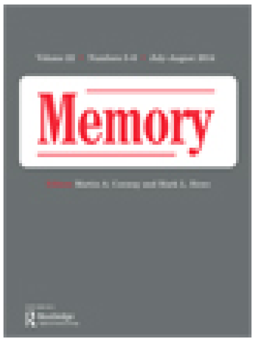

Memory

ISSN: 0965-8211 (Print) 1464-0686 (Online) Journal homepage: http://www.tandfonline.com/loi/pmem20

Is retrieval the key? Metamemory judgment and testing as learning strategies

Elçin Akdoğan, Marie Izaute, Jean-Marie Danion, Pierre Vidailhet & Elisabeth Bacon

MEMORY, 2015
http://dx.doi.org/10.1080/09658211.2015.1112812

Routledge
Taylor & Francis Group

Is retrieval the key? Metamemory judgment and testing as learning strategies

Elçin Akdoğan[a], Marie Izaute[b], Jean-Marie Danion[a,c], Pierre Vidailhet[a,c] and Elisabeth Bacon[a]

[a]INSERM U-1114, Fédération de Médecine Translationnelle de Strasbourg (FMTS), University of Strasbourg, France; [b]LAPSCO, UMR CNRS 6024, Clermont University, Clermont Ferrand, France; [c]Psychiatry Department, University Hospital, Strasbourg, France

ABSTRACT

Re-reading is the most common learning strategy, albeit not a very efficient one. Testing is highly efficient, but not perceived by students as a learning strategy. Prospective judgment-of-learning (JOL) reflect the learner's impression of subsequently being able to retrieve the ongoing learning in a cued-recall task. Estimating JOL involves attempting to retrieve the information, as in testing. The few studies that have explored the potential mnemonic benefit of JOL have yielded contradictory results. Our aim was to compare JOL and testing with re-study and to examine the impact of these strategies according to the relative difficulty of the material (cue-target association strength) in two experiments. After a first encoding phase, participants re-studied, provided JOL, or took a test. Forty-eight hours later, they participated in a final cued-recall test, during which their confidence level judgments were collected. The main result was that delayed JOL behaved in the same way as testing, and both yielded better performances than re-study when material was of moderate difficulty. The easy or very difficult material revealed no differences between these strategies. JOL is proposed as an alternative to testing when faced with difficult material.

ARTICLE HISTORY
Received 22 October 2014
Accepted 21 October 2015

KEYWORDS
Memory; metamemory; JOL; testing; metacognition

The majority of students spontaneously resort to re-reading as their common learning strategy (Karpicke, Butler, & Roediger III, 2009). However, according to a review of the efficiency of learning techniques, re-reading is regarded as having "low utility" for learning, compared with the "high utility" of practice testing (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013).

The testing effect suggests "taking a test on material can have a greater positive effect on future retention of that material than spending an equivalent amount of time restudying the material" (Roediger III & Karpicke, 2006b, p. 181). The growing literature on the testing effect, which is regarded as more advantageous for long-term rather than short-term retention (Roediger III & Karpicke, 2006a; Toppino & Cohen, 2009; Wheeler, Ewers, & Buonanno, 2003), has repeatedly highlighted the technique's efficiency in both laboratory conditions (Carrier & Pashler, 1992; Carpenter, Pashler, Wixted, & Vul, 2008; Roediger III & Karpicke, 2006a; Wheeler et al., 2003), as well as real-world settings (Agarwal, Karpicke, Kang, Roediger III, & McDermott, 2008; Larsen, Butler, & Roediger III, 2008).

Surprisingly, in their article reviewing 10 learning techniques, Dunlosky et al. (2013) omitted to evaluate the self-monitoring approach despite the fact that it "prescribes how people can regulate study across to-be-learned materials in a manner that enhances the efficacy of learning" (Dunlosky et al., 2007, p. 70). It is an approach where learners monitor which material has been learned, and which material has not been well learned. Following a learning session, the judgment-of-learning (JOL) reflects in particular the learner's impression of subsequently being able to retrieve the ongoing learning in a cued-recall task. Accurate monitoring signals one's ability to discriminate between those items that have been well learned and those that have not. The highest degree of accuracy is obtained with a delayed cue-only JOL (Dunlosky & Nelson, 1992). Delayed JOLs are JOLs expressed after a time-lag between the study and the JOL estimate. Cue-only JOLs are JOLs requested in the presence of the cue alone, rather than the cue and target together (Dunlosky & Nelson, 1992; Kelemen & Weaver III, 1997; Nelson & Dunlosky, 1991). Delayed JOLs are accurate predictors of actual memory, and this type of monitoring has been used successfully to enhance learning when monitoring is followed by re-learning (Dunlosky et al., 2007; Dunlosky, Hertzog, Kennedy, & Thiede, 2005).

According to Nelson and Dunlosky (1991), prior to their introspective JOL assessment, participants attempt to retrieve the target answer. In other words, their judgment will be based on their ability to retrieve the target information. This behaviour is reported by the participants themselves. In the study conducted by Nelson and Dunlosky (1991), 19 of the 20 adults taking part reported having tried to recall the response when rating delayed

CONTACT Marie Izaute marie.izaute@univ-bpclermont.fr

JOLs. This finding is also in line with the observation that a more accurate JOL is obtained when it is made in the presence of the cue alone rather than the word pair (Dunlosky & Nelson, 1992). When the cue is presented alone, the subject can attempt to retrieve the target, whereas when it is accompanied by the target, such an attempt is no longer feasible. Consequently, retrieval practice could be an underlying mechanism common to both the testing and delayed JOLs. Two recent studies looked at the potential memory benefits of JOL. In the first, the authors compared a sequence of four study sessions with a study-JOL-study-JOL sequence (Sundqvist, Todorov, Kubik, & Jönsson, 2012). In the second study, the effect of four study-test sessions was compared with that of four study-JOL sessions (Jönsson, Hedner, & Olsson, 2012). In the final test, a week later, neither study observed any difference between the groups. The observation that JOL monitoring was no different to either re-study (Sundqvist et al., 2012) or testing (Jönsson et al., 2012) is perplexing, since testing has consistently been shown to produce better performance than re-study (Roediger III & Karpicke, 2006b). However, to the best of our knowledge, no study has directly compared delayed JOL and testing with a re-study condition.

Moreover, there is little information about the effectiveness of these strategies for material of different degrees of difficulty. The relative difficulty of retrieving information at the time of testing has been shown to be beneficial for later retrieval (Pyc & Rawson, 2009; Roediger III & Karpicke, 2006b) and is known as the retrieval effort hypothesis, which postulates that the more effort put in at encoding, the greater the testing effect (Bjork, 1994; Carpenter, 2009; Pyc & Rawson, 2009). For this reason, we also manipulated the strength of the cue–target association. We assumed that a weaker cue–target association would lead participants to make more of an effort at encoding and would hence produce better learning efficiency.

The aim of this study was therefore to compare both delayed-JOL monitoring and testing with a re-study condition. If JOL monitoring involves an attempt to retrieve information before expressing a JOL, we should obtain memory performance similar to that obtained with testing. Also, insofar as testing repeatedly produces better performance than re-study, we predict that JOL monitoring should produce better performance than re-study. The strength of the cue–target association was strong associates and weak associates for the first experiment and weak associates and non-associates for the second experiment. To gain some insight into the quality of participants' memory performance assessment following each strategy, we also collected their own retrospective assessment of how confident they were of their answers after the final memory test.

In keeping with the known, long-term benefits of the testing effect (Roediger III & Butler, 2011; Toppino & Cohen, 2009), we opted for a retention interval of 48 hours.

Experiment 1

Material and methods

Participants

A total of 111 undergraduates at Clermont-Ferrand University were required to take part in this experiment as part of their course. They were randomly assigned to one of the three experimental conditions and tested in small groups of five or six.

Material

Items were chosen from a French database of word association norms established for 366 concrete object names (Ferrand & Alario, 1998). The items to be learned consisted of 60 word pairs, half with a strong association (rocket–space) with a 45.9% mean probability of producing the target when given the cue, and half with a weak association (lettuce–rabbit) with an average cue-to-target strength of 3.1%.

Procedure

Each participant was given a booklet containing the relevant pages about each condition: re-study, single delayed JOL, or testing. They were informed that the task included a total of 60 word pairs (cue–target) and instructed to study all the word pairs with a view to being able to recall the second word (target) when cued with the first (cue) in the final test, two days later. The order in which the word pairs appeared was randomised for each new presentation. The encoding phase was common to all three conditions. Each word pair was projected onto a screen for 5 seconds and, after a 1.5 seconds blank screen, was followed by the next pair. There was then a 4-minute interval which was filled with a non-verbal distraction task. In the re-study condition, word pairs were presented for a second time. In the testing and delayed-JOL conditions and also in the final test, each cue word was presented by itself for 5 seconds, with a 3 seconds blank screen separating one cue word from the next. The extra time (1.5 seconds) compared to a (re)learning condition was provided to allow participants to write down either the JOL (delayed-JOLs condition) or the recall attempt (testing condition). In the delayed-JOL condition, participants were asked to rate their JOLs on a scale of 0–100% with increments of 20%. Their rating was to reflect how confident they felt about recalling the second word of each pair when prompted with the first one. In the testing condition, they had to write down the target word if they knew it. In the final test, they were provided with the cue and asked to recall the target; then they were asked to judge how confident they were in each of the answers they had given.

Results and discussion

An alpha level of .05 was used. Effect sizes are denoted by partial eta squared (η_p^2) for analysis of variances (ANOVAs), Cohen's *d* (*d*) for the comparisons; a posteriori comparisons were done with corrected Bonferroni tests.

Recall

Below, we start by reporting the first recall for the testing condition and then the recall performances in the final tests for all conditions.

First recall for testing

A univariate ANOVA was carried out on the first recall performances in the Testing condition. The strong associates produced more correct answers ($M = 77.4$, SD = 17.2) than the weak associates ($M = 59.5$, SD = 21.3), $F(1, 66) = 14.5$, MSE = 376.8, $p < .001$, $\eta_p^2 = .18$.

Final test

A two-way ANOVA, Condition (3) × Association (2), was carried out on recall performance, with condition as a between-participant factor and association as a within-participant factor. The analysis yielded a significant effect for condition $F(2, 108) = 5.1$, MSE = 546.1, $p < .01$, $\eta_p^2 = .09$, for association, $F(1, 108) = 391.4$, MSE = 92.9, $p < .001$, $\eta_p^2 = .78$, as well as a significant interaction $F(1, 108) = 5.5$, MSE = 92.9, $p = .005$, $\eta_p^2 = .09$ (see Table 1). Participants gave a higher percentage of correct answers with strong associates ($M = 76.5$, SD = 14.5) than weak associates ($M = 50.9$, SD = 21.8). Regarding weak associates, the percentages of correct answers for the testing condition ($M = 58.3$, SD = 22.8) and delayed-JOL condition ($M = 53.5$, SD = 19.2), $d = .23$, were not different. However, both the delayed-JOL condition, $p < .001$, $d = .79$, and testing condition, $p < .05$, $d = .62$, yielded higher percentages of correct answers than the re-study condition ($M = 41.2$, SD = 20.6). There was no significant difference between conditions in terms of the correct answers with strong associates (testing, $M = 80.4$, SD = 15.8, delayed JOL, $M = 76.7$, SD = 13.2, and re-study, $M = 72.9$, SD = 14.1). So the memory performances following delayed JOL, and testing conditions produced better performances than the re-study condition, but only for items with weak cue-to-target association strength, that is, difficult items.

JOL and recall for monitoring condition

Mean predicted recall (JOL) and actual recall percentages as a function of strength of association were analysed in the delayed-JOL condition. In the procedure proposed by Koriat (1997), estimates and memory performance are treated as a repeated factor (labelled measure). A two-way ANOVA, Measure (JOL vs. recall) × Association (strong vs. weak), yielded $F(1, 39) = .2$, MSE = 216.3, $p = .7$, ns, $\eta_p^2 = .004$, for measure, $F(1, 39) = 220.6$, MSE = 10941.9, $p < .001$, $\eta_p^2 = .85$, for association, and $F(1, 39) = 41.7$, MSE = 1756.9, $p < .001$, $\eta_p^2 = .52$, for the interaction. JOLs were significantly higher ($M = 59.2$, SD = 15.4) than the percentage of correct answers ($M = 53.5$, SD = 19.2) for weak associates, whereas they were significantly lower ($M = 69.1$, SD = 13.7) than the percentage of correct answers ($M = 76.7$, SD = 13.2) for strong associates.

Confidence level in the different learning conditions relative to semantic association

A two-way ANOVA, Condition (3) × Association (2), on confidence level judgments yielded a main effect for condition, $F(2, 108) = 9.5$, MSE = 3923, $p < .001$, and for association, $F(1, 108) = 24$, MSE = 1351, $p < .001$. Relative to the testing condition ($M = 79.7$, SD = 13.0), participants were reliably less overconfident both for the delayed-JOL condition ($M = 70.7$, SD = 15.3) and for the re-study condition ($M = 64.9$, SD = 17.4), both $p < .001$. Participants were reliably more confident for strong associates ($M = 74.0$, SD = 15.5) than for low associates ($M = 69.0$, SD = 17.0). The interaction was not reliable, $F(2, 108) = 1.5$, MSE = 86, $p = .2$.

The results of Experiment 1 revealed that both delayed JOL and testing had an advantage over re-learning as learning strategies, but that this advantage was restricted to moderately difficult material and not observed for easy material. However, a confounding factor was the extra time given for the JOLs and testing. Hence, Experiment 2 attempts to replicate the findings of Experiment 1 while equating the encoding time across conditions. We expected that the JOL condition and testing condition would still outperform relearning. Moreover, we wondered whether the retrieval effort hypothesis might have some limitations, and what would happen in the presence of some additional difficulty, that is, with very difficult material, in other words non-associated word pairs (Carpenter, 2009). So difficulty was again manipulated, but weak- and non-associated pairs were used in Experiment 2.

Experiment 2

Material and methods

Participants

A total of 88 undergraduates at Clermont-Ferrand University were required to take part in this second experiment as part of their course. They were randomly assigned to one of the three experimental conditions.

Material

Weakly associated items were the same as in Experiment 1. The other half of the items consisted of the same first words of the strongly associated items from Experiment 1 but with a second word not associated with the first one.

Procedure

The procedure was the same as in Experiment 1 with one exception. The encoding phase, JOL phase (delayed-JOL condition), testing phase, and cued-recall phase were common to all three conditions; more specifically, each word pair or cue word was projected onto a screen for

Table 1. Percentage of final recall and confidence level for all learning conditions (Testing, Delayed JOL, Re-study) and for association with strong and weak associates (standard deviations in parentheses) for Experiments 1 and 2.

	Testing	Delayed JOL	Re-study
Experiment 1			
Correct answer			
Strong associates	80.4 (15.8)	76.7 (13.2)	72.9 (14.1)
Weak associates	58.3 (22.8)	53.5 (19.2)	41.2 (20.6)
Confidence level			
Strong associates	81.1 (13.4)	74.2 (13.2)	67.4 (17.1)
Weak associates	78.3 (12.6)	67.1 (16.6)	62.4 (17.6)
Experiment 2			
Correct answer			
Weak associates	45.2 (16.2)	43.3 (13.0)	30.3 (19.6)
Non-associates	14.4 (11.0)	12.6 (8.6)	9.4 (6.1)

Note: Insofar as participants gave few answers in Experiment 2, we cannot calculate confidence level means (for non-associated words participants averaged near 10% of answers, corresponding to only three items).

5 seconds and, after a 3 seconds blank screen, was followed by the next pair.

Results and discussion

Recall

First recall for testing

A univariate ANOVA was carried out on the first recall performances in the Testing condition. There were more correct answers with the weak associates ($M = 51.03$, SD = 16.2) than the non-associates ($M = 22.2$, SD = 13.25), $F(1, 56) = 55.1$, MSE = 219.2, $p < .001$, $\eta_p^2 = .49$

Final test

The two-way ANOVA, Condition (3) × Association (2), yielded $F(2, 85) = 6.8$, MSE = 253.5, $p = .002$, $\eta_p^2 = .14$, for condition, $F(1, 85) = 327.9$, MSE = 100.9, $p < .001$, $\eta_p^2 = .8$, for association, and $F(1, 85) = 4.9$, MSE = 100.9, $p = .009$, $\eta_p^2 = .1$, for the interaction. Participants gave a higher percentage of correct answers with weak associates ($M = 39.2$, SD = 17.8) than non-associates ($M = 12.0$, SD = 8.9). Regarding weak associates, there was no difference in the percentage of correct answers between the testing condition ($M = 45.2$, SD = 16.2) and delayed-JOL condition ($M = 43.3$, SD = 13.0), $d = .13$. However, both the delayed JOL, $p = .004$, $d = .78$, and testing, $p < .001$, $d = .82$, conditions produced a higher percentage of correct answers than the re-study condition ($M = 30.3$, SD = 19.6). There was no significant difference between the conditions in terms of correct answers with non-associates (testing, $M = 14.4$, SD = 11, delayed JOL, $M = 12.6$, SD = 8.6, and re-study, $M = 9.4$, SD = 6.1). So, here again, the delayed JOL and testing conditions outperformed the re-study condition, but only for moderately difficult pairs.

JOL and recall for monitoring condition

The two-way ANOVA, Measure (JOL vs. recall) × Association (weak vs. non-), yielded $F(1, 26) = 100.2$, MSE = 96.4, $p < .001$, $\eta_p^2 = .8$, for measure, $F(1, 26) = 219.4$, MSE = 109.5, $p < .001$, $\eta_p^2 = .89$, for association, and $F(1, 26) = .95$, MSE = 23.8, $p = .4$, ns, $\eta_p^2 = .03$, for the interaction. JOLs were significantly higher ($M = 46.9$, SD = 18.4) than the percentage of correct answers ($M = 25.6$, SD = 19.6).

Confidence level

Insofar as participants gave few answers in Experiment 2 (for non-associated words they averaged 10% of answers, corresponding to only three items), confidence levels were considered unreliable.

General discussion

The aim of this study was to use a very straightforward experimental paradigm to compare the learning efficiency of delayed JOL and testing with that of a re-study condition while manipulating the difficulty of the material to be learned. After an initial encoding phase with word pairs common to all conditions, participants were randomly assigned to one of the three conditions. In the delayed-JOL condition, they were asked to express an item-by-item JOL when prompted with the first word, in the testing condition, they had to retrieve the second word when prompted with the first one, and in the re-study condition they studied the material for a second time. The final memory assessment session took place 48 hours later. Experiment 1 compared strong with weak associates and Experiment 2 weak associates with non-associates. The main observation was that in both experiments the participants in the delayed-JOL and testing conditions outperformed the participants in the re-study condition. This was only the case, however, for items with weak cue-to-target association strength, that is, difficult items. No difference was observed between the three conditions as regards strong cue-to-target association strength, that is, easy items, or items without any association, that is, very difficult items. In these experiments then, the difference between the strategies emerges only when the item pairs are semantically related but only moderately so, such that some learning effort is required. The retrieval effort hypothesis (Carpenter, 2009; Pyc & Rawson, 2009) was initially proposed for the testing effect in terms of the degree of difficulty of the material to be learned/retrieved but it also seems to apply to the delayed-JOL condition, suggesting that both testing and delayed JOL are more effortful than relearning. These findings are consistent with the hypothesis that suggests that retrieval practice underlies both the delayed-JOL and testing effects. Studies of the testing effect as an overt retrieval practice provide another example of that. In some of the studies, overt retrieval was compared with covert retrieval during which participants are asked to retrieve the answer mentally in their heads without actually producing it; this is similar to the hypothesised retrieval attempt made before expressing a JOL. The studies revealed a strong testing effect for both overt and covert retrieval practices (Carpenter et al., 2008), but, in addition, recent studies have also found them to be equivalent in terms of their retention benefits (Putnam & Roediger III, 2013; Smith, Roediger III, & Karpicke, 2013). Recent neuroimagery data also support the incorporation of retrieval processes in JOLs (Do Lam et al., 2012).

This effort on the part of learners seems to be efficient; however, only when the items are of moderate difficulty since we observed no difference between the three conditions with respect to word pairs with strong cue-to-target association strength, that is, the easy items. Moreover, the efficiency of the difficulty introduced in a given learning task by a given learning technique seems to have its limitations, insofar as for the non-associated pairs (very difficult), the added difficulty introduced by the testing or delayed JOL did not increase learning efficiency relative to re-learning. So, the retrieval effort hypothesis seems to have some limitations, given that when some extrinsic difficulty (to make a JOL or to re-learn) was combined with an intrinsic high degree of difficulty (no association between cue and target) for the very difficult pairs the re-learning strategy did not show inferior performance. However, a possible floor effect cannot be ruled out, and one may ask whether additional JOL or testing trials might allow for better performances than additional re-learning sessions.

Accordingly, the specific efficiency patterns we observed with testing and delayed JOL could be the result of our manipulating cue-to-target strength, given that previous studies found no reliable difference between JOL and re-study (Sundqvist et al., 2012), or between JOL and testing (Jönsson et al., 2012).

To summarise, delayed JOL behaved in the same way as testing, and both yielded better performances than re-study when material was of moderate difficulty. As far as we know, this is the first time JOL monitoring has produced memory outcomes as efficient as testing when compared with the re-study condition and with respect to difficult material. A further interesting observation was that following a single learning phase a single JOL proved to be an efficient learning strategy, and one no different, statistically, from a single test session (Carpenter, 2009; Karpicke & Roediger III, 2007). For Experiment 1, JOL monitoring also produces a greater introspective sensitivity to difficulty. This condition could therefore make learners more aware of the relative difficulties of the material to be learned. Nevertheless, Kornell and Bjork (2008) observed that, using their JOL, participants could drop items from further re-study even if they felt they would not be able to retrieve it in a later test. Studies have reported that students adopt self-testing as a memory diagnostic tool rather than as a memory enhancement strategy (Hartwig & Dunlosky, 2012; Karpicke et al., 2009; Kornell & Son, 2009). Taken together, our results show that JOL monitoring shows potential as an alternative to testing. Since it does not involve overt testing, it might also be perceived as less stressful (Beilock, 2008; Ryan & Ryan, 2005; Schmader, Johns, & Forbes, 2008), and students might thus be more likely to adopt it as a learning strategy.

Acknowledgement

The authors would like to thank Sarah Ehi, Graziella Fafournoux, Ninon Hussong, Vanessa Jalicon, and Lolita Lorgeray for their help with data collection, and Gillian Wakenhut for correcting our English.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research project received support funding from the Institut National de la Santé et de la Recherche Médicale (INSERM), the University Hospital of Strasbourg, Strasbourg University, the Centre National de la Recherche Scientifique (CNRS), and the Agence Nationale de la Recherche TDE EMCO 2011, Blaise Pascal University.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger III, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*(7), 861–876. doi:10.1002/acp.1391

Beilock, S. L. (2008). Math performance in stressful situations. *Current Directions in Psychological Science*, *17*(5), 339–343. doi:10.1111/j.1467-8721.2008.00602.x

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). Cambridge: MIT Press.

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. doi:10.1037/a0017021

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, *36*(2), 438–448.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–642.

Do Lam, A. T. A., Axmacher, N., Fell, J., Staresina, B. P., Gauggel, S., Wagner, T., … Weis, S. (2012). Monitoring the mind: The neurocognitive correlates of metamemory. *PLoS ONE*, *7*(1), e30009. doi:10.1371/journal.pone.0030009

Dunlosky, J., Cavallini, E., Roth, H., McGuire, C. L., Vecchi, T., & Hertzog, C. (2007). Do self-monitoring interventions improve older adult learning? *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *62* (Spec No 1), 70–76.

Dunlosky, J., Hertzog, C., Kennedy, M. R. T., & Thiede, K. W. (2005). The self-monitoring approach for effective learning. *Cognitive Technology*, *10*, 4–11.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, *20*(4), 374–380. doi:10.3758/BF03210921

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning and comprehension: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. doi:10.1177/1529100612453266

Ferrand, L., & Alario, F.-X. (1998). Normes d'associations verbales pour 366 noms d'objets concrets. *L'année Psychologique*, *98*(4), 659–709. doi:10.3406/psy.1998.28564

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and sheduling related to achievement? *Psychonomic Bulletin & Review*, *19*, 126–134. doi:10.3758/s13423-011-0181-y

Jönsson, F. U., Hedner, M., & Olsson, M. J. (2012). The testing effect as a function of explicit testing instructions and judgments of learning. *Experimental Psychology (formerly Zeitschrift für Experimentelle Psychologie)*, *59*(5), 251–257. doi:10.1027/1618-3169/a000150

Karpicke, J. D., Butler, A. C., & Roediger III, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory (Hove, England)*, *17*(4), 471–479. doi:10.1080/09658210802647009

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

Kelemen, W. L., & Weaver III, C. A. (1997). Enhanced memory at delays: Why do judgments of learning improve over time? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(6), 1394–1409. doi:10.1037/0278-7393.23.6.1394

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. doi:10.1037/0096-3445.126.4.349

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits – and costs – of dropping flashcards. *Memory*, *16*(2), 125–136. doi:10.1080/09658210701763899

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*(5), 493–501.

Larsen, D. P., Butler, A. C., & Roediger III, H. L. (2008). Test-enhanced learning in medical education. *Medical Education*, *42*(10), 959–966. doi:10.1111/j.1365-2923.2008.03124.x

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The « delayed-JOL effect ». *Psychological Science*, *2*(4), 267–270. doi:10.1111/j.1467-9280.1991.tb00147.x

Putnam, A. L., & Roediger III, H. L. (2013). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition*, *41*(1), 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*(4), 437–447. doi:10.1016/j.jml.2009.01.004

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. doi:10.1016/j.tics.2010.09.003

Roediger III, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. doi:10.1111/j.1467-9280.2006.01693.x

Roediger III, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. doi:10.1111/j.1745-6916.2006.00012.x

Ryan, K. E., & Ryan, A. M. (2005). Psychological processes underlying stereotype threat and standardized math test performance. *Educational Psychologist*, *40*(1), 53–63. doi:10.1207/s15326985ep4001_4

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review*, *115*(2), 336–356. doi:10.1037/0033-295X.115.2.336

Smith, M. A., Roediger III, 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*, *39*(6), 1712–1725. doi:10.1037/a0033569

Sundqvist, M. L., Todorov, I., Kubik, V., & Jönsson, F. U. (2012). Study for now, but judge for later: Delayed judgments of learning promote long-term retention. *Scandinavian Journal of Psychology*, *53*(6), 450–454. doi:10.1111/j.1467-9450.2012.00968.x

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, *56*(4), 252–257. doi:10.1027/1618-3169.56.4.252

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory (Hove, England)*, *11*(6), 571–580. doi:10.1080/09658210244000414