



The effect of practice test modality on perceived mental effort and delayed final test performance

Leonora Coppens , Mario de Jonge , Tamara van Gog & Liesbeth Kester

To cite this article: Leonora Coppens , Mario de Jonge , Tamara van Gog & Liesbeth Kester (2020) The effect of practice test modality on perceived mental effort and delayed final test performance, Journal of Cognitive Psychology, 32:8, 764-770, DOI: [10.1080/20445911.2020.1822366](https://doi.org/10.1080/20445911.2020.1822366)

To link to this article: <https://doi.org/10.1080/20445911.2020.1822366>



Published online: 17 Sep 2020.



Submit your article to this journal [↗](#)



Article views: 151





View related articles [↗](#)



View Crossmark data [↗](#)



The effect of practice test modality on perceived mental effort and delayed final test performance

Leonora Coppens , Mario de Jonge, Tamara van Gog and Liesbeth Kester 

Department of Education, Utrecht University, the Netherlands

ABSTRACT

Taking a test on studied materials results in better delayed recall performance than restudying (a.k.a. the *testing effect*). A common finding in testing effect research is that the effect depends on test format: the magnitude of the testing effect differs between free-recall, cued-recall, and recognition testing. This is explained by the effortful retrieval hypothesis: effortful successful retrieval results in better memory for an item than less effortful successful retrieval. However, the assumption that successful retrieval on different types of tests requires different levels of effort has not yet been tested. To test this assumption, we measured perceived mental effort on different test formats. Participants indicated free-recall was more effortful than cued-recall, and cued-recall more effortful than recognition. Furthermore, cued and free-recall yielded better cued-recall performance on a one-week delayed test than restudy or recognition. The results support the assumption that different practice test formats require different levels of mental effort.

ARTICLE HISTORY

Received 16 December 2019
Accepted 7 September 2020

KEYWORDS

mental effort; test format;
effortful retrieval; retrieval
practice

Decades of research have shown that taking a test on studied material enhances the retention of successfully retrieved information compared to restudying the material (a.k.a. the testing effect; for overviews see Adesope et al., 2017; Roediger & Karpicke, 2006; Rowland, 2014). One of the prevailing hypotheses put forward to explain this beneficial effect of retrieval practice on longer-term retention is the effortful retrieval hypothesis (Pyc & Rawson, 2009). According to the effortful retrieval hypothesis (Pyc & Rawson, 2009), given that a retrieval attempt is successful, retrieval that took more mental effort is more beneficial for retention than retrieval that took little mental effort. This hypothesis stems from the desirable difficulties-framework (Bjork, 1994), which states that difficult but successful processing is more beneficial for learning than easy and successful processing. The effortful retrieval hypothesis is typically investigated by using different types of retrieval practice tests that are presumed to differ in the amount of mental effort they require the learner to invest (Kang et al., 2007; Stenlund et al., 2016). However, the underlying assumption that different test formats elicit different levels of mental effort has not yet been directly tested. The present study aims to do so.

1.1. Effortful retrieval

To explain why difficult but successful processing is more effective for longer-term retention than easy and successful processing, Bjork and Bjork (1992) differentiate between retrieval strength and storage strength. Storage strength refers to the degree to which an item has been “learned”, or durably remembered (i.e. the greater the storage strength, the better the retention). Retrieval strength refers to the accessibility of an item (i.e. the greater the retrieval strength, the more easily accessible the item). They argue that retrieval strength of retrieved items and the increase in storage strength due to retrieving it are negatively correlated. Thus, the harder it is to access an item in long-term memory, the better this item will be retained once it is successfully retrieved (e.g. Glover, 1989; Pyc & Rawson, 2009).

The mechanisms underlying the effect of retrieval effort are further explained by elaborative-processing views on retrieval (e.g. Carpenter, 2009; Carpenter & DeLosh, 2006; Glover, 1989; Whitten & Leonard, 1980). According to these views, the memory search for a specific item activates related items in memory (i.e. elaborative information). So, a retrieval attempt causes the activation of elaborative information

related to the target item. This activated information then provides multiple pathways to access the target item on future retrieval attempts and, therefore, strengthens the retention of the target item. According to these elaborative-processing views on retrieval, the more mental effort is needed to find the target item (i.e. the lower the retrieval strength), the larger the activated elaborative structure and, thus, the more pathways that lead to the target information on future retrieval attempts (i.e. the higher the storage strength).

The effortful retrieval hypothesis is often tested by varying test formats. For instance, Stenlund et al. (2016) let students study a section of a textbook and then had them take practice tests on the information in the form of multiple-choice questions (presumed to require less effort investment) or short answer questions (presumed to require more effort investment), or had them reread factual statements. The final test consisted of multiple choice or short answer questions and was administered after 5 min, after one week and after four weeks. Short answer questions and multiple-choice questions produced more learning than rereading, and short answer questions produced better performance on the short answer final test than multiple choice questions. Thus, the more effortful the practice test method presumably was, the better participants' retention was. However, although it is a central assumption in the effortful retrieval hypothesis, it is still an open question whether different test formats actually elicit different levels of perceived retrieval effort, as this was not assessed by Stenlund and colleagues (2016).

Although it is generally found that cued and free recall tests yield better performance on a later recall test than recognition tests (e.g. Carpenter & DeLosh, 2006; Hogan & Kintsch, 1971), there are mixed findings on the difference in effectiveness between cued and free recall. In a meta-analysis, Rowland (2014) found that cued recall yielded larger testing effects than free recall, when including all testing effect studies. However, this could be because cued recall more often includes feedback and performance on cued recall tests is usually higher than on free recall tests. When looking at only studies where participants received feedback or had an initial test performance of 0.75 or higher, there was no significant testing effect difference between free and cued recall.

1.2. Effort of different test formats

Endres and Renkl (2015) measured perceived mental effort in a study using different forms of

retrieval practice. After studying three texts, participants took a practice test consisting of short-answer questions on one text, a test of free-recall questions on the second text, and reread the third text. They rated their perceived mental effort after each test and reread phase. Endres and Renkl found evidence of a testing effect: tested information was remembered better than reread information after a one-week delay. Interestingly, the effect of testing disappeared when the authors statistically controlled for perceived mental effort. This suggests that for learning texts, the effect of testing is dependent on the required mental effort. While the authors set out to investigate the elaborative retrieval hypothesis, their results also seem to provide some evidence that different test formats require different levels of retrieval effort.

However, Endres and Renkl (2015) did not find differences in retrieval effort between the different practice formats. Furthermore, perceived mental effort was only measured once after each entire test, instead of after each test item. Not only could this affect the mental effort rating (i.e. research has shown that one overall rating after all items is typically higher than an average of ratings obtained immediately after each item; Schmeck et al., 2015; Van Gog et al., 2012), but it also means there was no way to distinguish between successfully answered test questions and questions that were not or incorrectly answered. The present study was designed to be able to investigate differences in retrieval effort between testing formats on successful trials, by measuring perceived mental effort after each trial. It has long been clear that there are more differences between test formats than just their presumed effort. For instance, Hogan and Kintsch (1971) argued that the processes of recognition and recall are fundamentally different, and that a free recall test requires both recall (to produce initial candidate answers) and recognition (to select the most suitable answer), whereas a recognition test requires only recognition. However, because the focus of the current paper is on effort, we will focus on the difference in effort between test formats.

1.3. The present study

The main aim of the present experiment was to test the assumption that different practice test formats require different levels of perceived mental effort

investment. We had participants learn word pairs through three retrieval practice strategies, namely, recognition, cued recall and free recall, and compared this to a restudy control condition. We measured perceived mental effort immediately after each item during retrieval practice using Paas' (1992) 9-point rating scale. For completeness, we also tested performance on a cued retention test one week later. Based on the effortful retrieval hypothesis, we expected that free recall would lead to the highest perceived mental effort on successful trials and the best retention of successfully retrieved word pairs, followed by cued recall, followed by recognition.

2. Method

2.1. Participants and design

Forty-four participants were recruited through social networks; they participated voluntarily and did not receive a reward for participation. Data from two participants had to be excluded because of a technical error during the experiment, which resulted in a final sample of 42 participants (31 female, 11 male, age 18 - 80 years, *Mage* 34.5). The study had a within-subjects design with four retrieval practice conditions: 1) no retrieval practice (i.e. restudy), 2) recognition, 3) cued recall, and 4) free recall. The order of the conditions was randomised for each participant and the assignment of word pairs to the conditions was counterbalanced across participants. The study was conducted in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. We obtained informed consent from all participants prior to the experiment.

2.2. Materials

The entire experiment was programmed and run in Gorilla.sc (www.gorilla.sc/about).

2.2.1. Word pairs

We used Carpenter's (2009) "weak cue - target" list of 48 word pairs. We randomly selected 24 word pairs from this list to serve as the learning material in the learning phase. Eighteen other word pairs from the list were used as lures in the recognition test. The word pairs were translated to Dutch.

2.2.2. Learning phase

In study and restudy trials, participants were instructed to try to memorize the word pairs. One word pair at a time was presented in the centre of the screen until the participant clicked a button to continue.

In recognition trials, word pairs (mixed with lures) were presented one by one and participants were instructed to indicate whether they had seen the word pair before in the experiment, by clicking a "yes" or "no" button. No feedback was provided.

In cued recall trials, only the first word of each pair was presented and the participant was instructed to type the second word to the right of the first word. The trial ended when the participant clicked a button to continue. No feedback was provided.

In free recall trials, participants saw an empty text box in the centre of the screen and were instructed to enter the word pairs they had just learned one by one. After entering a word pair, they clicked a button to get a new empty text box. No feedback was provided.

2.2.3. Final test

The final test was identical to the cued recall test in the learning phase: the first word of each pair was presented and the participant tried to complete the word pair, typing in their answer. As described in the Introduction, the final test was of secondary interest and only added for completeness. We used a cued recall final test for all items because this enabled us to measure the testing effect emerging from all intervening test formats (cf. Glover, 1989, Experiment 4).

2.2.4. Mental effort rating

Paas' (1992) subjective 9-point rating scale was used to measure perceived mental effort. It ranges from (1) very, very low mental effort to (9) very, very high mental effort. Immediately after each restudy or practice test trial, we instructed participants as follows: "Please indicate on a scale from 1 to 9 how much mental effort you invested in studying the previous word pair" (restudy), "Please indicate on a scale from 1 to 9 how much mental effort you invested in recognising the previous word pair" (recognition practice test) or "Please indicate on a scale from 1 to 9 how much mental effort you invested in retrieving the previous word pair from memory" (cued and free recall practice tests).

Participants clicked on one of 9 buttons numbered 1–9 to indicate their mental effort.

2.3. Procedure

Participants were tested on a laptop or pc, individually in a quiet room. The learning phase started with a brief introduction of the experimental task, in which participants were instructed to try to remember the word pairs for a later (unspecified) test. The entire learning phase was self-paced. The word pairs were presented in four blocks. In each block a set of six word pairs was practiced. Each block started with an instruction that was specific for the condition in which the word pairs were to be practiced. The six word pairs in that block were then presented for study once, one word pair at a time, with each word pair being shown until the participant clicked a button to continue. When all six word pairs of a set had been studied once, the participant engaged in restudying, recognition, cued recall, or free recall of that set of word pairs (depending on the condition the set was assigned to) three times (e.g. study – recognition – recognition – recognition). The six word pairs of the set were presented one by one in a new random order each of those three times. After each restudy/practice trial, participants rated their invested mental effort on the Paas scale (so for each word pair, a participant gave three mental effort ratings in total). When all six word pairs in a block had been studied and practiced three times, the next block started until all 24 word pairs had been practiced.

After one week, participants were asked to log in for the final test. This test started with a brief welcome after which the participants read the instruction for the cued recall test (ca. 30 s). All cues were presented one by one on the screen and the participants tried to complete the word pairs, typing in their answers.

2.4. Scoring

Each correctly recognised or recalled word pair on the practice tests and each correctly recalled target on the final cued recall test was rewarded a full point. Recall of only a target or a cue in the free recall practice test was given a half point. Minor

typing errors, in which one letter was missing or in the wrong place, were corrected. False recognition of a lure in the recognition test resulted in a one-point deduction. Two raters independently scored all cued recall and free recall answers and a high degree of inter-rater reliability was found: Spearman's rho was $\rho = .992$ for cued recall during the learning phase, $\rho = .954$ for free recall during the learning phase, and $\rho = 1.000$ for the final test. The range of potential scores per condition was 0–18 on the practice tests, and 0–6 on the final test. We converted the scores to proportions for ease of comparison of practice test and final test performance.

3. Results

Descriptives of perceived mental effort and final test performance are given in Table 1. Because the variables were not normally distributed¹, the results were analysed with Friedman tests, with Wilcoxon signed rank tests to follow up on any significant overall effects.

3.1. Learning phase

3.1.1. Perceived mental effort

There was a significant effect of condition on perceived mental effort invested in successful trials (i.e. correctly recognised/recalled word pairs on the practice tests) during learning, $X^2(2) = 33.78$, $p < .001$. Participants reported the highest mental effort on free recall trials, followed by cued recall trials, followed by recognition trials (Table 1). Follow-up Wilcoxon signed rank tests with an adjusted alpha of .0167 indicated that perceived mental effort significantly differed between all test formats: recognition vs. cued recall: $Z = 3.93$, $p < .001$; cued recall vs. free recall, $Z = 3.27$, $p = .001$; recognition vs. free recall: $Z = 5.08$, $p < .001$.

3.1.2. Performance on the practice test trials

There was a significant effect of condition on performance on the practice test trials during the learning phase, $X^2(2) = 56.60$, $p < .001$. Performance in the learning phase was best on recognition trials, followed by cued recall trials, followed by free recall trials (Table 1). Follow-up Wilcoxon signed rank tests with an adjusted alpha of .0167 indicated

¹The non-normal (skewed) distribution of mental effort scores may have been caused by the setup of the experiment. Because the retention interval was longer, the current experiment included more repetitions than the experiment of Carpenter (2009), in which these word pairs were previously used. It is possible that the materials were (perceived as) easy to learn in the current experiment because of the additional exposure in the learning phase, which caused the skewed distribution of scores.

Table 1. Median (first; third Quartile) effort ratings during the learning phase and proportion correctly recalled word pairs during the learning phase and on the final test, per condition.

	Restudy <i>Mdn</i> (Q1;Q3)	Recognition practice test <i>Mdn</i> (Q1;Q3)	Cued recall practice test <i>Mdn</i> (Q1;Q3)	Free recall practice test <i>Mdn</i> (Q1;Q3)
Proportion correct practice test	-	1.00 (0.94; 1.00)	0.94 (0.65; 1.00)	0.40 (0.22; 0.72)
Proportion correct final cued recall test	0.33 (0.00; 0.50)	0.33 (0.17; 0.54)	0.58 (0.29; 0.83)	0.50 (0.17; 0.88)
Perceived mental effort during correct practice test trials	-	1.12 (1.00; 1.29)	1.47 (1.14; 3.05)	3.00 (1.67; 5.12)
Perceived mental effort during restudy trials *	1.17 (1.00; 2.10)	-	-	-
Perceived mental effort during learning phase on word pairs correct on practice test and correct on final test	-	1.11 (1.00; 1.33)	1.47 (1.13; 2.33)	3.20 (1.67; 5.90)
Perceived mental effort during learning phase on word pairs correct on practice test but not on final test	-	1.09 (1.00; 1.27)	1.53 (1.00; 2.46)	2.86 (1.50; 5.50)

Note. *For completeness, we also had participants rate their perceived mental effort during restudy trials. The question posed to participants after restudy trials was “how much effort did you invest in studying this item”, instead of “how much effort did you invest in retrieving this item” after test trials.

that performance during the learning phase differed between all test formats: recognition vs. cued recall: $Z = 3.59, p < .001$; cued recall vs. free recall, $Z = 4.50, p = .001$; recognition vs. free recall: $Z = 5.51, p < .001$.

3.2. Final test performance

There was a significant effect of condition on the number of correctly recalled word pairs on the final test, $X^2(3) = 19.49, p < .001$. Follow-up Wilcoxon signed rank tests with an adjusted alpha of .008 indicated a significant difference in performance on the final test between restudy and cued recall, $Z = 3.35, p = .001$, between restudy and free recall, $Z = 3.26, p = .001$, and between recognition and cued recall, $Z = 2.91, p = .004$. There were no significant differences between restudy and recognition, $Z = 1.52, p = .128$, between recognition and free recall, $Z = 2.532, p = .011$, or between cued recall and free recall, $Z = 0.27, p = .789$. Participants remembered the word pairs better after cued recall than restudy or recognition, and better after free recall than restudy.

3.3. Exploratory analysis: mental effort during practice tests on items remembered vs. forgotten on the final test

As an exploratory analysis, we investigated whether participants reported to have invested more effort in “remembered items”, that is, correct practice test trials on items that were also correctly recalled on the final test one week later than in “forgotten items”, that is, correct practice test trials on items that were not correctly recalled on the final test. A

difference in invested mental effort between remembered (i.e. higher mental effort) and forgotten items (i.e. lower mental effort) would be in line with the effortful retrieval hypothesis, which predicts that retrieval effort is beneficial for retention. We used Wilcoxon signed ranks tests with an adjusted alpha of .0167. Descriptives are given in the last two rows of Table 1. There was no difference in perceived mental effort between remembered and forgotten items in any of the conditions; recognition: $Z = 0.02, p = .983$; cued recall: $Z = 0.09, p = .931$; free recall: $Z = 0.16, p = .875$.

3.4. Data sharing

The data that support the findings of this study are openly available from the Open Science Framework at https://osf.io/4h2mz/?view_only=69c142063517408ea7b04998aaddf8e9.

4. Discussion

The aim of the present experiment was to test the assumption that different practice test formats require different levels of mental effort investment. To this end, we had participants learn word pairs through repeated study, recognition tests, cued recall, and free recall, and participants rated their mental effort directly after every trial. Furthermore, we measured retention of the word pairs through a cued recall test one week after learning. Based on the effortful retrieval hypothesis, we expected that free recall tests would yield the highest perceived mental effort and the best retention of

successfully retrieved word pairs, followed by cued recall, followed by recognition.

We indeed found differences in perceived mental effort between the conditions: successful free recall trials were rated as more effortful than cued recall, and successful recognition trials were rated as less effortful than successful cued recall trials. These results support the assumption that these different test formats require different levels of retrieval effort.

In line with our expectations, we also found a testing effect: on the final test after one week, word pairs were remembered better when they had been practiced with cued recall than through restudy or recognition, and better when they had been practiced with free recall than with restudy. Thus, our findings provide more direct evidence, from perceived mental effort ratings, for the effortful retrieval hypothesis (compared to prior research in which effort differences among practice formats were only assumed but not measured; Stenlund et al., 2016, or not measured after each item; Endres & Renkl, 2015).

However, we did not find full support for the effortful retrieval hypothesis. According to this hypothesis, word pairs should also be remembered better when they have been retrieved with more effort. Free recall is considered to be more effortful than cued recall or recognition, and our participants indeed indicated having invested significantly more mental effort in free recall, but we found no difference in final test performance between free recall and cued recall or recognition. A possible explanation for this finding is that performance on the free recall trials in the learning phase was very low (median proportion correct: 0.4). In future studies, performance on free recall trials could be boosted by using even shorter lists (e.g. lists of three items that are immediately repeated, instead of the six items in a set in the current experiment). Another explanation could be transfer-appropriate processing (Morris et al., 1977): the match between initial and final test format could have improved performance on items in the cued recall condition, although there is little support for the transfer-appropriate processing account of the testing effect (e.g. Carpenter & DeLosh, 2006; Glover, 1989).

Moreover, our exploratory analysis failed to show a difference in perceived mental effort among remembered and forgotten items while, based on the effortful retrieval hypothesis, mental effort should have been higher for remembered items than for forgotten items. A result that may seem

surprising is that for items practiced through free recall, performance was better on the final test than on the practice tests. This finding can be explained by the fact that the final test was a cued recall test. It is likely that these items were not learned well enough (i.e. the retrieval strength was not high enough) to retrieve them on a free recall practice test. On the final test however, when provided with a strong retrieval cue (i.e. the first word of each pair) the items could be retrieved. A more difficult (e.g. free recall) final test could solve this problem in future studies and could also further clarify differences between conditions (bifurcation; Halamish & Bjork, 2011).

In conclusion, the results of the present experiment provide support for the assumption that different practice test formats require different levels of mental effort. Students and teachers seeking to boost longer-term retention would be well-advised to select more effortful retrieval practice formats like cued or free recall practice over less effortful ones such as recognition.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Acknowledgments

The authors would like to thank Michelle Janssen for her assistance with collecting the data.

ORCID

Leonora Coppens  <http://orcid.org/0000-0002-9846-3534>
Liesbeth Kester  <http://orcid.org/0000-0003-0482-0391>

References

- Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, 87(3), 659–701. <https://doi.org/10.3102/0034654316689306>
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.
- Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes*, 2, 35–67.
- 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(6), 1563–1569. <https://doi.org/10.1037/a0017021>
- 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(2), 268–276. <https://doi.org/10.3758/BF03193405>
- Endres, T., & Renkl, A. (2015). Mechanisms behind the testing effect: An empirical investigation of retrieval practice in meaningful learning. *Frontiers in Psychology*, 6, 1054. <https://doi.org/10.3389/fpsyg.2015.01054>
- Glover, J. A. (1989). The “testing” phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, 81(3), 392–399. <https://doi.org/10.1037/0022-0663.81.3.392>
- Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(4), 801–812. <https://doi.org/10.1037/a0023219>
- Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning & Verbal Behavior*, 10(5), 562–567. [https://doi.org/10.1016/S0022-5371\(71\)80029-4](https://doi.org/10.1016/S0022-5371(71)80029-4)
- Kang, S. H. K., McDermott, K. B., & Roediger, H. L. I. I. (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. <https://doi.org/10.1080/09541440601056620>
- Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behavior*, 16(5), 519–533. [https://doi.org/10.1016/S0022-5371\(77\)80016-9](https://doi.org/10.1016/S0022-5371(77)80016-9)
- Paas, F. G. W. C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics - a cognitive-load approach. *Journal of Educational Psychology*, 84(4), 429–434. <https://doi.org/10.1037/0022-0663.84.4.429>
- 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(4), 437–447. <https://doi.org/10.1016/j.jml.2009.01.004>
- Roediger, H. L. I.I.I., & Karpicke, D. J. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1(3), 181–210. <https://doi.org/10.1111/j.1745-6916.2006.00012.x>
- Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, 140(6), 1432–1463. <https://doi.org/10.1037/a0037559>
- Schmeck, A., Opfermann, M., Van Gog, T., Paas, F., & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: Differences between immediate and delayed ratings. *Instructional Science*, 43(1), 93–114. <https://doi.org/10.1007/s11251-014-9328-3>
- Stenlund, T., Sundström, A., & Jonsson, B. (2016). Effects of repeated testing on short- and long-term memory performance across different test formats. *Educational Psychology*, 36(10), 1710–1727. <https://doi.org/10.1080/01443410.2014.953037>
- Van Gog, T., Kirschner, F., Kester, L., & Paas, F. (2012). Timing and frequency of mental effort measurement: Evidence in favor of repeated measures. *Applied Cognitive Psychology*, 26(6), 833–839. <https://doi.org/10.1002/acp.2883>
- Whitten, W. B., & Leonard, J. M. (1980). Learning from tests: Facilitation of delayed recall by initial recognition alternatives. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2), 127–134. <https://doi.org/10.1037/0278-7393.6.2.127>