 of van Gog and Sweller 2015). Thus, the available prior outcomes are too extensive to describe at length here. As luck would have it, Rowland (2015) recently conducted a meta-analysis of the testing effect literature, specifically focusing on comparisons of testing vs. restudy and examining type of material as a moderator. The mean weighted effect size was similar for prose (g=0.58, based on k=23 effect size estimates) vs. paired associates (g=0.59, k=71) and stronger than for word lists (g=0.39, k=58). Rowland also reported outcomes for the subset of studies that either included feedback and/or reported at least 75 % performance on the practice tests, given that these conditions represent a more level playing field with respect to re-exposure to target information (i.e., all information is re-exposed

[3] A few other studies that included delayed criterion tests compared example-problems to problems only, and the general finding is that example-problem conditions outperform problem-only conditions (e.g., Carroll 1994; Salden, Aleven, Renkl, and Schwonke 2009; Ward and Sweller 1990). Note that the analog comparison in the testing effect literature would be study-test conditions vs. test-only conditions, and research has consistently shown study-test to outperform test only. Thus, the modal outcomes in the worked example and testing effect literatures are consistent with one another on this front (i.e., the directional effect of study-test vs. test only does not appear to be moderated by material complexity).

[4] Darabi et al.'s article stated that initial instruction included description of how to troubleshoot malfunctions of components in the plant, but a reviewer noted that they did not state explicitly whether these instructions specifically described to how to solve troubleshooting problems like those encountered in the practice phase. With less specific initial instruction, one would arguably expect a weaker testing effect, to the extent that problem-solving performance during practice would be lower (i.e., less effective practice tests).

during restudy, whereas in the absence of feedback, only correctly retrieved information is re-exposed during practice testing). Under these conditions, estimated effects for prose were even more impressive (g=0.73, k=13, vs. g=0.69, k=47 for paired associates or g=0.64, k=27 for word lists). Thus, the testing effect would appear to be alive and well for text materials.

In sum, the weight of the available evidence (from the target articles and from prior research) does not support the conclusion that the testing effect is not obtainable for complex materials. Although the magnitude of the effect is disappointingly small in some cases, with very few exceptions, it is consistently positive. If testing is often but not always substantially better than restudy, the prescriptive conclusions for teachers and students remain unchanged: Testing is still the strategy that has the highest likelihood of paying off.

In closing, I again point to the importance and timeliness of this special issue, which highlights the need for more empirical and theoretical work on test-enhanced learning for complex materials. Although the weight of the evidence still favors practice testing as an effective learning technique for complex materials, more research is clearly needed to further examine when and why these effects may be limited, which in turn can inform efforts to optimize test-enhanced learning for educationally relevant materials and tasks.

Acknowledgments The author would like to thank John Dunlosky for the helpful input on the content of this article.

References

Carroll, W. M. (1994). Using worked examples as an instructional support in the algebra classroom. *Journal of Educational Psychology, 86*, 360–367.

Cortina, J. M., & Nouri, H. (2000). *Effect size for ANOVA designs*. Thousand Oaks: Sage.

Darabi, A. A., Nelson, D. W., & Palanki, S. (2007). Acquisition of troubleshooting skills in a computer simulation: worked example vs. conventional problem solving instructional strategies. *Computers in Human Behavior, 23*, 1809–1819.

de Jonge, M., 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.*

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

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

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

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.*

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning (NCER 2007–2004)*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: how much is enough? *Journal of Experimental Psychology: General, 140*, 283–302.

Roediger, H. L. I. I. I., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation, 55*, 1–36.

Rowland, C. A. (2015). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin.*

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

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.*

Ward, M., & Sweller, J. (1990). Structuring effective worked examples. *Cognition and Instruction, 7*, 1–3.