Does Individual Performance Feedback Increase the Use of Retrieval Practice?

Luotong Hui¹ · Anique B. H. de Bruin¹ · Jeroen Donkers¹ · Jeroen J. G. van Merriënboer¹

Accepted: 1 February 2021 / Published online: 20 March 2021
© The Author(s) 2021

Abstract

The testing effect—the power of retrieval practice to enhance long-term knowledge retention more than restudying does—is a well-known phenomenon in learning. However, retrieval practice is hardly appreciated by students and underutilized when studying. One of the reasons is that learners usually do not experience immediate benefits of such practice which often present only after a delay. We therefore conducted 2 experiments to examine whether students choose retrieval practice more often as their learning strategy after having experienced its benefits. In Experiment 1, students received individual feedback about the extent to which their 7-day delayed test scores after retrieval practice differed from their test scores after restudy. Those students who had actually experienced the benefits of retrieval practice appreciated the strategy more and used it more often after receiving feedback. In Experiment 2, we compared the short-term and long-term effects on retrieval practice use of individual performance feedback and general instruction about the testing effect. Although both interventions enhanced its use in the short term, only the individual feedback led to enhanced use in the long term by those who had actually experienced its benefits, demonstrating the superiority of the individual feedback in terms of its ability to promote retrieval practice use.

Keywords Retrieval practice · Testing effect · Feedback · Mental effort · Perceived utility

Introduction

Retrieval practice has been confirmed to significantly enhance learning compared with repeated study (or “restudy”). This phenomenon is known as “the testing effect” (Carpenter,

✉ Luotong Hui
l.hui@maastrichtuniversity.nl

¹ Department of Educational Development & Research, School of Health Professions Education, Faculty of Health, Medicine and Life Sciences (FHML), Maastricht University, P.O. Box 616, 6200 MD Maastricht, the Netherlands

2009; Grimaldi & Karpicke, 2014; Rawson & Dunlosky, 2012; Roediger III & Karpicke, 2006a). The testing effect has been found not only in laboratory settings during the study of word pairs (Pyc & Rawson, 2010; Runquist, 1983), face-name pairs (Carpenter & DeLosh, 2005), symbols or features (Coppens et al., 2011; Jacoby et al., 2010), and text (Endres & Renkl, 2015) but also in real course settings across different age groups (Carpenter et al., 2017; McDaniel et al., 2007).

Although studies have shown that some students use retrieval practice, they mostly do so to assess their knowledge level *at the end* of a study phase rather than *during* this phase to maximize learning (Geller et al., 2018; Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007). Indeed, when making learning strategy decisions, students do not always prefer retrieval practice to restudy, despite the powerful memory benefits it can bring when used as a study tool (Ariel & Karpicke, 2018; Clark & Svinicki, 2015; Tullis et al., 2013). This is probably because learners base their learning strategy decisions on *perceived* learning (also called judgments of learning; JOLs), which can differ greatly from *actual* learning (Kirk-Johnson et al., 2019; Koriat, 1997; Schwartz et al., 1997). More specifically, as students cannot objectively gauge their future learning outcomes, they judge their learning through the subjective feelings generated by the learning strategy used to study the learning material (e.g., feelings of familiarity/fluency) (Bjork et al., 2013; Kirk-Johnson et al., 2019; Yan et al., 2016). Students do not experience immediate feelings of fluency from retrieval practice use, and its long-term memory benefits often present after a delay. Consequently, they do not feel that the strategy has enhanced their learning as much as it actually did (i.e., low JOLs after retrieval practice) (Bjork et al., 2013; Finn & Tauber, 2015). During restudy, on the other hand, students do experience immediate feelings of fluency which they mistake for actual learning, leading to the misconception that the strategy has enhanced their learning (i.e., high JOLs after restudy) (Bjork et al., 2013; Finn & Tauber, 2015). As a result of such erroneous retrieval practice beliefs and ensuing perceptions of reduced utility, learners are disinclined to use retrieval practice (Kirk-Johnson et al., 2019). This tendency is confirmed by the expectancy-value model, which predicts that perceived utility influences an individual's behavior to produce positive end states (Rokeach, 1973). Hence, reduced perceived utility may pose an additional barrier to retrieval practice use.

Over the last decade, researchers have actively promoted the use of effective learning strategies such as retrieval practice to enhance learning. However, rather than concentrating on perceived learning as its fundamental factor, these studies focused on how to change erroneous beliefs (see studies Donker et al., 2014; Hattie et al., 1996; Tullis et al., 2013). Since students did not actually experience the benefits of retrieval practice in these studies, it is doubtful whether they would use this strategy more frequently, even after changing their beliefs. This is probably because their perceived learning after retrieval practice is low, and this low perceived learning still dominates their decisions on retrieval practice.

According to Toppino et al. (2018), the underuse of retrieval practice could be down to a metacognitive bias in which learners do not adequately balance the risks of retrieval failure against the potential benefits of retrieval practice. If students experience benefits from retrieval practice use (i.e., the testing effect), for instance, by observing performance feedback shows that this strategy leads to better learning compared to restudy, they might use it more often (Einstein et al., 2012) because their actual learning affects their perceived learning. Similarly, students' actual learning could impact their beliefs in retrieval practice. In light of these considerations, we expect that students who have not experienced the benefit of retrieval practice will stick to their previous strategies. In the current study, experiencing the benefits of

retrieval practice means that (a) students profit from the testing effect and (b) are actually aware of this profit.

Although experiencing the benefit of retrieval practice may have the potential to enhance its use, only a handful of studies have been conducted to investigate this. In an effort to enhance retrieval practice use, DeWinstanley and Bjork (2004) asked students the following question about their actual learning after retrieval practice and restudy: “What did you notice about your performance on the previous memory test?” Some students reported results that were at odds with their actual learning outcomes, which demonstrate that they had not effectively experienced the benefits of retrieval practice. Nevertheless, students did become more engaged in retrieval practice after noticing the testing effect. In a similar fashion, Einstein et al. (2012) asked students to take a delayed test after retrieval practice and restudy and subsequently presented them with the test scores of the whole class for comparison. After they had explained the testing effect to students and gave them the average actual scores of the class which showed a testing effect, 82% of students reported that they were more or much more likely to use retrieval practice. Likewise, Dobson and Linderholm (2015) found that students who had been exposed to the testing effect and encouraged to use retrieval practice performed better on a subsequent test, supposedly because they applied more retrieval practice. Finally, Carpenter et al. (2017) exposed students in an introductory biology course to optional online review questions that could be studied by either retrieval practice or restudy. After students had learned that the average actual scores after retrieval practice were higher than the scores after restudy, the proportion of students choosing retrieval practice increased significantly in the following review phase.

None of the studies discussed above had students experience the benefit of retrieval practice by having them compare their personal learning outcomes after retrieval practice and after restudy, in order to increase their actual strategy use. We therefore conducted two experiments to investigate how the provision of feedback about actual learning outcomes, showing that retrieval practice outperforms restudy, affects learning strategy decisions, i.e., retrieval practice decisions. In Experiment 1, we measured whether students use retrieval practice more often after being given individual feedback about their actual learning outcomes showing that retrieval practice outperforms restudy. In Experiment 2, we investigated whether individual feedback about the testing effect has a stronger impact on retrieval practice decisions than general feedback about the testing effect. To our knowledge, this study is the first to promote the use of retrieval practice by letting students directly experience its benefits.

Experiment 1

The aim of Experiment 1 was twofold. First, we sought to investigate whether students would use retrieval practice more often after experiencing its benefits by receiving individual feedback about their actual learning outcomes showing that retrieval practice outperforms restudy. Our second aim was to explore the relationship between mental effort and the testing effect, because previous studies have been inconclusive in this regard. It is suggested that retrieval practice requires higher mental effort than restudy, in turn leading to better learning. Such enhanced effort, as mentioned above, might also prevent students from actually using the strategy (Karpicke & Roediger III, 2007; Kirk-Johnson et al., 2019; Rowland, 2014). In his review study, Rowland indeed argued that effortful processing contributed to the testing effect (2014). At the same time, Endres and Renkl (2015) found that retrieval practice did not require

any more investment from students than restudy did. With Experiment 1 we hoped to bring more clarity in this regard.

Participants studied half of the images of human anatomical structures with their Latin name using retrieval practice, the other half using restudy. In the subsequent study of new image-name pairs, participants were free to choose either one of these strategies. After a 7-day interval, a test was taken, and participants were informed of their performance after each learning strategy use. Finally, participants were again free to choose a strategy to study new image-name pairs. We hypothesized that:

1. Hypothesis 1: Students' scores for the delayed test would be higher after retrieval practice than after restudy (the testing effect).
2. Hypothesis 2: Students would have higher JOLs after restudy than after retrieval practice (JOLs).
3. Hypothesis 3: Students who had experienced the testing effect would significantly change their beliefs about retrieval practice (retrieval practice beliefs).
4. Hypothesis 4: Students who had experienced the testing effect would more often select retrieval practice compared to those who had not (retrieval practice decisions).

Method

Participants

Participants were 68 university students from nonmedical faculties (M age = 20.6 y, SD = 2.5, 56% female), 29 of whom had studied some human anatomy before university and 39 never had. Fifty-six students were non-native English speakers. Since 54 students had never studied Latin and 65 had never studied ancient Greek, the majority was not familiar with the anatomical names which are often closely related to Latin and/or ancient Greek.

Memory Materials

Seventy-two image-name pairs of human anatomical structures were used. Images were obtained from a human anatomy study website for medical students (www.anatomylearning.com). The structure names counted 7 to 10 letters. These 72 anatomical image-name pairs were organized into 18 units. Thus, each unit consists of four image-name pairs. In a restudying unit, these image-name pairs were shown in four consecutive rounds, and participants were asked to memorize them. In a retrieval practice unit, the image-name pairs were only shown in two consecutive rounds, followed by two rounds of retrieval practice (cued recall). The 18 units were displayed in the same order to all participants.

Design

We used a single-factor, within-subjects design. The single factor was “learning strategy” (retrieval practice vs. restudy). Each participant studied half of the units 1–10 in the restudy condition (restudy units), the other half in the retrieval practice condition (retrieval practice units), and the outcome variable was delayed recall performance after two conditions. For units 11–18, participants were free to choose one of the strategies, and the outcome variable was the percentage of retrieval practice choice.

Measures

Delayed Recall test As a measure of recall performance, we used the percentage of anatomical names answered correctly at the 7-day delayed test after restudy and retrieval practice.

Mental Effort We measured mental effort after each restudy unit and retrieval practice unit by asking the question: “You studied four image-name pairs by restudy/self-testing. How much mental effort did you invest from 1 to 9?”, using Paas’ 9-point rating scale (1 = very, very little mental effort, 9 = very, very much mental effort; see Fig. 1a) (Paas, 1992). We preferred the term “self-testing” in communicating with participants as this is more common language than “retrieval practice.” We consequently calculated the average mental effort that students had rated for restudy units and retrieval practice units.

JOLs Students judged their learning after each restudy unit and retrieval practice unit by answering the question: “How many pairs do you think you will still remember after one week?” (see Fig. 1b). We consequently calculated the average percentage of pairs that students had entered for restudy units and retrieval practice units.

Retrieval Practice Beliefs We measured students’ beliefs about retrieval practice by asking the question: “How effective is self-testing in helping you to memorize the anatomic image-name pairs from 1 (extremely ineffective) to 7 (extremely effective)?” (see Fig. 1c).

Retrieval Practice Decisions To measure students’ learning strategy decisions on retrieval practice, we asked the multiple-choice question: “Now that you have studied this item twice, do you want to: A. continue restudying or B. continue self-testing?” Students were free to choose a learning strategy to study the image-name pairs. We consequently calculated the percentage of retrieval practice choice.

Effort Beliefs We measured participants’ effort beliefs using Blackwell (2002) 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). This scale contains four positive items (e.g., “When something is hard, it just makes me want to work more on it, not less”) and five negative items (e.g., “If you’re not good at a subject, working hard won’t make you good at it”). To create a measure of positive effort beliefs, we reversely scored the negative items and merged them with the positive ones (McDonald’s $\omega_1 = .74$, $M = 4.96$, $SD = .66$).

Procedure

The procedure is depicted in Fig. 1d. The experiment consisted of two parts (day 1 and day 8) that were separated by a 7-day interval. Participants were seated in a computer room in groups of 5 to 15 students. They answered questions about demographics, knowledge of Greek and Latin, and prior knowledge of anatomy at the start of the first day. Then they were informed to study the names of human anatomical images and informed that they will be tested for image-name pairs they learned during each phase, and those who performed at the top 50% would receive a small additional financial compensation. Task difficulty and learning strategy sequence were counterbalanced. As shown in Fig. 1d, 50% of the participants started with

Fig. 1 **a** mental effort measure, **b** JOLs measure, **c** retrieval practice beliefs measure, **d** procedure of Experiment 1, **e** individual performance feedback for students who had benefited from retrieval practice. X and Y indicate their actual scores after restudy and retrieval practice respectively. $(Y-X)/8 \times 100$ shows how the score difference was calculated across two strategies. All those values were calculated automatically according to participants' delayed test performance. ME, mental effort, JOLs, judgments of learning

restudy condition for unit 1 followed by retrieval practice condition for the next unit, and another 50% started with retrieval practice condition for unit 1 followed by restudy condition.

Participants started with a learning strategy orientation phase in which they practiced in the restudy and retrieval practice conditions on units 1 and 2. As Fig. 2 shows, for a unit under the restudy condition (a restudy unit), four image-name pairs appeared for 8 s on the computer screen in a serial order. The presentation was repeated for four rounds. Participants completed a 15-s distracter activity after the second learning round for each unit, in which they were required to write down as many country names as possible that are located in different regions of the world (e.g., “Let’s play a game by writing down the names of the countries that surround the Arctic Ocean. You have 15 s to write down as many as you can. We will count the number

Fig. 2. Display of the Latin square across four learning rounds of a restudy unit (above) and a retrieval practice unit (below). All images were obtained from www.anatomylearning.com

of countries that you know”). The four image-name pairs in each unit were randomized in a Latin square across the four learning rounds. For a unit under the retrieval practice condition (a retrieval practice unit), four image-name pairs appeared for 8 s in a serial order for two repeated presentation rounds. Followed by a 15-s distracter, participants received two rounds of presentation of the images with the first letter of its anatomical name. Here, they had 8 s to complete the name of each image in learning rounds 3 and 4. No feedback was provided.

This orientation phase was followed by a learning strategy application phase in which participants studied eight units (units 3–10), four in the restudy condition and another four in the retrieval practice condition (see Fig. 1d). After each unit, we asked participants to rate their mental effort and provide JOL ratings. Upon completion of this application phase, we first administered the retrieval practice beliefs scale to participants, after which a 3-min distracter was shown in which they were asked to watch cartoon videos from the Boomerang UK YouTube channel. Participants then took an immediate test consisting of 16 image-name pairs in total (eight from the restudy units, two out of each unit; eight from the retrieval practice units, also two out of each unit). These pairs were randomly chosen from the units, and all participants received the same pairs. During this test, 16 images were displayed one by one together with the first letter of their name which participants had to complete. This test had no time limits, nor did we provide participants with information about the correct answer. The day ended with the first learning strategy decisions phase in which they studied new image-name pairs from units 11–14. Participants first studied unit 11. Four image-name pairs from this unit appeared for 8 s in a serial order for two repeated presentation rounds. After the second round of each image-name pair, participants were asked to select a strategy for continuing study by answering: “Now that you have studied this item twice, do you want to: A. continue restudying or B. continue self-testing?” Hence, after the 15-s distracter in learning rounds 3 and 4, they used the chosen strategy to study each pair. For example, if a participant chose “A. continue restudying” for an image-name pair, this image together with its name would be presented in learning rounds 3 and 4. If a participant chose “B. continue self-testing,” the image with the first letter of its name would be presented in learning rounds 3 and 4. He or she had 8 s to complete the name. This process was repeated until participants finished all units from this phase (units 12–14).

On day 8, participants first took a delayed test consisting of 16 image-name pairs that had not been used in the immediate test to measure their delayed recall performance. The procedure of this delayed test was the same as for the immediate test. We then provided them with individual feedback which showed their 7-day delayed test scores (their actual learning) pertinent to retrieval practice and restudy and how these scores differed across the two learning strategies used. Figure 1e shows feedback for students who had obtained a higher score after retrieval practice. For students who had obtained the same or a higher score after restudy, the last sentence of the feedback was changed to “However, based on scientific results, we still highly recommend that you use self-testing, because this strategy enhances long-term memory (>1 week) best.”

A second learning strategy decisions phase followed in which participants selected a strategy to study the image-name pairs of units 15–18. This phase was identical with the learning strategy decisions phase 1. As a last step, students completed the effort beliefs and again the retrieval practice beliefs scales. Data Analysis

We used SPSS, version 25, to analyze the data. To compare performance between the two learning strategies as well as the retrieval practice beliefs before and after the individual feedback, we ran paired-samples *t*-tests. A simple linear regression was used to predict the delayed test score based on participants’ invested mental effort. Finally, we performed repeated measures analyses of variance (ANOVAs) to compare the retrieval practice decisions before and after receiving feedback. The significance level was set at .05.

Results

We had to delete the 7-day delayed test data of five participants due to programming errors. In addition, we had to remove all JOLs data of five students because they typed in the incorrect answer to one or more questions (e.g., they typed in “e” rather than a number). Due to time conflicts or illness, three participants dropped out of the 7-day delayed test. See Table 1 for a summary of the results. Our preliminary analyses showed that the learning strategy order, knowledge of Greek and Latin, and prior knowledge of anatomy did not affect delayed test scores.

Testing Effect and Mental Effort

Hypothesis 1 predicted that students’ scores for the delayed test would be higher after retrieval practice than after restudy. A paired-samples *t*-test revealed that students indeed performed significantly better after retrieval practice than after restudy ($t(59) = 1.99, p = .05, d = .25$) (see Table 1), confirming the testing effect.

As mental effort may play a role in the testing effect, we performed an exploratory analysis to investigate the relationship between mental effort and the testing effect. We found that participants rated higher mental effort during retrieval practice than during restudy ($t(67) = 1.97, p = .05, d = .18$). There was no correlation between effort beliefs and mental effort during retrieval practice ($r = -.04, p = .76$) or restudy ($r = .06, p = .66$), which suggests that effort beliefs were not related to the amount of effort invested during the use of either strategy. From this, we conclude that the additional mental effort invested for retrieval practice was probably intrinsic to the strategy itself. However, a regression analysis showed that higher mental effort during retrieval practice did not predict a higher test score ($F(1, 58) = .10, p = .75$), which suggests that mental effort alone did not explain the testing effect.

JOLs

Hypothesis 2 predicted that students would have higher JOLs after restudy than after retrieval practice. We found that participants’ JOLs were the same across both strategies ($t(62) = .78, p = .44$).

Table 1 Means (and standard deviations) of the variables measured in Experiment 1

	Retrieval practice	Restudy	<i>t</i>	<i>p</i>
Initial learning performance ($n = 68$)	64.15 (24.26)			
Immediate test performance ($n = 68$)	47.61 (27.43)	38.97 (29.97)	2.92**	.01
Delayed test performance ($n = 60$)	31.67 (22.72)	26.25 (21.43)	1.99*	.05
Mental effort ($n = 68$)	5.94 (1.35)	5.68 (1.51)	1.97*	.05
JOLs ($n = 63$)	48.51 (22.67)	50.10 (24.45)	-.78	.44

Performance and JOLs are expressed as percentages. Initial learning performance indicates the test performance when participants were tested for the first time when learning the image-name pairs from retrieval practice units (which means learning round 3 in this study)

JOLs = judgments of learning

* $p \leq .05$, ** $p \leq .01$

Retrieval Practice Beliefs

Hypothesis 3 predicted that students who had experienced the testing effect would significantly change their beliefs about retrieval practice. While 25 students obtained a higher score after retrieval practice (retrieval benefit), the scores of 21 students remained the same after retrieval practice and after restudy (same score), and yet another 14 students obtained a higher score after restudy (restudy benefit). Figure 3a displays participants' retrieval practice beliefs before and after individual feedback, illustrating that students who had benefited from retrieval practice positively changed their retrieval practice beliefs ($t(24) = 4.26, p < .001, d = .56$). This was not the case for students who had obtained the same score ($t(20) = 1.43, p = .17$), nor for those who had benefited from restudy ($t(13) = -.69, p = .50$).

Retrieval Practice Decisions

Hypothesis 4 predicted that students who had experienced the testing effect would more often select retrieval practice compared to those who had not. Figure 3b depicts participants' retrieval practice choice before and after receiving the individual feedback. We grouped participants based on the strategy they benefited from. Consequently, we found an interaction effect between the groups and retrieval practice decisions that was statistically significant ($F(2, 57) = 8.42, p < .001, \eta_p^2 = .23$). More specifically, the retrieval-benefit group revealed a significant increase in retrieval practice choice after receiving feedback ($p < .01$); the same-score group showed no difference in strategy choice between the two measurements ($p = .90$); and the restudy-benefit group significantly reduced retrieval practice choice after feedback ($p = .01$), meaning that they more often chose restudy.

Before receiving individual feedback (day 1), all of the participants were expected to choose the same amount of retrieval practice. However, Fig. 3b indicates that the restudy-benefit group chose more retrieval practice than the others at day 1. Considering the small sample size of this group, it could be due to one or two students not giving good faith responses which lead to this unstable result. There is a big confidence interval for this group, 95% CI [56.13, 89.40] (see Supplementary information).

Fig. 3. **a** Participants' retrieval practice beliefs rated on a 7-point Likert scale before (day 1) and after (day 8) individual feedback. **b** The percentage of retrieval practice (RP) choice of each participant group (retrieval-benefit, same-score, and restudy-benefit group) before (day 1) and after (day 8) individual feedback

Discussion

With the overarching aim to promote retrieval practice use, Experiment 1 served two main purposes: (1) to investigate the effect of experiencing the benefit of retrieval practice by receiving individual performance feedback showing retrieval practice outperforms restudy and (2) to explore the relationship between mental effort and the testing effect. Our results indeed showed that retrieval practice has a benefit compared to restudy, confirming Hypothesis 1. We also found that participants invested more mental effort during retrieval practice than during restudy. Yet, as in Endres and Renkl's study (2015), in our experiment mental effort alone could not explain the testing effect. These results contradict Rowland's previous finding that effortful processing during retrieval practice contributed to the testing effect (2014). Considering our small sample size, which resulted in an even smaller retrieval-benefit group ($n = 25$), perhaps a larger sample would have produced more stable results.

Another finding is that we could not confirm Hypothesis 2 predicting that students would have higher JOLs after restudy than after retrieval practice. This seems to conflict with the results of previous studies in which restudy led to feelings of fluency (Kornell & Son, 2009; Roediger III & Karpicke, 2006b). We will therefore examine this more in detail in Experiment 2.

Next, our results confirmed Hypothesis 3 predicting that students who had experienced the testing effect would significantly change their retrieval practice beliefs. Students rated retrieval practice as more effective after they had experienced the testing effect, suggesting that the feedback on actual learning outcomes showing the testing effect significantly impacted not only students' retrieval practice decisions but also their beliefs.

Moreover, we found that, after receiving feedback on actual learning outcomes for retrieval practice and restudy, students who had experienced the testing effect chose retrieval practice more often than those who had not benefited from retrieval practice, confirming Hypothesis 4. Students who had obtained the same or a higher score after restudy, however, still continued to use their previous strategy and chose restudy more often, even though we explained the advantage of retrieval practice in the individual feedback. Hence, consistent with previous studies (DeWinstanley & Bjork, 2004; Einstein et al., 2012), our findings suggest that individual performance feedback significantly impacts students' retrieval practice decisions, *provided* they actually benefited from the strategy. Without intervention, students intend to use retrieval practice only when they are certain they will answer correctly (Kornell & Bjork, 2007; Vaughn & Kornell, 2019). In our study, almost half of the participants benefited from retrieval practice even without knowing whether their answer was correct or not during retrieval practice. Although learning strategy decisions were made on the item-level and the individual performance feedback was global (i.e., across all items after each strategy), this feedback may make students aware of the global mismatch between their perceived learning and actual learning. This global awareness may have influenced decisions on the item-level, which means learners are more inclined to test themselves also on items where they were uncertain of the correct answer. This may also explain why individual performance feedback increased retrieval practice choice only for those who experienced the benefit of retrieval practice.

Nevertheless, a study by Ariel and Karpicke (2018) showed that students may not need to experience the actual benefits of retrieval practice to increase its use. In their study, general instruction stressing the mnemonic benefits of retrieval practice over restudy promoted the use of retrieval practice, even in the long term. Indeed, according to Donker et al. (2014), such general instruction about why and how to use certain learning strategies is a prerequisite for

applying learning strategies. However, we agree with Fishbein and Ajzen (1975) that having this knowledge does not warrant its application in real life. It seems that Ariel and Karpicke's findings do not agree with Fishbein and Ajzen (1975), nor with our results from Experiment 1. We started from the assumption that individual feedback would be more promising than general instruction because the former might enhance students' perceived utility of retrieval practice by demonstrating its actual utility. At the same time, with knowledge gained from the general instruction, university students could still consider retrieval practice as not useful to them, since "knowledge" can be interpreted in many ways (Schommer, 1990). As the perceived utility of a certain learning strategy influences a shift to this strategy (Berger & Karabenick, 2011; Escribe & Huet, 2005), general instruction may not be sufficiently powerful to impact retrieval practice decisions. In conclusion, these findings seem to suggest that individual feedback may be more powerful than general instruction in influencing future retrieval practice decisions.

Experiment 2

In Experiment 2, we aimed to measure two interventions, specifically the provision of individual performance feedback showing the actual testing effect and general instruction about the testing effect (hereinafter referred to as "general feedback"), in terms of their short-term and long-term effect on increasing retrieval practice choice. Since we wanted to investigate the long-term effects as well, Experiment 2 spanned 3 days that were again separated by a 7-day interval (day 1, day 8, and day 15). The rest of the procedure was roughly similar to that of Experiment 1, the only difference being that on day 8, after the delayed test, we divided participants into two groups: group 1 received feedback on their individual performance; group 2 received general feedback explaining how retrieval practice was superior to restudy. The learning strategy decisions procedure continued as in Experiment 1 and was repeated on day 15. In addition to retesting Hypotheses 1 and 2 from Experiment 1, we hypothesized that:

1. Hypothesis 5: Individual feedback would promote the use of retrieval practice among students who had experienced the testing effect, in both the short and long term, and that the general feedback would not (retrieval practice decisions).
2. Hypothesis 6: Compared with the general-feedback group, participants from the individual-feedback group who had experienced the testing effect would consider retrieval practice as more useful (perceived utility).

Method

Participants

Participants were 124 university students (M age = 20.3 y, SD = 2.6, 71% female). We determined the sample size based on the effect size of the testing effect from Experiment 1 (d = .25), meaning that we needed 101 students to achieve a power of at least .80. We oversampled since we expected a high dropout rate due to the long duration of the study (3 weeks). The demographic structure was similar to that of Experiment 1.

Memory Materials

In addition to the 18 units of Experiment 1, we used another four units in Experiment 2 to measure participants' learning strategy decisions on day 15.

Design

We used a mixed two-factor design for Experiment 2, combining a between-subjects factor (individual vs. general feedback) with the same within-subjects factor as in Experiment 1 (retrieval practice vs. restudy). This means that participants either received individual feedback or general feedback. Similar to Experiment 1, for the within-subjects factor, each participant studied half of units 1–10 in the restudy condition, the other half in the retrieval practice condition. For units 11–18 and the newly added four units on day 15, participants were free to choose one of the strategies.

Measures

Retrieval Practice Decisions We used the same method as in Experiment 1 to measure participants' retrieval practice decisions in the short (day 8) and long (day 15) term. We again calculated the percentage of retrieval practice choice on these respective days.

Perceived Utility Instead of measuring participants' retrieval practice beliefs on days 1 and 8, we measured their perceived utility on days 1 and 15 by asking the question “To what extent do you think self-testing is useful in helping you memorize the anatomic image-name pairs?” to be scored on a 7-point rating scale (1 = extremely useless, 7 = extremely useful).

The other variables were measured in the same way as described in Experiment 1.

Procedure

As mentioned, the experiment spanned 3 days: day 1, day 8, and day 15. On day 1, the procedure was identical to day 1 of Experiment 1, except that participants rated perceived utility rather than retrieval practice beliefs. On day 8, all participants took a delayed test, after which they were divided into two groups: individual-feedback group and general-feedback group. As in Experiment 1, the individual-feedback group received individual performance feedback, whereas the general-feedback group received general feedback. The individual feedback was identical to the feedback in Experiment 1, the only difference being that the last sentence now read “because this strategy enhances long-term memory (>2 days) best” (changed from “because this strategy enhances long-term memory [>1 week] best”). The general feedback differed from the individual feedback in that it presented the average rate at which retrieval practice was found superior to restudy, as calculated from the studies by Grimaldi and Karpicke (2014) and Roediger III and Karpicke (2006b), instead of providing students' actual learning outcomes. Figure 4 details the general feedback that the general-feedback group received. The rest of day 8 was the same as day 8 of Experiment 1. On day 15, we measured performance on the image-name pairs studied on day 8 (this performance will not be reported). Then, the two groups completed another learning strategy decisions phase with four new units. Similar to the learning strategy decisions phase 1 and 2 in Experiment 1,

Fig. 4. General feedback in Experiment 2

participants were asked to select a strategy for each pair for continuing study after learning the second round of each image-name pair. Hence, in learning rounds 3 and 4, they used the chosen strategy to study each pair. The experiment ended with participants rating the perceived utility scale again.

Data Analysis

The data analysis methods were similar to those of Experiment 1. We used paired-samples t -tests to analyze how each feedback condition influenced students' decisions in the short term and long term compared to their decisions at the baseline of day 1. Repeated measures ANOVAs were used to measure retrieval practice utility as perceived by the general-feedback group and the participants from the individual-feedback group who had actually experienced the testing effect.

Results

Three participants dropped out of the day 8 measurement and another 3 dropped out of the day 15 measurement. We removed all JOL data of one participant because this participant gave incorrect answers to the JOLs-related questions. See Table 2 for a summary of the results.

Testing Effect and Mental Effort

As in Experiment 1, Hypothesis 1 predicted the testing effect which was confirmed again ($t(120) = 3.53, p < .001, d = .36$). There was no significant difference in invested mental effort between retrieval practice and restudy ($t(123) = .80, p = .42$), which was not consistent with the results from Experiment 1. Furthermore, mental effort during retrieval practice could not predict the test score after retrieval practice ($F(1, 179) = 1.10, p = .30$). These results again suggest that effortful processing does not directly contribute to the testing effect.

Table 2 Means (and standard deviations) of the variables measured in Experiment 2

	Retrieval practice	Restudy	<i>t</i>	<i>p</i>
Initial learning performance (<i>n</i> = 124)	60.84 (22.50)			
Immediate test performance (<i>n</i> = 124)	41.23 (26.34)	35.28 (24.18)	2.86**	.01
Delayed test performance (<i>n</i> = 121)	29.75 (21.67)	22.42 (19.02)	3.53***	.00
Mental effort (<i>n</i> = 124)	5.83 (1.44)	5.76 (1.45)	.80	.42
JOLs (<i>n</i> = 123)	44.44 (21.04)	51.45 (21.47)	-5.25***	.00
Perceived utility (retrieval benefit, <i>n</i> = 31)				
Before feedback	5.55 (1.31)	}	.71	.48
After feedback	5.74 (1.15)			
Perceived utility (restudy benefit, <i>n</i> = 13)				
Before feedback	5.92 (.90)	}	-2.80*	.02
After feedback	5.08 (1.51)			
Perceived utility (same score, <i>n</i> = 17)				
Before feedback	5.53 (1.51)	}	-1.29	.22
After feedback	5.06 (1.68)			
Perceived utility (general feedback, <i>n</i> = 60)				
Before feedback	5.81 (1.28)	}	-2.21*	.03
After feedback	5.39 (1.67)			

Performance and JOLs are expressed as percentages. Initial learning performance indicates the test performance when participants were tested for the first time when learning the image-name pairs from retrieval practice units (which means learning round 3 in this study). JOLs = Judgments of Learning; retrieval benefit = students whose scores were higher after retrieval practice than after restudy; restudy benefit = higher scores after restudy.

* $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

JOLs

As in Experiment 1, Hypothesis 2 predicted that students would provide higher JOLs after restudy than after retrieval practice. We found that students had significantly higher JOLs after restudy than after retrieval practice ($t(122) = 5.25, p < .001, d = .33$), which was not consistent with the results from Experiment 1.

Retrieval Practice Decisions

Hypothesis 5 predicted that individual feedback would promote retrieval practice use among students who had experienced the testing effect, in both the short and long term, and that general feedback would not. In the group that received individual feedback, 31 students had gained a retrieval benefit, 17 students had obtained the same score, and 13 students had gained a restudy benefit. Based on these results, we divided students into three corresponding subgroups: a retrieval-benefit, same-score, and restudy-benefit group, respectively. Figure 5 depicts the average percentage of retrieval practice choice at baseline (day 1) and in the short term (day 8) and the long term (day 15) by students from the general-feedback group and the three individual-feedback subgroups. In the short term, both the general-feedback group ($t(59) = 2.43, p = .02, d = .29$) and the retrieval-benefit group ($t(30) = 2.26, p = .03, d = .34$) chose retrieval practice more often on day 8 compared to day 1 at baseline. In the long term, however, only the retrieval-benefit group chose retrieval practice more often on day 15 compared to the baseline ($t(30) = 2.49, p = .02, d = .44$), which was not the case for the general-feedback group ($t(58) = 1.80, p = .08$). More specifically, the percentage of retrieval practice choice on day 15 slightly dropped for the general-feedback group, while for the retrieval-benefit group it slightly increased, although there was no statistical difference in retrieval practice choice on day 15 between these two groups ($t(88) = .68, p = .50$). Hence, our

Fig. 5. The percentage of retrieval practice (RP) choice of each participant group (retrieval-benefit, same-score, restudy-benefit group, and gen. FB) before (day 1) and after the two types of feedback in both the short (day 8) and long (day 15) term. “Gen. FB,” general-feedback group

results could not confirm the short-term effect predicted by Hypothesis 5, but they did confirm the predicted long-term effect.

For the same-score group, we found that neither in the short ($t(16) = .23, p = .82$) nor in the long term ($t(15) = -.16, p = .88$) the feedback affected retrieval practice decisions. For the restudy-benefit group, the feedback did not affect retrieval practice decisions in the short ($t(12) = .00, p = 1$) or in the long ($t(11) = .28, p = .79$) term. The short-term effect was not consistent with our finding from Experiment 1 that participants who had benefited from restudy chose restudy more often on day 8. Given the small sample of the restudy-benefit group in both experiments, it is difficult to reach a unified conclusion although participants who benefited from restudy might have displayed a tendency to choose restudy more often. Note that individual feedback in general did not have an impact on retrieval practice decisions. When students were not split up into the three subgroups, individual feedback did not influence retrieval practice decisions both in the short term ($t(60) = 1.67, p = .10$) and in the long term ($t(58) = 1.70, p = .09$). Moreover, when the individual-feedback group was not split up according to learning strategy benefit, there was no difference between the two types of feedback in increasing the retrieval practice choice in the short term ($t(119) = .65, p = .52$) or the long term ($t(116) = .04, p = .97$).

Perceived Utility

Repeated measures ANOVAs did not confirm Hypothesis 6, which predicted that participants from the individual-feedback group who had experienced the testing effect would consider retrieval practice as more useful than would those from the general-feedback group, ($F(1, 88) = 3.45, p = .07$), although there was a tendency in the expected direction. We explored the perceived utility rating for restudy-benefit and same-score groups. The results showed that the utility of retrieval practice as perceived by the restudy-benefit participants decreased significantly before and after the individual feedback ($t(11) = -2.80, p = .02, d = -.68$). In the same-score group, by contrast, perceived utility did not change before or after the individual feedback ($t(16) = -1.29, p = .22$). Combined, these results suggest that feedback on individual performance impacted perceived utility but only to a limited extent.

In Experiment 2, we measured whether and how the provision of individual or general feedback about the benefits of retrieval practice would impact the use of retrieval practice in the short and long term. Additionally, we measured JOLs and the relationship between the testing effect and mental effort. We confirmed both Hypothesis 1 that retrieval practice would outperform restudy and Hypothesis 2 that students would provide higher JOLs after restudy than after retrieval practice, which we will discuss in the next section.

We also found that both types of feedback inspired students from general-feedback and retrieval-benefit groups to engage in retrieval practice more often in the short term, partially dismissing Hypothesis 5. Not surprisingly, however, only individual feedback led to an enhanced use of retrieval practice in the long term, provided that participants had actually experienced the testing effect. More details will be discussed in the next section.

Furthermore, Hypothesis 6 was not confirmed, which predicted that participants from the individual-feedback group who had experienced the testing effect would consider retrieval practice as more useful than would the general-feedback group. Since the general-feedback group who had not received any actual performance feedback gave lower utility ratings on day 15 than day 1, we might presume that the image-name pairs were too difficult. Students might have thought that whichever strategy they chose, it would not help them learn. This could also

explain why students from the individual-feedback group who had experienced the testing effect did not give significantly higher utility ratings. We therefore suggest to investigate in future research how individual performance feedback impacts perceived utility and subsequent retrieval practice decisions by choosing tasks of medium difficulty. Although individual performance feedback did not significantly impact perceived utility, students from the individual-feedback group who had experienced the testing effect did choose retrieval practice more often. Hence, contrary to Rokeach's finding (1973) that perceived utility influences an individual's behavior, our results suggest that being informed about the results of a certain behavior (i.e., higher scores after retrieval practice) has a stronger influence on subsequent behavior (i.e., choosing retrieval practice) than perceived utility has.

General Discussion

Across two experiments, the present study investigated how the provision of individual performance feedback showing that retrieval practice outperforms restudy affects students' retrieval practice decisions. We found such individual feedback to significantly impact students' retrieval practice decisions in both the short and long term, provided they actually benefited from retrieval practice. In addition, this study confirmed the testing effect, by demonstrating the mnemonic advantage of retrieval practice as predicted by Hypothesis 1 in both Experiments 1 and 2.

Unlike previous studies that involved images (e.g., Carpenter, 2009; Coppens et al., 2011; Jacoby et al., 2010) or word pairs (e.g., Ariel & Karpicke, 2018), however, we did not obtain a large effect size: across both experiments, only about half of the students benefited from retrieval practice. A potential reason could be that, in order to control for prior knowledge, we selected students who were not familiar with human anatomy. Although low prior knowledge is an advantage in certain learning conditions (Kalyuga, 2007; Mayer & Gallini, 1990), little prior knowledge may have constrained students' elaborative semantic processing during study, in turn stifling the benefits of retrieval practice (Carpenter et al., 2016). Lack of prior knowledge may also increase task difficulty, further hindering students benefit from retrieval practice. In our study, students' initial learning was lower than 65% across two experiments (see the initial learning performance in Tables 1 and 2), which indicates that the task they received was, indeed, difficult. In his meta-analysis (2014), Rowland explained that the reliable testing effect was found if students' initial learning is in between 50 and 75%, and greater retrieval benefit if it is higher than 75%, otherwise students may benefit more from restudy. This also explains why about 26% of the students benefited from restudy, a finding that is echoed in previous studies.

Effortful processing is a contested explanation for the testing effect (Roediger III & Karpicke, 2006a; Rowland, 2014). In Kirk-Johnson and colleagues' study (2019), for instance, participants reported that retrieval practice was effort-demanding. However, our data did not provide a strong indication that participants *invest* more mental effort during retrieval practice than during restudy. These findings echo those of Endres and Renkl (2015), who also measured invested mental effort. Our argument in the "Introduction" section stated that retrieval practice *requires* more mental effort. The way to phrase the mental effort question is important because students respond differently when asking them for the effort "invested" or the effort "required" (Koriat et al., 2014). We therefore invite future researchers to rephrase the mental effort scale or to measure mental effort objectively by collecting physiological data, for

instance, using eye trackers. This may provide a better basis for investigating the relationship between mental effort and the testing effect.

We were able to confirm Hypothesis 2 which predicted that students would have higher JOLs after restudy than after retrieval practice in Experiment 2, but not in Experiment 1. This inconsistency may be due to the fact that we asked participants how many pairs out of four they would remember, instead of a percentage like most other studies did. The measurement (0 to 4) may make it difficult to distinguish the difference between the two strategies. Despite the potential difficulty, we could still confirm Hypothesis 2 after increasing the sample size in Experiment 2. The finding from Experiment 2 resonates with previous studies about feelings of fluency during restudy (Ariel & Karpicke, 2018; Bjork et al., 2013; Kornell & Son, 2009; Roediger III & Karpicke, 2006b). Learners mistook these feelings for “learning” causing them to have higher perceptions of learning after restudy than after retrieval practice, and they are not aware of this, in turn hindering the future use of retrieval practice (Kirk-Johnson et al., 2019). Providing feedback about the perceived learning as well as the actual learning may make students realize the mismatching of the perceived learning and the actual learning. Showing this mismatching may have a stronger influence on retrieval practice decisions than only showing the actual learning as we did. This could therefore be an interesting topic of future studies.

The results from Experiment 2 showed that both individual and general feedback caused students (i.e., general-feedback and retrieval-benefit groups) to choose retrieval practice more often in the short term. In the long term, however, only individual feedback made students engage in retrieval practice more often, provided they had actually experienced the testing effect, confirming Hypothesis 5 about the long-term impact of individual feedback on retrieval practice decisions. As previous studies (Carpenter et al., 2017; DeWinstanley & Bjork, 2004; Dobson & Linderholm, 2015; Einstein et al., 2012) were already keen to point out, learners use a certain strategy more often if they have noticed the actual benefits of that strategy. Our results indeed suggest that experiencing the testing effect, by observing that retrieval practice outperforms restudy, significantly impacts subsequent retrieval practice decisions in both the short and long term. The individual feedback may have opened students’ eyes to the fact that their actual learning did not match their perceived learning (Dobson & Linderholm, 2015), in turn inducing them to base learning strategy decisions on their objective rather than subjective learning (Kirk-Johnson et al., 2019).

However, our results also showed that individual feedback did not have an impact on improving students’ choice for retrieval practice, when including students who had benefited from restudy and students who had the same score. Considering only about half of the students from the individual-feedback group experienced the testing effect, this result could have been expected. With more students benefitting from retrieval practice like most of the memory studies showed, we would expect that the individual feedback will have a stronger impact than general feedback on improving the use of retrieval practice. Moreover, our study only showed students individual feedback once. The results of the general-feedback group implicate that the restudy-benefit and same-score groups would benefit more from the general feedback, at least in the short term. However, repeated reinforcement is needed to change behavior (Bai & Podlesnik, 2017). With more students *repeatedly* seeing that they obtain higher scores after retrieval practice than after restudy, individual feedback may show its superiority. Comparing the effect on individual and general feedback across multiple sessions could be an interesting direction for future studies. Moreover, studies show that students seek feedback about correct answers after retrieval practice (e.g., Dunlosky & Rawson, 2015). This implies that providing

feedback about correct answers may make retrieval practice more appealing. Future studies could combine individual performance feedback and feedback about correct answers to explore a more effective learning strategy intervention.

As mentioned, the general feedback was not able to promote the use of retrieval practice in the long term. Worse still, the group receiving general feedback in Experiment 2 considered retrieval practice even less useful after having tried the strategy. Since this group had only received instruction about the testing effect and no information about their actual test results, they may have appreciated that retrieval practice is an effective learning strategy and tested its use in the short term. When reconsidering its utility after a while, they may have concluded that retrieval practice was not applicable to them, because in absence of objective learning results they based their judgments on perceived learning. As a result, these students kept their original poor perceptions of the strategy's utility, reverting to their previous strategy in the long term. This finding that general feedback did not impact retrieval practice decisions in the long term does not agree with a study by Ariel and Karpicke (2018). In this study, the instruction explained that students should successfully retrieve each Lithuanian-English pair at least three times before the final test, which was administered 45 min after initial learning. The clearly quantified times of using retrieval practice (three times) and the relatively short interval between the initial learning phase and the final test (45 min) may have directed and motivated students to use retrieval practice. These features might have caused a long-term effect in use of retrieval practice, which were not present in our study. However, the final test was conducted only 45 min after the initial learning in this study. Since spaced retrieval practice works better than crammed retrieval practice, it will be interesting to measure the effect of Ariel and Karpicke's general instruction when there is a longer time interval between the initial learning and the final test (e.g., a 7-day interval).

Our study has some limitations. First, the image-name pairs used were difficult to memorize. The fact that a relatively large group of students did not experience the testing effect could be a result of this. More importantly, it may also play an essential role in influencing initial learning and further influencing retrieval practice decisions. However, to make our two experiments comparable, we did not change tasks when we conducted Experiment 2. We suggest that future studies should investigate how task characteristics and initial learning impact learning strategy decisions, by manipulating task difficulty and by using multiple learning strategy decision rounds. Second, as stated before, the way we phrased the mental effort question might have misled participants to judge the effort required by retrieval practice and restudy. We therefore welcome rephrasing the mental effort question or the use of alternative methods, such as the ones previously suggested, to measure mental effort in future studies. Third, all participants received either individual or general feedback across the two experiments. We did not compare results to a no feedback group. The task experience from using retrieval practice and restudy may increase the choice of retrieval practice even without receiving feedback, although previous research in this area (e.g., Karpicke, 2009; Kirk-Johnson et al., 2019) shows that this is not a likely explanation. These studies show that students perceive retrieval practice as less effective than restudy as stated in the "Introduction" section. However, to measure the extent of the performance feedback effect, it is advisable to include a control group without feedback in future research.

Our study was the first to use individual performance feedback that showed the benefits of retrieval practice to change learning behavior. Moreover, such feedback showed significant promise as a tool to promote retrieval practice use among students who had benefited from this strategy. We found that experiencing the benefit of retrieval practice by witnessing its actual

learning effects significantly enhanced its use. Educational practitioners seeking to promote the use of certain strategies may want to make their students directly experience the benefits of these strategies as this has been proven to be an effective intervention.

Conclusion

The present study was the first to underscore the value of individual performance feedback showing the benefits of retrieval practice as a tool to promote its use. Experiencing the testing effect, by receiving individual performance feedback showing that retrieval practice outperforms restudy, impacts students' subsequent retrieval practice decisions, not only in the short but also in the long term. Rather than being impacted by the general instructions we give them, students will choose the learning strategy they benefited from.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10648-021-09604-x>.

Acknowledgements The authors would like to thank Dr. Wisnu Wiradhany for his invaluable feedback on the “Method” section during the revision of this manuscript.

Funding The research project was funded by the China Scholarship Council (File No. 201708120059).

Declarations

Ethics Approval All procedures were approved by the Health, Medicine and Life Sciences Ethics Review Committee of Maastricht University (approval number: FHML-REC/2019/006(2)).

Consent to Participate Informed consent was obtained from all individual participants included in the study.

Conflict of Interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ariel, R., & Karpicke, J. D. (2018). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied*, 24(1), 43–56.
- Bai, J. Y., & Podlesnik, C. A. (2017). No impact of repeated extinction exposures on operant responding maintained by different reinforcer rates. *Behavioural processes*, 138, 29–33.
- Berger, J.-L., & Karabenick, S. A. (2011). Motivation and students' use of learning strategies: Evidence of unidirectional effects in mathematics classrooms. *Learning and Instruction*, 21(3), 416–428.

- Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual review of psychology*, 64(1), 417–444. <https://doi.org/10.1146/annurev-psych-113011-143823>.
- 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.
- Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 19(5), 619–636.
- Carpenter, S. K., Lund, T. J., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review*, 28(2), 353–375.
- Carpenter, S. K., Rahman, S., Lund, T. J., Armstrong, P. I., Lamm, M. H., Reason, R. D., & Coffman, C. R. (2017). Students' use of optional online reviews and its relationship to summative assessment outcomes in introductory biology. *CBE—Life Sciences Education*, 16(2), ar23.
- Clark, D. A., & Svinicki, M. (2015). The effect of retrieval on post-task enjoyment of studying. *Educational Psychology Review*, 27(1), 51–67.
- Coppens, L. C., Verkoeijen, P. P., & Rikers, R. M. (2011). Learning Adinkra symbols: The effect of testing. *Journal of Cognitive Psychology*, 23(3), 351–357.
- DeWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & cognition*, 32(6), 945–955.
- Dobson, J. L., & Linderholm, T. (2015). Self-testing promotes superior retention of anatomy and physiology information. *Advances in Health Sciences Education*, 20(1), 149–161.
- Donker, A. S., De Boer, H., Kostons, D., Van Ewijk, C. D., & van der Werf, M. P. (2014). Effectiveness of learning strategy instruction on academic performance: A meta-analysis. *Educational Research Review*, 11, 1–26.
- Dunlosky, J., & Rawson, K. A. (2015). Do students use testing and feedback while learning? A focus on key concept definitions and learning to criterion. *Learning and Instruction*, 39, 32–44.
- Einstein, G. O., Mullet, H. G., & Harrison, T. L. (2012). The testing effect: Illustrating a fundamental concept and changing study strategies. *Teaching of Psychology*, 39(3), 190–193.
- Endres, T., & Renkl, A. (2015). Mechanisms behind the testing effect: an empirical investigation of retrieval practice in meaningful learning. *Frontiers in psychology*, 6, 1054.
- Escribe, C., & Huet, N. (2005). Knowledge accessibility, achievement goals, and memory strategy maintenance. *British Journal of Educational Psychology*, 75(1), 87–104.
- Finn, B., & Tauber, S. K. (2015). When confidence is not a signal of knowing: How students' experiences and beliefs about processing fluency can lead to miscalibrated confidence. *Educational Psychology Review*, 27(4), 567–586.
- Fishbein, M., & Ajzen, I. (1975). *Intention and behavior: An introduction to theory and research*. Addison-Wesley, Reading, MA.
- Geller, J., Toftness, A. R., Armstrong, P. I., Carpenter, S. K., Manz, C. L., Coffman, C. R., & Lamm, M. H. (2018). Study strategies and beliefs about learning as a function of academic achievement and achievement goals. *Memory*, 26(5), 683–690.
- Grimaldi, P. J., & Karpicke, J. D. (2014). Guided retrieval practice of educational materials using automated scoring. *Journal of Educational Psychology*, 106(1), 58–68.
- Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, 19(1), 126–134.
- Hattie, J., Biggs, J., & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. *Review of educational research*, 66(2), 99–136.
- Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(6), 1441.
- Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, 19(4), 509–539.
- Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of experimental psychology: General*, 138(4), 469–486.
- Karpicke, J. D., & Roediger III, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(4), 704.
- Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive psychology*, 115, 101237.

- 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.
- Koriat, A., Nussinson, R., & Ackerman, R. (2014). Judgments of learning depend on how learners interpret study effort. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(6), 1624.
- Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, 14(2), 219–224.
- Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, 17(5), 493–501.
- Mayer, R. E., & Gallini, J. K. (1990). When is an illustration worth ten thousand words? *Journal of Educational Psychology*, 82(4), 715–726.
- McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, 19(4-5), 494–513.
- Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, 84(4), 429–434.
- Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, 330(6002), 335–335.
- Rawson, K. A., & Dunlosky, J. (2012). When is practice testing most effective for improving the durability and efficiency of student learning? *Educational Psychology Review*, 24(3), 419–435.
- Roediger III, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on psychological science*, 1(3), 181–210.
- Roediger III, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, 17(3), 249–255.
- Rokeach, M. (1973). *The nature of human values*. Free press.
- 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.
- Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory & cognition*, 11(6), 641–650.
- Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of educational psychology*, 82(3), 498–504.
- Schwartz, B. L., Benjamin, A. S., & Bjork, R. A. (1997). The inferential and experiential bases of metamemory. *Current Directions in Psychological Science*, 6(5), 132–137.
- Toppino, T. C., LaVan, M. H., & Iaconelli, R. T. (2018). Metacognitive control in self-regulated learning: Conditions affecting the choice of restudying versus retrieval practice. *Memory & Cognition*, 46(7), 1164–1177.
- Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & cognition*, 41(3), 429–442.
- Vaughn, K. E., & Kornell, N. (2019). How to activate students' natural desire to test themselves. *Cognitive research: principles and implications*, 4(1), 35.
- Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of experimental psychology: General*, 145(7), 918–933.