Journal of Memory and Language 73 (2014) 99–115

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Testing facilitates the regulation of subsequent study time

Nicholas C. Soderstrom *, Robert A. Bjork

Department of Psychology, University of California, Los Angeles, USA

ARTICLE INFO

Article history:
Received 26 November 2013
revision received 24 February 2014
Available online 27 March 2014

Keywords:
Testing
Metacognition
Test-potentiated learning
Study-time allocation
Self-regulated learning

ABSTRACT

We examined how testing potentiates self-regulated learning and alleviates the foresight bias—an illusion of competence that arises from information being present during study but absent at test—and whether such benefits can transfer to non-tested material. After studying paired associates that varied in difficulty, participants either restudied or were tested on all the pairs (Experiment 1); were tested on only half of the pairs (Experiment 2); or were tested on half of the pairs and restudied the remaining pairs (Experiments 3 and 4). All items were then restudied at participants' own pace before a final cued-recall test. In Experiment 1, interim tests enhanced the effectiveness of subsequent study time and alleviated the foresight bias, whereas interim restudying had no such benefits. Experiments 2, 3, and 4 demonstrated that such test-potentiated self-regulated learning can transfer to non-tested items if restudied intermixed with items that were tested. The results demonstrate yet another practical benefit of testing and suggest that retrieval practice can foster metacognitive sophistication among learners, serving as an experience-based debiasing procedure.

Introduction

As the popularity of online courses grows and technological advances usher in new means of instruction, learning is increasingly occurring in unsupervised settings outside of the classroom. Consequently, learners are becoming more autonomous agents in their own education, responsible for initiating and managing their own learning. Such self-regulated learning has been the focus of much empirical work (for a comprehensive review, see Bjork, Dunlosky, & Kornell, 2013), and the current study contributes to this important line of research. In particular, we examined how testing informs learners' subsequent self-regulated study behavior.

Benefits of testing

Decades of research overwhelmingly supports the notion that tests, far from acting merely as passive assessments of what has been learned, confer myriad direct and indirect benefits to learning (see, e.g., Carpenter, 2012; Roediger, Putnam, & Smith, 2011). To date, most of the empirical work on test-enhanced learning has emphasized the direct benefits of testing—that is, the learning effects driven by the test itself (i.e., *the testing effect*; for a comprehensive review, see Roediger & Karpicke, 2006). Here, the retrieval practice that is promoted by testing acts as a "memory modifier" (Bjork, 1975) in the sense that it renders the successfully retrieved information more recallable in the future than if that same information received no retrieval practice or was permitted additional study time. The testing effect has emerged as one of the most robust and reliable effects in all of memory research.

More relevant to the current study is the growing literature suggesting that tests can also have indirect, or

* Corresponding author. Address: Department of Psychology, 1285 Franz Hall, UCLA, Los Angeles, CA 90095-1563, USA.
E-mail address: nsoderstrom@psych.ucla.edu (N.C. Soderstrom).

http://dx.doi.org/10.1016/j.jml.2014.03.003
0749-596X/

mediated, effects on learning. The learning benefits associated with testing can, for example, transfer to non-tested material, a phenomenon termed *retrieval-induced facilitation*. Chan, McDermott, and Roediger (2006; see also Chan, 2009; Chan, 2010) showed that testing participants after reading prose passages facilitated later memory for tested information, but also enhanced memory, albeit to a lesser extent, for related content that was presented in the passage but not tested.[1] Likewise, Little, Bjork, Bjork, and Angello (2012) demonstrated that multiple-choice tests can foster learning of previously tested information and, to a lesser extent, information related to plausible (i.e., competitive) incorrect alternatives. Competitive alternatives, Little et al. argue, induce students to recall why those alternatives are incorrect, thus leading to retrieval-induced facilitation of those items.

Tests can also enhance the effectiveness of subsequent encoding, a phenomenon termed *test-potentiated learning.* This was first demonstrated by Izawa (1966; see also Izawa, 1968; Izawa, 1970; Izawa, 1971) who showed that increasing the number of tests facilitates subsequent encoding when the tested material is restudied, a finding that has been recently replicated and extended (Arnold & McDermott, 2013; Grimaldi & Karpicke, 2012; Hays, Kornell, & Bjork, 2013; Karpicke, 2009; Karpicke & Roediger, 2007; Kornell, Hays, & Bjork, 2009). Arnold and McDermott (2013), for example, varied the number of interim tests between the initial study of Russian–English word pairs (e.g., *medved—bear*) and their subsequent restudy. A variety of analyses were then conducted to isolate the indirect effects of the tests, the most crucial of which examined the number of previously unrecalled items that were acquired during restudy trials—that is, items that were not successfully retrieved immediately *before* restudy but were successfully retrieved immediately *after* restudy. Consistent with the notion that tests potentiate subsequent learning, more previously unrecalled items were acquired during restudy trials when more interim tests had been taken.

Test-potentiated learning also appears to transfer to new and more complex materials. Wissman, Rawson, and Pyc (2011) had participants study expository texts on, for example, the US labor market or greenhouse gases, which were divided into three sections. Participants in the interim-test condition attempted to recall each section's material before studying the next section, whereas participants in the no-interim-test condition were only prompted to recall the third (final) section. Across five experiments, recall of the final section was greater when interim tests had been taken for the previous two sections compared to when no such tests had been taken. That is, interim tests enhanced the learning of new material. Speculating on the possible mechanisms, Wissman et al. suggested that tests may engender more effective subsequent encoding strategies (see also Pyc & Rawson, 2010; Pyc & Rawson, 2012). Alternatively—or, additionally—tests may protect against proactive interference (Szpunar, McDermott, & Roediger, 2008) and/or reduce mind-wandering (Szpunar, Khan, & Schacter, 2013). Regardless of the responsible mechanism(s), however, it is clear that taking tests can enhance subsequent learning of tested and non-tested material.

Finally—and undoubtedly a contributing factor to the potentiating effects of testing—tests also bestow a metacognitive benefit to the learner. Generally speaking, tests are useful internal assessment tools in the sense that they allow one to become aware of gaps in one's knowledge, which, in turn, can then aid in the assessment of whether information is likely to be remembered in the future (e.g., Nelson & Dunlosky, 1991; see Rhodes & Tauber, 2011). If permitted practice with multiple study-test phases, for example, the accuracy of participants' memory predictions often improve markedly (e.g., Benjamin, 2003; Castel, 2008; Finn & Metcalfe, 2007; King, Zechmeister, & Shaughnessy, 1980; Koriat, 1997; Koriat, Sheffer, & Ma'ayan, 2002), presumably because prior testing equips the learner with information that is diagnostic of future recall—namely, whether specific items were previously recalled or not. Indeed, recent survey studies have shown that learners engage in self-testing primarily to assess their own learning (e.g., Hartwig & Dunlosky, 2012; Kornell & Son, 2009). Without a testing experience, learners are prone to metacognitive illusions that are marked by dissociations between predicted and actual memory performance (e.g., Benjamin, Bjork, & Schwartz, 1998; Castel, McCabe, & Roediger, 2007; Hertzog, Dunlosky, Robinson, & Kidder, 2003; Koriat, Bjork, Sheffer, & Bar, 2004; Soderstrom & McCabe, 2011; Zechmeister & Shaughnessy, 1980).

The foresight bias

For current purposes, we focus on one metacognitive illusion that was initially demonstrated in a study by Koriat and Bjork (2005), who showed that learners, when not permitted a testing experience, are susceptible to a *foresight bias*, which refers to an overestimation of one's future memory performance brought about by the inherent discrepancy between study and test situations. More specifically, prospective judgments that are solicited during study are typically made in conjunction with information that is absent, but needs to be recalled, during testing. Koriat and Bjork had participants study and make memory predictions for paired associates that differed in their associative direction. For *forward* pairs there existed, according to word association norms, a stronger association from the cue word to the target word than from the target word to the cue word, whereas for *backward* pairs, the opposite was true. To illustrate, the pair *umbrella–rain* is considered a forward pair because the likelihood of *umbrella* eliciting *rain* is very high, whereas the likelihood of *rain* eliciting *umbrella* is very low. Conversely, the pair *rain–umbrella* is considered a backward pair because the stronger association is from the target word (*umbrella*) to the cue word (*rain*). Participants, presumably basing their predictions on the overall fluency (i.e., ease of processing) that derived from studying each pair as a whole, were insensitive to

[1] There is also a large literature on *retrieval-induced forgetting* showing that retrieval of tested material can impair memory for non-tested material (see Anderson, 2003; Anderson, Bjork, & Bjork, 1994). Whether retrieval induces memory facilitation or forgetting may depend on how well the to-be-remembered materials are integrated and the length of delay between retrieval practice and the final test (see Chan, 2009).

associative directionality, giving equivalent memory predictions for forward and backward pairs when asked to assess the likelihood that they would be able to recall the target word when later presented with the cue word. As predicted by the word association norms, however, final recall favored the forward pairs, and thus the foresight bias was confirmed.

Subsequent research by Koriat and Bjork (2006a; Koriat and Bjork, 2006b) examined debiasing procedures—one theory-based and one experience-based—intended to alleviate the foresight bias. As the name implies, the theory-based debiasing procedure involved helping participants formulate a general theory pertaining to the conditions that lead to the foresight bias, which was explicitly explained to participants after an initial study-test phase. The experience-based debiasing procedure, on the other hand, involved simply permitting the learner study-test practice, which, as previously discussed, can by itself substantially improve metacognitive accuracy. Koriat and Bjork (2006b) found that both the theory- and experience-based debiasing procedures sensitized participants to the distinction between forward and backward associates when the same pairs were restudied. Specifically, although forward and backward pairs received equivalent predictions and study-time allocation during the first study phase, backward pairs were given lower predictions and allocated more study time than forward pairs during the subsequent, post-test restudy phase.

Importantly, Koriat and Bjork (2006b) found that only theory-based debiasing transferred to new items, suggesting that the success of testing as an experience-based debiasing procedure is item-specific, and thus does not help learners formulate a rule that can be applied beyond the original learning context. As Koriat and Bjork (2006b) put it: "...participants' study-test experience equips them with useful mnemonic cues about the recallability of different items to the extent of improving their monitoring on a repeated study of these items but provides them with little insight into what they have learned from study-test practice" (p. 1139). Although such a conclusion is reasonable given Koriat and Bjork's (2006b) results, it remains an open question whether experience-based debiasing of the foresight bias is indeed restricted to tested information because no research, to our knowledge, has investigated the potential of such debiasing on non-tested material that is restudied amongst material that was tested. If successful experience-based debiasing is limited to tested information, then we would not expect learners to sensitize to associative directionality for any type of non-tested material. However, if the crucial determinant is reexposure to material after a test, then it is possible that learners can gain an appreciation for the forward-backward distinction, even for items that were not tested.

Theories of study-time allocation

A wealth of research on self-regulated learning—including Koriat and Bjork's (2006b) previously described study investigating the foresight bias—has focused on how learners choose to allocate their study time (for reviews, see Dunlosky & Ariel, 2011; Son & Kornell, 2008). Given that present study is also concerned with this issue, we thought it useful to briefly review the two dominant theories of study-time allocation, both of which drew heavily upon Nelson and Leonesio's (1988) *monitoring-affects-control* hypothesis (see also, Nelson & Narens, 1990). Briefly stated, this hypothesis posits that learners monitor, or reflect upon, their own learning and then use this monitoring to control, or modify, their subsequent behavior.

One model of study-time allocation that directly captures the monitoring-control relationship is the *discrepancy-reduction model* (Dunlosky & Hertzog, 1998; Dunlosky & Thiede, 1998), which asserts that learners will study difficult items (or at least those items perceived as more difficult) longer than easy items in an attempt to reduce the discrepancy between what has been learned and what is sought to be learned. Stated in their own words, "An item will be continued to be studied...until the person's perceived degree of learning meets or exceeds the norm of study" (Dunlosky & Thiede, 1998, p. 1024). Data supporting this model include negative correlations between study time and judged item difficulty, a relationship that was found in most of the early studies investigating study-time allocation (e.g., Nelson, Dunlosky, Graf, & Narens, 1994; for a review, see Son & Metcalfe, 2000).

To accommodate later research showing that learners' study decisions can hinge on whether time constraints are imposed (e.g., Son & Metcalfe, 2000), the *region of proximal learning model* (Metcalfe, 2002; Metcalfe & Kornell, 2005) was developed, which, like the discrepancy-reduction model, generally predicts that people will selectively allocate more time to items that they believe they have not learned. However, the region of proximal learning model departs from the discrepancy-reduction model by emphasizing that the perceived rate of learning is the crucial determinant of how long one will persist in studying to-be-learned information: namely, learners will study material until the benefits of studying are perceived no more, or at least when the rate of learning is substantially reduced. Indeed, Metcalfe and Kornell (2005) showed that, under some circumstances, it is the moderately difficult items, rather than the most difficult items, that are studied the longest. Items of moderate difficulty, it was argued, are associated with longer durations of perceived learning compared to both easy items, which are learned quickly, and extremely difficult items, which are too difficult to learn.

For current purposes, the important point is that the two prevailing theories of study-time allocation, while certainly differing in some aspects, both assume that people will spend a disproportionate amount of time studying information that is perceived to be unlearned.[2] Without an objective measure of what has been learned, however, learners are limited to basing their study-time decisions on what they *think* they know, rather than what they *actually* know. An explicit memory test may be particularly useful in providing the learner with more objective information about what has been learned, which would

[2] Another, more recently formulated model of study-time allocation—the *agenda-based regulation model* (Ariel, Dunlosky, & Bailey, 2009; see also Dunlosky & Ariel, 2011)—argues a major role for one's agendas, or goals, when allocating study time.

better advise learners' subsequent study decisions as compared to if these decisions were exclusively made on subjective grounds.

The present study

The present study was conducted to examine how retrieval practice influences learners' subsequent self-allocated study time and to determine whether experience-based debiasing of the foresight bias can, under some circumstances, generalize to non-tested items. We note that the foresight bias typically refers to the accuracy of predictive judgments, which we did not solicit in the current study; however, as Koriat and Bjork (2006b) demonstrated, study time can be used as a proxy for predictive judgments and thus we felt justified investigating the foresight bias as indexed by study time. Given the previously discussed benefits of testing—particularly, that tests potentiate learning and that retrieval practice improves metacognitive awareness—we predicted that interim tests would enhance the efficiency and effectiveness of learners' subsequent self-paced study and could alleviate the foresight bias. Tests reveal to the learner what has and has not been learned, and thus they should selectively direct subsequent restudy efforts toward items that were not successfully recalled during the interim test. Furthermore, given the ever-growing literature showing that the benefits of testing transfer beyond tested information, we anticipated that non-tested material, if restudied amongst material that had been tested, would also reap the benefits of the testing experience.

Our general methodology is illustrated in Fig. 1. Participants initially studied a list of paired associates comprised of three levels of difficulty (forward, backward, and unrelated word pairs; see Koriat & Bjork, 2005). Next, participants were either tested on all of the items or restudied all of them (Experiment 1); were tested on only half of the pairs (Experiment 2); or were tested on half of the pairs and restudied the remaining pairs (Experiment 3 and 4). In all experiments, all 36 pairs were then restudied at a pace controlled by the participants, who were instructed to study the pairs long enough such that the second word could be recalled if given the first word. Lastly, a final test of all of the items was administered. Such a procedure permitted an examination of how an interim test influences subsequent study behavior of both tested and non-tested items, and whether such influences are beneficial to the learner.

Experiment 1

Experiment 1 was conducted to determine how interim testing and interim restudying influence subsequent study-time allocation. After studying a mixture of forward, backward, and unrelated paired associates, one group of participants was tested on all of the pairs (interim-test group), whereas the other group restudied the pairs again (interim-restudy group). Both groups then restudied all of

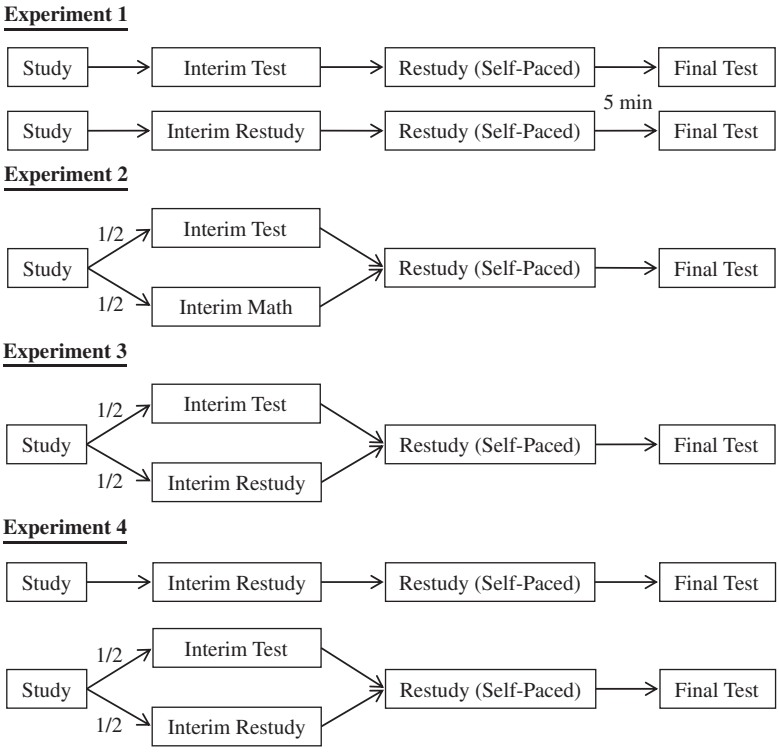

Fig. 1. Schematic of basic experimental procedures for Experiments 1–4.

the pairs one last time at their own pace before a final test. Immediately after the final test, a follow-up question was asked regarding the study strategies participants' used during the self-paced study phase. We expected that the interim-test group would show relatively more effective subsequent study-time allocation and, unlike the interim-restudy group, would become sensitive to the subtle, yet important, distinction between forward and backward associates.

Method

Participants, design, and materials

Thirty-six undergraduates at the University of California, Los Angeles (UCLA) participated for partial course credit. We used a between-subjects design such that participants were assigned to either the interim-restudy or interim-test condition (18 participants in each). The materials were 36 word pairs composed of 12 forward associates, 12 backward associates, and 12 unrelated items (according to Nelson, McEvoy, & Schreiber, 1998). To determine which items would be designated forward and backward associates, we first compiled a list of 24 pairs with asymmetric associations (i.e., pairs with strong forward associations but weak backward associations). We then divided these pairs into two sets of 12 that were equated in terms of their average forward and backward associative strength. For one set, the average forward and backward associative strength was .544 and .028, respectively; for the other set, the averages were .555 and .020, respectively. We then assigned one set to be forward associates (with the strongest association from the cue word to the target word) and one set to be backward associates (with the strongest association from the target word to the cue word), which was counterbalanced across participants. Twelve unrelated pairs (with zero associative strength) were also included.

Procedure

All participants first studied the 36 paired associates (12 forward, 12 backward, 12 unrelated) one at a time for 5 s each in a randomized order. The interim-restudy group then restudied all of the pairs again for 5 s each in a new randomized order, whereas the interim-test group took a cued-recall test. For the interim test, all 36 cue words were presented one at a time for 5 s each in a randomized order, with participants instructed to recall their corresponding target word in that time. Responses were typed and no corrective feedback was provided. Both groups then restudied all items again at their own pace. The instructions for this phase were as follows: "You will now study the word pairs one last time AT YOUR OWN PACE. Once you think that you have studied the pair long enough to be able to recall the second word if presented with the first word, press 'done' underneath the pair." The self-paced study phase was followed by a brief distractor task (5 min of Tetris) before a final cued-recall test in which all 36 cue words were presented one at a time for 5 s each in a randomized order, and participants were instructed to type in each target word within that time.

After the final test was completed, participants were asked a follow-up question regarding their study strategies during the self-paced study phase. The questionnaire was a modified version of the Personal Encoding Preferences Questionnaire (see Hertzog & Dunlosky, 2004) and was worded as follows:

> During the last study phase of the experiment when you studied each pair at your own pace, what study strategies did you use (mark all that apply)?

1. *Rote repetition* (saying the word pair over and over)
2. *Attentive reading* (reading over or saying the word pair once in your mind)
3. *Semantic reference* (relating the word pair to something of meaning in your life)
4. *Focal attention* (focusing on the word pair by looking or staring at it until you can see the pair clearly in your mind)
5. *Imagery* (imagining a scene using the two words as images in it)
6. *Sentence generation* (constructing a sentence using both of the words)
7. *Other strategy* (please explain).

Results and discussion

Interim-recall performance

As intended, there was an effect of item type (forward, backward, unrelated) on interim-recall performance as evidenced by a one-way repeated-measures analysis of variance (ANOVA), $F(2,34) = 30.90$, $p < .05$, $\eta_p^2 = .65$. Recall of forward pairs was higher than backward pairs (.73 vs. .45, respectively), $t(17) = 5.91$, $p < .05$, $d = 1.30$, and recall of backward pairs was marginally higher than unrelated pairs (.45 vs. 36, respectively), $t(17) = 1.62$, $p = .12$, $d = .32$.

Study-time allocation

Fig. 2 presents participants' mean subsequent study-time allocation for each item type after an interim test or interim restudy. Focusing first on each condition's overall mean study time, comparing interim restudy to interim test (overall), a 2 (condition: interim restudy vs. interim test) $\times$ 3 (item type: forward, backward, unrelated) mixed-model ANOVA revealed that the interim-test group spent more time studying items than the interim-restudy group, $F(1,34) = 7.73$, $p < .05$, $\eta_p^2 = .19$, and that study-time allocation was influenced by the type of item being studied, $F(2,68) = 25.46$, $p < .05$, $\eta_p^2 = .43$. However, these main effects were qualified by a reliable interaction, $F(2,68) = 8.39$, $p < .05$, $\eta_p^2 = .20$. Whereas no study time difference was found between conditions for forward pairs ($p > .05$), the interim-test group allocated more time than the interim-restudy group to backward pairs, $t(34) = 2.80$, $p < .05$, $d = .93$, and unrelated pairs, $t(34) = 3.05$, $p < .05$, $d = 1.02$.

Still comparing interim restudy to interim test (overall), we specifically examined whether an interim test sensitized learners to the distinction between forward and backward associative strength by conducting a 2 (condition: interim restudy vs. interim test) $\times$ 2 (item type: forward vs. backward) mixed-model ANOVA. This analysis

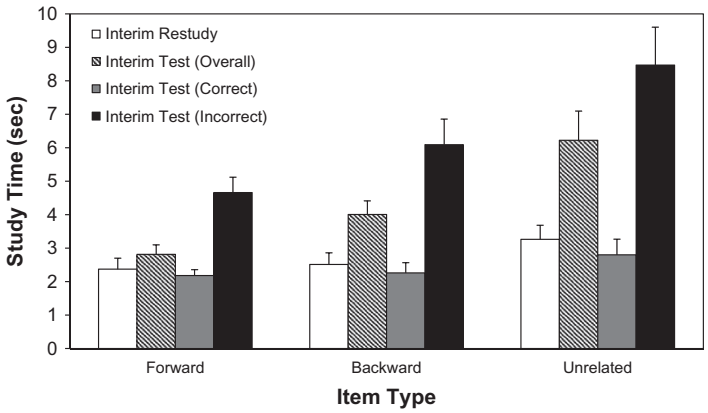

Fig. 2. Mean study-time allocation as a function of item type (forward, backward, and unrelated pairs) following interim restudy or an interim test in Experiment 1. For the interim-test condition, overall study time is reported as well as mean study times conditionalized on whether items were correctly or incorrectly recalled during the interim test. Error bars represent standard error of the means.

revealed a reliable interaction, $F(1,34) = 10.25$, $p < .05$, $\eta_p^2 = .23$. As evident in Fig. 2, the interim-test group devoted more time overall to backward pairs than forward pairs, $t(17) = 4.15$, $p < .05$, $d = .82$, whereas the interim-restudy group studied forward and backward pairs for the same duration ($p > .05$). Thus, taking an interim test alleviated the foresight bias as indexed by study time allocation. It should be noted, however, that the interim-restudy group was not completely insensitive to item difficultly as unrelated pairs were allocated more time than both forward pairs, $t(17) = 4.58$, $p < .05$, $d = .56$, and backward pairs, $t(17) = 4.07$, $p < .05$, $d = .46$, in this condition. Participants in the interim-test group also allocated the most time overall to unrelated pairs, studying them longer than backward pairs, $t(17) = 3.50$, $p < .05$, $d = .82$.

We next examined how performance on the interim test influenced participants' subsequent self-paced study time. Fig. 2 presents the interim-test group's mean subsequent study times conditionalized on whether items were correctly or incorrectly recalled during the interim test. (We note that incorrect responses for all experiments included errors of omission and commission, although the former was far more common). A 2 (interim-recall accuracy: correct vs. incorrect) × 3 (item type: forward, backward, unrelated) repeated-measures ANOVA revealed that, unsurprisingly, items recalled incorrectly on the interim test were allocated more subsequent study time than items recalled correctly, $F(1,17) = 48.90$, $p < .05$, $\eta_p^2 = .74$, and that there was an effect of item type, $F(2,34) = 9.62$, $p < .05$, $\eta_p^2 = .35$. However, these effects were qualified by a reliable interaction, $F(2,34) = 5.21$, $p = .05$, $\eta_p^2 = .24$. For correct items, no differences in study time were found across item type, $F(2,34) = 1.45$, $p > .05$, $\eta_p^2 = .08$, perhaps suggesting that participants quickly terminated study after realizing that these items had been successfully recalled during the interim test. For incorrect items, however, study time increased monotonically with item difficulty, $F(2,34) = 8.39$, $p < .05$, $\eta_p^2 = .33$. Backward pairs were studied marginally longer than forward pairs, $t(17) = 1.69$, $p = .10$, $d = .55$, and unrelated pairs were studied longer than backward pairs, $t(17) = 2.93$, $p < .05$, $d = .59$. Finally, it is clear from Fig. 2 that participants in the interim-restudy condition did not restudy items any longer than interim-test participants restudied items that were correctly recalled during the interim test ($F < 1$).

Final-recall performance

The proportions of items recalled on the final cued-recall test by the interim-restudy and interim-test groups are shown in Fig. 3 as a function of item type (the dashed lines embedded in the interim-test group's data bars represent performance on the interim-recall test). A 2 (condition: interim restudy vs. interim test) × 3 (item type: forward, backward, unrelated) repeated-measures ANOVA revealed two significant effects: The interim-test group recalled more items overall than the interim-restudy group, $F(1,34) = 6.37$, $p < .05$, $\eta_p^2 = .16$, and recall was affected by item type, $F(2,68) = 21.64$, $p < .05$, $\eta_p^2 = .39$.

It is also clear from Fig. 3 that the interim-test group greatly profited from the self-paced study phase. A 2 (recall: interim test vs. final test) × 3 (item type: forward, backward, unrelated) repeated-measures ANOVA revealed that the interim-test group recalled more items during the final-recall test than the interim-recall test, $F(1,34) = 76.92$, $p < .05$, $\eta_p^2 = .82$. This was true for forward pairs, $t(17) = 5.80$, $p < .05$, $d = 1.26$, backward pairs, $t(17) = 7.19$, $p < .05$, $d = 1.51$, and unrelated pairs, $t(17) = 7.71$, $p < .05$, $d = 1.13$. Furthermore, the interim-test group's substantial gains in acquiring new items during the self-paced study phase did not come at the expense of previously recalled items: 98% of the forward pairs, 96% of the backward pairs, and 97% of the unrelated pairs recalled during the interim test were also recalled during the final recall test.

Study strategies

Fig. 4 presents participants' self-reported study strategies used during the self-paced study phase (reported after the final test). The vertical dashed line equally divides the strategies into what previous experimental research has established as relatively ineffective (rote repetition, attentive reading, focal attention) and effective (semantic reference, imagery, sentence generation) study strategies

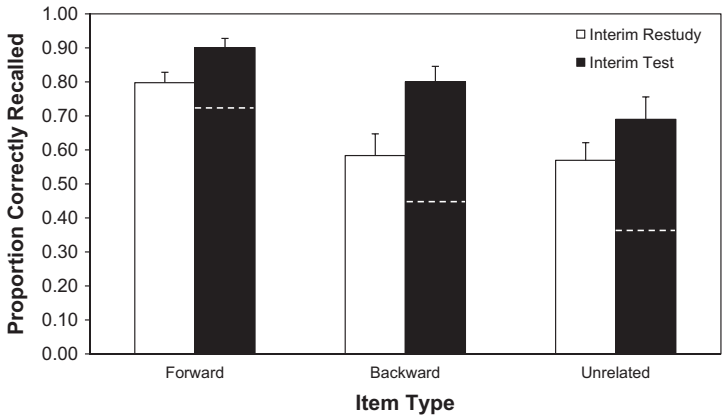

Fig. 3. Mean proportion of items correctly recalled on the final cued-recall test as a function of item type (forward, backward, and unrelated pairs) following interim restudy or an interim test in Experiment 1. The dashed lines embedded in the interim-test data bars denote interim-recall performance. Error bars represent standard error of the means.

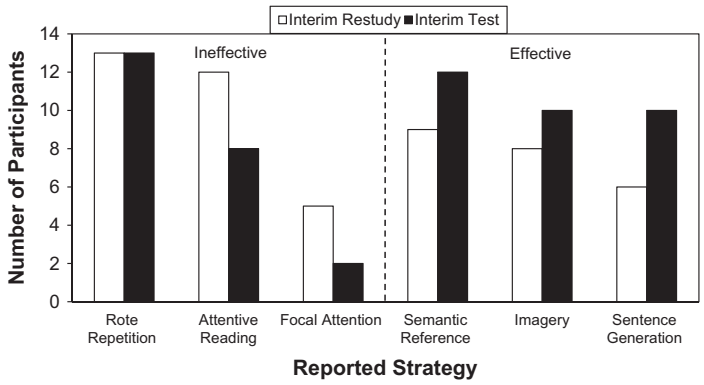

Fig. 4. Number of participants (out of 18 in each condition) who reported using each encoding strategy during the self-paced study phase following interim restudy or an interim test in Experiment 1. Strategies to the left of the vertical dashed have been shown by previous research to be ineffective; strategies to the right of the dashed line have been shown to be effective. Note that participants could indicate using more than one strategy as they were instructed to 'mark all that apply'.

for paired-associate learning. With the exception of rote rehearsal, ineffective strategies appeared to be reported more often by interim-restudy participants than interim-test participants, whereas effective strategies appeared to be reported by more interim-test participants than interim-restudy participants. Furthermore, of all reported strategies (53 in the interim-restudy group; 55 in the interim-test group), the interim-restudy group reported using 57% ineffective and 43% effective strategies, whereas the interim-test group reported 42% ineffective and 58% effective strategies. Thus, not only did taking an interim test lead participants to effectively devote more time to difficult items, it also appeared to foster more effective study strategies.

Experiment 1 revealed several findings of interest. First, the interim-test group spent more time studying items during the self-paced study phase than the interim-restudy group and, as predicted, taking an interim test alleviated the foresight bias. Unsurprisingly, the interim-test group allocated most of their subsequent study time to items that were not successfully recalled on the interim test with study time for these items increasing monotonically with item difficulty. For items recalled correctly on the interim test, however, no study time differences were found as a function of item difficulty, suggesting that participants quickly terminated study after realizing that these items had already been recalled during the interim test. Without a testing experience to draw upon, participants in the interim-restudy condition did not restudy items any longer than interim-test participants restudied previously recalled items, perhaps reflecting an illusion of knowing brought about by the subjective fluency in which items were processed upon the their third presentation. A follow-up questionnaire revealed that the quality of restudying also favored the interim-test group, whose participants reported using relatively more effective encoding strategies during the self-paced study phase. Consequently, final recall in Experiment 1 favored the interim-test condition across all items, demonstrating yet another instantiation of test-potentiated learning.

Experiment 2

Can the test-potentiated self-regulated learning demonstrated in Experiment 1 transfer to previously presented, non-tested material? To illustrate, consider an instructor who gives a quiz at the end of lecture covering half of the lecture's content. Experiment 1 showed that subsequent self-regulated learning of the quizzed material will be enhanced, but might the non-quizzed material also profit from the quiz? Given recent findings that have suggested that tests can enhance the learning of tested and non-tested information (see Carpenter, 2012), we predicted that it would. To explore this possibility, Experiment 2, as shown in Fig. 1, employed a methodology similar to Experiment 1, except that all participants took an interim test that included only half of the initially studied items before restudying all of the items—tested and non-tested—at their own pace.

Method

Participants, design, materials, and procedure

Twenty-eight undergraduates at the University of California, Los Angeles (UCLA) participated for partial course credit. The design, materials, and procedure were similar to Experiment 1 with the exception that all participants took an interim-test covering half of the initially studied items. For those items that were not tested during the interim, simple math problems (e.g., $8 \times 3 = ?$) were performed in their place in order to equate procedure time with Experiment 1. Like the interim-test trials, the math problems were fixed-paced at 5 s each, during which time participants typed in their responses. To determine which items would be tested during the interim and which would not, we first randomly divided the items from each item type (forward, backward, and unrelated pairs) into two sets of six. Counterbalanced across participants, one set of each item type was designated interim-test items, whereas the other set was designated non-interim-test items, in which math problems were performed in their place. The interim-test trials and math problems were randomly intermixed during the interim phase.

Results and discussion

Interim-recall performance

As intended, interim recall was affected by item type, $F(2,54) = 18.34$, $p < .05$, $\eta_p^2 = .40$. Forward pairs were better recalled than backward pairs (.66 vs. .38, respectively), $t(27) = 6.20$, $p < .05$, $d = 1.21$, and backward pairs were better recalled—numerically, but not statistically—than unrelated pairs (.38 vs. .30, respectively) ($p > .05$).

Study-time allocation

Fig. 5 presents participants' mean subsequent study-time allocation for each item type after interim math or an interim test. Focusing first on each condition's overall study time, comparing interim math vs. interim test (overall), a 2 (condition: interim math vs. interim test) $\times$ 3 (item type: forward, backward, unrelated) repeated-measures ANOVA showed that interim-math items (i.e., non-tested items that were replaced by math problems during the interim phase) were devoted marginally more subsequent restudy time than interim-test items, $F(1,27) = 4.14$, $p = .06$, $\eta_p^2 = .13$, and that study time for both interim-math and interim-test items was influenced by item type, $F(2,54) = 19.37$, $p < .05$, $\eta_p^2 = .42$. Unlike Experiment 1, the interaction was not reliable ($F < 1$).

A more focused 2 (condition: interim math vs. interim test) $\times$ 2 (item type: forward vs. backward) ANOVA on each condition's overall study time was conducted to examine whether the foresight bias was alleviated for non-tested (interim math) items when restudied amongst items that had been tested. Consistent with this notion, a main effect of item type, $F(1,27) = 8.07$, $p < .05$, $\eta_p^2 = .23$, but not a reliable interaction ($F < 1$), was found. Indeed, as evident in Fig. 5, for interim-test items more subsequent study time was allocated to backward pairs than forward pairs, $t(27) = 1.75$, $p = .09$, $d = .29$, which was also true for interim-math items, $t(27) = 2.49$, $p < .05$, $d = .43$. For completeness, we note that unrelated pairs were allocated more time than backward pairs for both interim-test items, $t(27) = 3.29$, $p < .05$, $d = 64$, and interim-math items, $t(27) = 3.88$, $p < .05$, $d = .41$.

Fig. 5 also presents interim-test items' subsequent study times conditionalized on whether items were correctly or incorrectly recalled during the interim test. A 2 (interim-recall accuracy: correct vs. incorrect) $\times$ 3 (item type: forward, backward, unrelated) repeated-measures ANOVA revealed that more time was allocated to previously unrecalled items than previously recalled items, $F(1,27) = 26.07$, $p < .05$, $\eta_p^2 = .49$, and that there was an effect of item type, $F(2,54) = 4.99$, $p < .05$, $\eta_p^2 = .16$. These effects, however, were qualified by a reliable interaction, $F(2,54) = 5.44$, $p < .05$, $\eta_p^2 = .17$. For correct items, no differences in study time were found across item type, $F(2,54) = 1.01$, $p > .05$, $\eta_p^2 = .04$. For incorrect items, however, there was an effect of item type, $F(2,54) = 5.55$, $p < .05$, $\eta_p^2 = .17$, such that no difference was found comparing forward and backward pairs ($p > .05$), but unrelated pairs were studied longer than both forward pairs, $t(27) = 2.69$, $p < .05$, $d = .43$, and backward pairs, $t(27) = 2.91$, $p < .05$, $d = .45$.

It is also clear from Fig. 5 that interim-math items were subsequently studied longer than correctly recalled interim-test items, $F(1,27) = 37.54$, $p < .05$, $\eta_p^2 = .58$. Follow-up t-tests showed this to be the case for forward pairs, $t(27) = 3.29$, $p < .05$, $d = .65$, backward pairs, $t(27) = 4.76$, $p < .05$, $d = 1.16$, and unrelated pairs, $t(27) = 6.50$, $p < .05$, $d = 1.63$. Furthermore, the study time differences between interim-math items and correctly recalled interim-test items expanded as a function of item type, $F(2,54) = 16.13$, $p < .05$, $\eta_p^2 = .37$. Compared to incorrectly recalled interim-test items, however, interim-math items were allocated significantly less time, $F(1,27) = 9.23$, $p < .05$, $\eta_p^2 = .26$. Follow-up tests showed this to be true for forward pairs, $t(27) = 3.21$, $p < .05$, $d = .60$, and unrelated pairs, $t(27) = 2.71$, $p < .05$, $d = .34$, but not for backward pairs ($p > .05$). Thus, generally speaking, subsequent study times for interim-math items fell in between study times for items recalled correctly and incorrectly on the interim test.

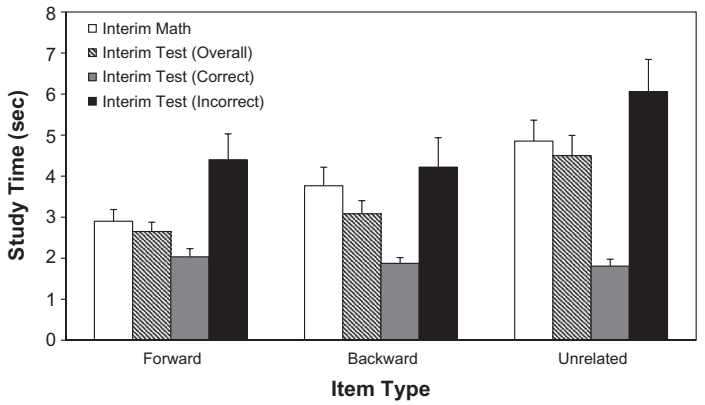

Fig. 5. Mean study-time allocation as a function of item type (forward, backward, and unrelated pairs) following interim math or an interim test in Experiment 2. For the interim-test items, overall study time is reported as well as mean study times conditionalized on whether items were correctly or incorrectly recalled during the interim test. Error bars represent standard error of the means.

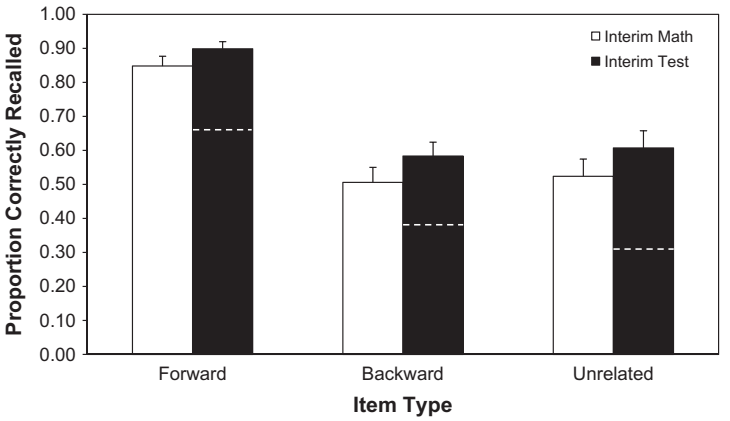

Fig. 6. Mean proportion of items correctly recalled on the final cued-recall test as a function of item type (forward, backward, and unrelated pairs) following interim math or an interim test in Experiment 2. The dashed lines embedded in the interim-test data bars denote interim-recall performance. Error bars represent standard error of the means.

Final-recall performance

Fig. 6 shows the proportion of items correctly recalled on the final cued-recall test across item type as a function of whether the items were tested during the interim or instead had math problems performed in their place (the dashed lines embedded in the interim-test items' data bars represent their performance on the interim-recall test). A 2 (condition: interim math vs. interim test) × 3 (item type: forward, backward, unrelated) repeated-measures ANOVA revealed two significant effects: Interim-test items were better recalled, in general, than interim-math items, $F(1,27) = 7.02$, $p < .05$, $\eta_p^2 = .21$, and recall was influenced by item type, $F(2,54) = 52.73$, $p < .05$, $\eta_p^2 = .66$. For interim-test items, item type influenced final recall, $F(2,54) = 22.97$, $p < .05$, $\eta_p^2 = .46$, such that forward pairs were recalled better than backward pairs, $t(27) = 7.79$, $p < .05$, $d = 1.86$; no difference was found between backward and unrelated pairs ($p > .05$). For interim-math items, final recall was influenced by item type in the same way, $F(2,54) = 35.24$, $p < .05$, $\eta_p^2 = .57$. That is, forward pairs were better recalled than backward pairs, $t(27) = 8.14$, $p < .05$, $d = 1.71$, but recall did not favor backward pairs over unrelated pairs ($p > .05$).

Similar to Experiment 1, Fig. 6 also shows that interim-test items in Experiment 2 profited substantially from the self-paced study phase. A 2 (recall: interim test vs. final test) × 3 (item type: forward, backward, unrelated) repeated-measures ANOVA revealed that more interim-test items were recalled during the final-recall test than the interim-recall test, $F(1,27) = 105.58$, $p < .05$, $\eta_p^2 = .80$. This was true for forward pairs, $t(27) = 5.30$, $p < .05$, $d = 1.35$, backward pairs, $t(27) = 6.84$, $p < .05$, $d = .91$, and unrelated pairs, $t(27) = 8.58$, $p < .05$, $d = 1.11$.

Experiment 2 was conducted to determine whether the test-potentiated self-regulated learning that was demonstrated for tested items in Experiment 1 could transfer to non-tested items. Indeed, testing half of the items during the interim enhanced study-time allocation for the non-tested items. Specifically—and unlike in Experiment 1—non-tested items in Experiment 2 were relieved of the foresight bias and were allocated more time than correctly recalled interim-test items. Consequently, final recall

performance of the non-tested items approached that of the interim-tested items, which is quite impressive given that the non-tested items were only studied twice, whereas interim-tested items were studied twice *and* were afforded the additional re-exposure time that accompanied successful recall on the interim test. This potentially unfair advantage for interim-tested items was addressed in Experiment 3.

Experiment 3

Experiment 2 showed that a testing experience can enhance study-time allocation of tested and non-tested material and provided the first empirical evidence, to our knowledge, that the experience-based debiasing procedure of testing can alleviate the foresight bias, as indexed by study time, for non-tested items. Experiment 3 sought to determine whether these novel findings would replicate after equating total exposure time for tested and non-tested items, which was not controlled for in Experiment 2. To this end, non-tested items in Experiment 3 were restudied amongst tested items during the interim phase (see Fig. 1).

Method

Participants, design, materials, and procedure

Twenty-six undergraduates at the University of California, Los Angeles (UCLA) participated for partial course credit. The design, materials, and procedure were identical to Experiment 2 with one exception: Whereas the non-tested items in Experiment 2 were replaced with math problems during the interim phase, non-tested items in Experiment 3 were restudied. That is, half of the initially studied items were tested and half were restudied prior to all of the items being studied one final time at participants' own pace. Interim-test and interim-restudy trials were randomly intermixed and were counterbalanced across participants.

Results and Discussion

Interim-recall performance

As expected, interim recall was affected by item type, $F(2,50) = 32.99$, $p < .05$, $\eta_p^2 = .57$. Forward pairs were better recalled than backward pairs (.73 vs. .47, respectively), $t(25) = 4.69$, $p < .05$, $d = 1.19$, and backward pairs were better recalled than unrelated pairs (.47 vs. .31, respectively), $t(25) = 2.98$, $p < .05$, $d = .65$.

Study-time allocation

Fig. 7 presents participants' mean subsequent study-time allocation for each item type after interim restudy or an interim test. First focusing on each condition's overall study time, comparing interim restudy vs. interim test (overall), a 2 (condition: interim restudy vs. interim test) × 3 (item type: forward, backward, unrelated) repeated-measures ANOVA revealed that interim-restudy items and interim-test items were subsequently studied for equal durations ($F < 1$), and that study time for both interim-restudy and interim-test items was influenced by item type, $F(2,50) = 16.10$, $p < .05$, $\eta_p^2 = .39$. Like Experiment 2, the interaction was not reliable $F(2,50) = 1.78$, $p > .05$, $\eta_p^2 = .06$.

A more focused 2 (condition: interim restudy vs. interim test) × 2 (item type: forward vs. backward) ANOVA on each condition's overall study time was conducted to examine whether the foresight bias was alleviated for non-tested (interim restudy) items—like it was in Experiment 2—when restudied amongst items that had been tested. Consistent with this notion, a main effect of item type, $F(1,25) = 14.69$, $p < .05$, $\eta_p^2 = .37$, but not a reliable interaction $F(1,25) = 1.29$, $p > .05$, $\eta_p^2 = .05$, was found. Indeed, as evident in Fig. 7, more overall subsequent study time was allocated to backward pairs than forward pairs for interim-test items, $t(25) = 2.64$, $p < .05$, $d = .52$, and interim-restudy items, $t(25) = 2.07$, $p < .05$, $d = .20$.

We next turn to the interim-test items' subsequent study times conditionalized on whether they were correctly or incorrectly recalled during the interim test, which are also presented in Fig. 7. A 2 (interim-recall accuracy: correct vs. incorrect) × 3 (item type: forward, backward, unrelated) repeated-measures ANOVA showed that more time was allocated to items recalled incorrectly than correctly, $F(1,25) = 33.93$, $p < .05$, $\eta_p^2 = .58$, and that there was an effect of item type, $F(2,50) = 4.56$, $p < .05$, $\eta_p^2 = .15$. These effects, however, were qualified by a reliable interaction, $F(2,50) = 8.78$, $p < .05$, $\eta_p^2 = .26$. For correct items, no differences in study time were found across item type ($F < 1$). For incorrect items, however, there was an effect of item type, $F(2,50) = 7.32$, $p < .05$, $\eta_p^2 = .23$, such that backward pairs were studied longer than forward pairs $t(25) = 2.60$, $p < .05$, $d = .46$; no difference was found comparing backward and unrelated pairs ($p > .05$).

It is also clear from Fig. 7 that interim-restudy items were subsequently studied longer than correctly recalled interim-test items, $F(1,25) = 10.89$, $p < .05$, $\eta_p^2 = .30$, and that such differences in study duration expanded as a function of item type, $F(2,50) = 4.29$, $p < .05$, $\eta_p^2 = .15$. Indeed, whereas only a marginal difference was found in favor of interim-restudy for forward pairs, $t(25) = 1.92$, $p = .07$, $d = .32$, clear differences were found for backward pairs, $t(25) = 2.21$, $p < .05$, $d = .60$, and unrelated pairs, $t(25) = 4.54$, $p < .05$, $d = 1.05$. Compared to incorrectly recalled interim-test items, however, interim-restudy items were allocated significantly less time, $F(1,25) = 8.39$, $p < .05$, $\eta_p^2 = .25$. This difference was only marginally significant comparing backward pairs, $t(25) = 1.97$, $p = .06$, $d = .51$, but was reliable for unrelated pairs, $t(25) = 3.13$, $p < .05$, $d = .66$. Thus, as in Experiment 2, subsequent study times for non-tested items in Experiment 3 fell in between study times for items recalled correctly and incorrectly on the interim test.

Final-recall performance

Fig. 8 shows the proportions of items correctly recalled on the final cued-recall test as a function of item type and whether the items were tested or restudied during the interim (the dashed lines embedded in the interim-test items' data bars represent their performance on the interim-recall test). In contrast to Experiments 1 and 2, a 2 (condition: interim restudy vs. interim test) × 3 (item type:

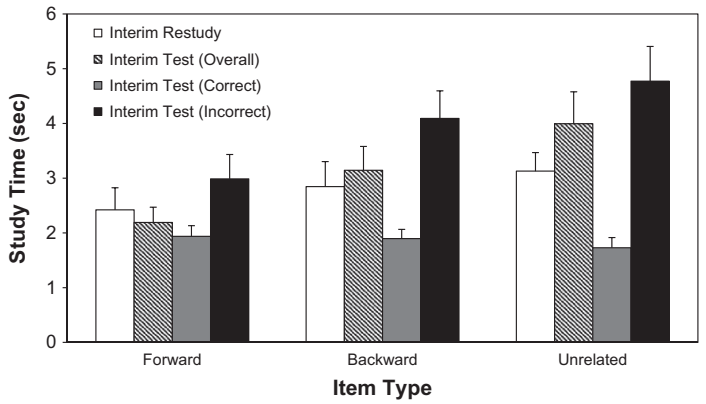

Fig. 7. Mean study-time allocation as a function of item type (forward, backward, and unrelated pairs) following interim restudy or an interim test in Experiment 3. For the interim-test items, overall study time is reported as well as mean study times conditionalized on whether items were correctly or incorrectly recalled during the interim test. Error bars represent standard error of the means.

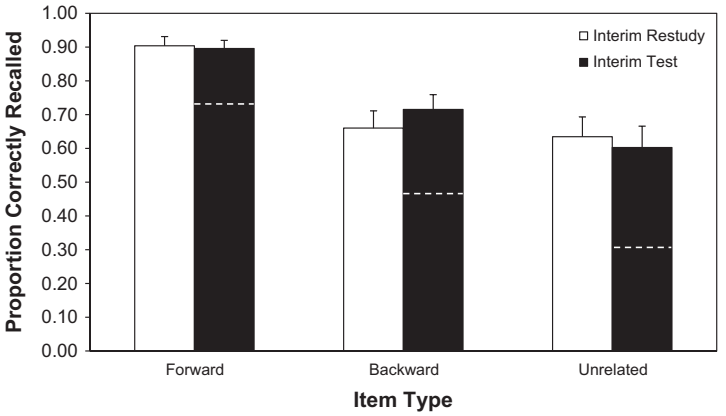

Fig. 8. Mean proportion of items correctly recalled on the final cued-recall test as a function of item type (forward, backward, and unrelated pairs) following interim restudy or an interim test in Experiment 3. The dashed lines embedded in the interim-test data bars denote interim-recall performance. Error bars represent standard error of the means.

forward, backward, unrelated) repeated-measures ANOVA revealed that interim-test items were not better recalled than interim-restudy items ($F < 1$). There was, however, a main effect of item type, $F(2,50) = 22.92$, $p < .05$, $\eta_p^2 = .48$. For interim-restudy items, item type influenced final recall, $F(2,50) = 16.44$, $p < .05$, $\eta_p^2 = .40$, such that forward pairs were recalled better than backward pairs, $t(25) = 5.20$, $p < .05$, $d = 1.22$, but no difference was found between backward and unrelated pairs ($p > .05$). For interim-test items, final recall was also influenced by item type, $F(2,50) = 15.98$, $p < .05$, $\eta_p^2 = 39$, such that forward pairs were recalled better than backward pairs, $t(25) = 4.06$, $p < .05$, $d = 1.05$, and backward pairs were recalled marginally better than unrelated pairs, $t(25) = 2.02$, $p = .05$, $d = .41$.

Similar to Experiments 1 and 2, Fig. 8 also shows that interim-test items in Experiment 3 greatly profited from the self-paced study phase. A 2 (recall: interim test vs. final test) $\times$ 3 (item type: forward, backward, unrelated) repeated-measures ANOVA revealed that more interim-test items were recalled during the final-recall test than the interim-recall test, $F(1,25) = 71.77$, $p < .05$, $\eta_p^2 = .74$. This was true for forward pairs, $t(25) = 4.75$, $p < .05$, $d = 1.05$, backward pairs, $t(25) = 5.94$, $p < .05$, $d = 1.03$, and unrelated pairs, $t(25) = 6.58$, $p < .05$, $d = 1.06$.

Replicating the results of Experiment 2, Experiment 3 showed that test-potentiated self-regulated learning can transfer to non-tested items, even after equating total exposure time of tested and non-tested material by permitting participants to restudy the non-tested items during the interim phase. Specifically, non-tested (interim restudy) items were relieved of the foresight bias and were allocated more time than correctly recalled interim-test items. As a consequence of equal exposure time and enhanced study-time allocation, non-tested items were rendered just as recallable during the final recall test as interim-test items.

Experiment 4

Experiment 3 demonstrated that test-potentiated self-regulated learning can transfer to non-tested items if such items are restudied intermixed with items that are tested, even after equating total exposure time of tested and non-tested material. Experiment 4 had two objectives.

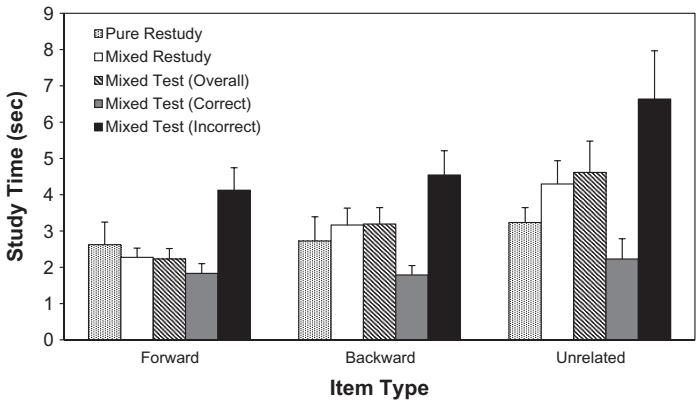

Fig. 9. Mean study-time allocation as a function of item type (forward, backward, and unrelated pairs) following pure interim restudy or interim restudy mixed with testing in Experiment 4. For the mixed-test items, overall study time is reported as well as mean study times conditionalized on whether items were correctly or incorrectly recalled during the interim test. Error bars represent standard error of the means.

First, we sought to replicate and extend the results of Experiment 3 by including a 'pure' restudy condition (like that included in Experiment 1) in which no items were tested during the interim (see Fig. 1). Including this pure restudy condition permits a direct comparison between items that are restudied amongst tested items and those that are not, thus providing an opportunity to examine the robustness of the transfer effects shown in Experiment 3. The second objective of Experiment 4 was to explore a possible mechanism for the transfer effects observed in Experiment 3. Specifically, the interim test may have encouraged participants to engage in covert self-testing during the self-paced restudy phase (e.g., by covering up, or directing attention away from, the target word). To examine this possibility, participants in Experiment 4 were asked the same follow-up question regarding their study strategies as participants in Experiment 1, except that self-testing was added as an answer option.

Method

Participants, design, materials, and procedure

Fifty-two undergraduates at the University of California, Los Angeles (UCLA) participated for partial course credit. Half (26) of the participants were assigned to the pure restudy condition, which was identical to the interim restudy condition in Experiment 1; the other half of the participants were assigned to the mixed restudy/test condition, which was identical to Experiment 3. After the final test was completed, participants were asked a follow-up question regarding their study strategies during the self-paced study phase. This question was identical to that used in Experiment 1, except that the following option was added: "*Self-testing* (testing yourself on the second word by covering it up or directing attention away from it)."

Results and discussion

Study-time allocation

Fig. 9 presents participants' mean subsequent study-time allocation for each item type after pure restudy or restudy mixed with testing. Our primary focus concerned the direct comparison of the non-tested restudied items in each condition ('pure restudy' vs. 'mixed restudy'), and thus we restrict our analyses to these items. We note, however, that the general pattern of results for the other types of items in the mixed restudy/test condition is consistent with those reported in Experiment 3. We included these data in Fig. 9 to convey this point.

We first conducted a 2 (restudy condition: pure restudy vs. mixed restudy) × 3 (item type: forward, backward, unrelated) mixed-model ANOVA, which revealed a main effect of item type, $F(2,100) = 10.72$, $p < .05$, $\eta_p^2 = .18$, and a marginally significant interaction, $F(2,100) = 3.04$, $p = .06$, $\eta_p^2 = .06$. More relevant for current purposes, we conducted a 2 (restudy condition: pure restudy vs. mixed restudy) × 2 (item type: forward vs. backward) mixed-model ANOVA to examine whether the foresight bias was alleviated for mixed restudy items (like in Experiment 3) but not for pure restudy items (like in Experiment 1). Indeed, a reliable interaction was found, $F(1,50) = 4.37$, $p < .05$, $\eta_p^2 = .08$. For mixed-restudy items, backward pairs were studied longer than forward pairs, $t(25) = 3.14$, $p < .05$, $d = .50$, whereas for pure-restudy items, forward and backward pairs were studied for the same duration ($p > .05$).

Final-recall performance

Fig. 10 shows the proportions of items correctly recalled on the final cued-recall test as a function of item type for the pure restudy condition and the mixed restudy/test condition. A 2 (restudy condition: pure restudy vs. mixed restudy) × 3 (item type: forward, backward, unrelated) mixed-model ANOVA revealed that mixed restudy items were recalled better than pure restudy items, $F(1,50) = 5.31$, $p < .05$, $\eta_p^2 = .10$, and a main effect of item type was also found, $F(2,100) = 26.32$, $p < .05$, $\eta_p^2 = .35$. The interaction was not reliable ($p > .05$). Finally—and replicating Experiment 3—no recall differences were found between the mixed restudy items and the mixed test items ($p > .05$).

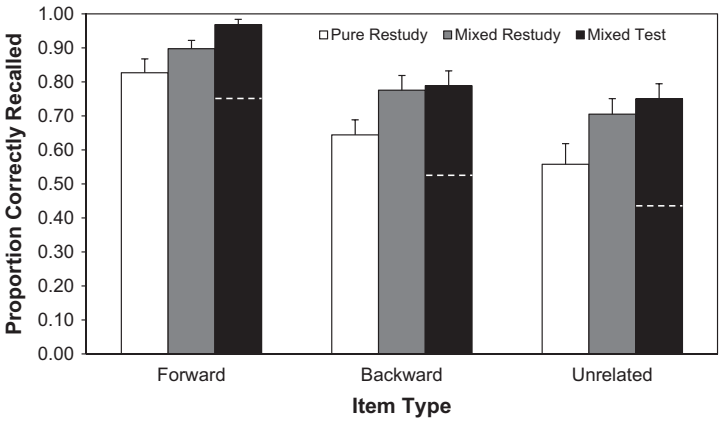

Fig. 10. Mean proportion of items correctly recalled on the final cued-recall test as a function of item type (forward, backward, and unrelated pairs) following pure interim restudy or restudy mixed with interim testing in Experiment 4. For the mixed condition, separate recall bars are presented for items that were restudied during the interim and items that were tested during the interim. The dashed lines embedded in the mixed-test data bars denote interim-recall performance. Error bars represent standard error of the means.

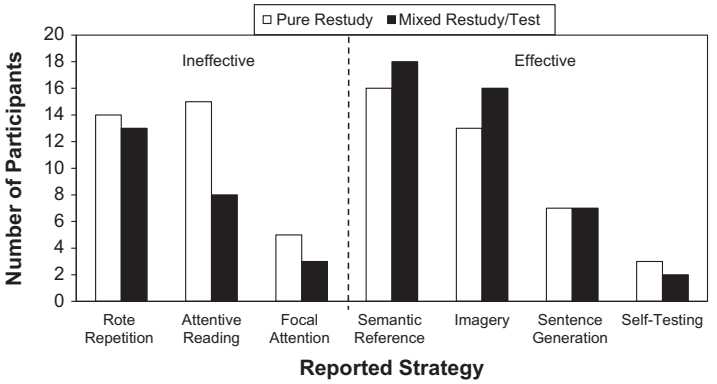

Fig. 11. Number of participants (out of 26 in each condition) who reported using each encoding strategy during the self-paced study phase following pure interim restudy or restudy mixed with interim testing in Experiment 4. Strategies to the left of the vertical dashed have been shown by previous research to be ineffective; strategies to the right of the dashed line have been shown to be effective. Note that participants could indicate using more than one strategy as they were instructed to 'mark all that apply'.

Study strategies

Fig. 11 presents participants' self-reported study strategies used during the self-paced study phase (reported after the final test). Ineffective strategies appeared to be reported more often, in general, by participants in the pure restudy condition than by participants in the mixed restudy/test condition. Conversely, the effective strategies of semantic reference and imagery appeared to be reported relatively more often by participants in the mixed restudy/test condition. Relatively few participants in either condition reported using self-testing.

Experiment 4 replicated and extended the results of Experiment 3 by including a 'pure' restudy condition in which no testing occurred during the interim phase. Similar to Experiment 3, items restudied intermixed with previously tested items were relieved of the foresight bias; whereas, consistent with Experiment 1, pure restudy items were not. Additionally, final recall favored mixed restudy items, which were as recallable during the final test as were items that were tested during the interim. Finally, Experiment 4 explored the potential role of self-testing in the transfer effects shown in Experiment 3. Inconsistent with this account, relatively few participants in the mixed restudy/test condition reported using self-testing during the self-paced study phase.

General discussion

Four experiments showed that prior testing potentiates self-regulated learning and, in particular, alleviates the foresight bias (Koriat & Bjork, 2005) when using study time as a proxy for predictive judgments (cf., Koriat & Bjork, 2006b). Experiment 1 revealed that learners, after an interim-test, selectively directed their restudy efforts to items that were not successfully recalled on the interim test with study time for these items increasing monotonically with item difficulty. Furthermore, the interim-test group became sensitive to the subtle, yet important, distinction between forward and backward paired associates by

allocating more restudy time to backward pairs than forward pairs, thus replicating Koriat & Bjork's (2006b) finding that study-test practice can alleviate the foresight bias. Participants in the interim-restudy condition, by contrast, were not sensitized to the difference between forward and backward pairs and did not restudy items any longer than interim-test participants restudied previously recalled items. Consequently, final recall favored the interim-test group across all three types of items (forward, backward, and unrelated pairs).

The interim-test group in Experiment 1 also reported using more effective strategies during the self-paced study phase than the interim-restudy group, a finding that aligns with the overall study time advantage of the interim-test group because effective strategies presumably require more time to implement than ineffective strategies. That taking an interim test engendered more effective subsequent encoding strategies is consistent with Pyc and Rawson's (2010; Pyc and Rawson, 2012) mediator-shift hypothesis, which states that retrieval failures encourage learners to shift to using more effective mediators during a restudy opportunity. Briefly, Pyc and Rawson (2012) had participants study Swahili–English word pairs, generating and reporting associative keywords for each pair (e.g., for *Mshoni–Tailor*, the learner might generate the keyword *shoe* because *Mshoni* sounds like shoe and a *Tailor* makes them). Items were then presented for either test-restudy practice or restudy practice alone, with participants again reporting the keyword used to encode each pair during each restudy opportunity. Supporting the mediator-shift hypothesis, a greater proportion of keywords were modified after test-restudy practice than restudy practice alone, with most of the keyword modifications occurring after retrieval failures. Couched in this language, our results make sense given that the interim-test group experienced retrieval failures during the prior interim test, whereas the interim-restudy group did not. Such retrieval failures may have enabled interim-test participants to evaluate and modify the strategies they used to originally encode the items (see Bahrick & Hall, 2005, for a similar strategy-shift account proposed to explain spacing effects).[3]

Can the test-enhanced study-time allocation demonstrated in Experiment 1 transfer to previously presented, non-tested material? Given the results from Experiments 2, 3, and 4, the answer is 'yes'. Specifically, non-tested items, when restudied amongst items that had been tested, were relieved of the foresight bias and were appropriated more study time than correctly recalled interim-test items. This was the case when non-tested items were replaced with math problems during the interim (Experiment 2) and when total exposure time of tested and non-tested items was equated (Experiments 3 and 4). That subsequent self-paced study of non-tested items was enhanced manifested in final recall performance. In Experiment 2, final recall only narrowly favored interim-tested items, which is impressive given that the non-tested items were only studied twice, whereas interim-tested items were not only studied twice but were also afforded additional exposure time during the interim test. When this unfair advantage was addressed in Experiments 3 and 4 by permitting participants to restudy the non-tested items during the interim phase, final recall of the non-tested items was equivalent to that of the interim-tested items. Thus, our results extend previous research showing that a testing experience can benefit the learning of non-tested material (e.g., Chan et al., 2006; Little et al., 2012; Wissman et al., 2011).

The differences in study-time allocation for non-tested items in Experiment 1 compared to Experiments 2, 3, and 4 speak to the notion that a testing experience fosters metacognitive sophistication among learners. Without a testing experience, learners are prone to metacognitive illusions, typically thinking that they know more than they actually do (for reviews, see Bjork, 1999; Bjork et al., 2013). One metacognitive illusion of particular interest in the current study was the foresight bias, an illusion of competence that arises from information being present during study but absent, yet solicited, at test (see Koriat & Bjork, 2005). As previously discussed, Koriat and Bjork (2006b) showed that the experience-based debiasing procedure of study-test practice alleviated the foresight bias for tested items (i.e., backward pairs were allocated more subsequent study time than forward pairs), but not for new items, concluding that testing alone does not help learners formulate a general rule that can transfer beyond tested information. However, results from the present Experiments 2, 3, and 4 cast doubt on such a strong conclusion. For non-tested items in those experiments, participants allocated more restudy time to backward pairs than to forward pairs, providing the first empirical evidence, to our knowledge, that experience-based debiasing of the foresight bias is not restricted to tested items. Thus, while it may be the case that study-test practice does not alleviate the foresight bias for *new* items, our results suggest that a testing experience does seem to equip the learner with a rule regarding forward and backward associates that can be applied to previously presented, non-tested items, provided that the non-tested items are restudied in the context of items that were tested.

Admittedly, there may be alternative explanations for our transfer effects. One possibility that was examined in Experiment 4 was that the interim test encourages participants to engage in covert self-testing during the self-paced restudy phase (e.g., by covering up, or directing attention away from, the target word). Such an account would be consistent with our finding that study time increased monotonically with item difficulty, and that final recall only narrowly favored interim-tested items in Experiment 2 and was equivalent across conditions in Experiment 3. However, given that relatively few participants reported using self-testing during the self-paced restudy phase in Experiment 4, we regard it as an unlikely mechanism responsible for the current transfer effects. This result

[3] Unlike Pyc and Rawson (2012), we solicited study strategies only once during the experiment and thus recognize that no strong conclusions regarding strategy shifts, *per se*, can be made. However, we find it reasonable to assume that by randomly assigning participants into our two conditions, the interim-test and interim-restudy groups most likely did not differ in their study strategies employed during the first study phase. Thus, the use of more effective strategies reported by the interim-test group does, in all likelihood, reflect a shift in study strategies.

is consonant with Karpicke's (2009) finding that learners, when regulating their own learning after retrieval practice, by and large do not choose to practice retrieval during subsequent study phases despite the fact that such a strategy would bolster learning. Nevertheless, we acknowledge that the explanation for our transfer effects—whether they can be attributed to a learned rule, covert self-testing, or some other mechanism—is not entirely clear and should be the focus of future research.

Another, more general metacognitive illusion was alleviated for non-tested items in the current study. Consider the finding that participants in the interim-restudy group in Experiment 1 restudied items no longer than interim-test participants restudied correctly recalled items. Such short study times in the interim-restudy group may have reflected an illusion of knowing, whereby having already been familiarized with the items during two subsequent study phases, participants experienced a heightened sense of processing fluency upon the items' third presentation and, consequently, erroneously predicted that the items had already been learned.[4] Such an illusion was ameliorated in Experiments 2, 3, and 4, in which all participants were tested on only half of the items during the interim. Specifically, restudy times for non-tested items fell in between restudy times for previously recalled and unrecalled interim-test items, reflecting what may have been a sensible strategy by the learner. To expand on this possibility, interim-test items were either previously recalled or not, which effectively informed learners of the items that needed further study. For non-tested items, however, the learner did not know whether these items would have been recalled *had* they been tested, an uncertainty that may have been reflected in participants studying non-tested items longer than previously recalled items, but not as long as previously unrecalled items.

Comparing the differences in study time allocation and final recall in Experiments 1 and 3, we are reminded of the powerful influence of one experimental factor: namely, whether variable(s) of interest are manipulated between- or within-subjects (see McDaniel & Bugg, 2008, for a general discussion on this issue). Experiments 1 and 3 were identical except that interim-testing vs. interim-restudying was manipulated between-subjects in Experiment 1 (i.e., one group of participants was tested on the items, whereas another group restudied the items) and within-subjects in Experiment 3 (i.e., all participants were tested on half of the items and restudied the remaining items). Unlike in Experiment 1, non-tested items in Experiment 3 profited from enhanced self-regulated learning, boosting their final recall to the level of tested items, thus eliminating the final recall advantage typically observed for tested items. Future research might examine the durability of these recall results, especially given that testing effects are often only revealed after relatively long retention intervals (see Roediger & Karpicke, 2006).

[4] Given that interim-restudy participants in Experiment 1 allocated more time to unrelated pairs than both forward and backward pairs suggests that the generally short restudy times in this group cannot be attributed to fatigue. If this were the case, we would expect short restudy times that were insensitive to item difficulty.

That a final recall advantage was observed for tested items between-subjects (Experiment 1), but not within-subjects (Experiments 3 and 4), is consonant with work on a closely related phenomenon, the generation effect, which is the finding that information generated from memory (e.g., generating the opposite word when presented with *hot-???*) is better remembered than information that is simply read (for a review, see Bertsch, Pesta, Wiscott, & McDaniel, 2007; Slamecka & Graf, 1978). DeWinstanley and Bjork (2004; see also Bjork, DeWinstanley, & Storm, 2007) had participants first read a passage that included both to-be-generated and to-be-read information, which resulted in learners experiencing the generation effect. Next, participants read a new passage—again containing to-be-generated and to-be-read information—but this time no generation effect materialized because learners applied a generation-based strategy to the to-be-read information. Critically, participants who were denied the experience of the generation effect did not show the enhanced encoding of subsequent to-be-read items. Thus, the experience of generation—like the experience of testing in the current study—enhanced subsequent encoding.

Given that participants in the current study chose to devote relatively more subsequent study time to previously unrecalled items and that study times were, in general, sensitive to item difficulty, we note that our results are generally consistent with extant theories of study time allocation—specifically, the discrepancy-reduction model (Dunlosky & Hertzog, 1998; Dunlosky & Thiede, 1998) and the region of proximal learning model (Metcalfe, 2002; Metcalfe & Kornell, 2005)—both of which assert that learners will allocate relatively more time to information that is perceived to be unlearned. The current study provides evidence that explicit memory tests are particularly effective in equipping learners with knowledge regarding what has and has not been learned and shows that such knowledge can inform subsequent study of non-tested material. Furthermore, that final recall profited from participants selectively directing their restudy efforts to unlearned material is consistent with previous work showing that learners can benefit more from making their own study decisions than if those decision are made for them (Kornell & Metcalfe, 2006; Nelson et al., 1994).

Concluding comment

In a review of the literature on study-time allocation, Son and Kornell (2008) noted, "The most important objective of research on study time allocation. . .is to uncover ways of improving efficiency" (p. 348). We agree with this statement and have uncovered one such way—namely, by providing learners with an interim-testing experience. Following such an experience, learners subsequently studied more efficiently and effectively than they did when no such test was taken. Furthermore, such test-potentiated self-regulated learning—including the alleviation of the foresight bias—can transfer to non-tested material when that material is restudied in the context of items that are tested. To illustrate a practical implication of the current work, consider an instructor who, after giving a full lecture, quizzes

his or her students on a subset of that lecture's content before dismissing the students from class. Our data suggest a provocative conclusion—namely, that when the students restudy that lecture content on their own, their self-regulated learning will be enhanced—for both the quizzed and non-quizzed material—compared to if no quiz was given.

Acknowledgments

Grant 29192G from the James S. McDonnell Foundation supported this research. We thank Lakshan Fonseka and Gayan Seneviratna for their help with data collection, and members of CogFog for insightful comments regarding this research.

References

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language, 49*, 415–445. http://dx.doi.org/10.1016/j.jml.2003.08.006.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1063–1087. http://dx.doi.org/10.1037/0278-7393.20.5.1063.

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General, 138*, 432–447. http://dx.doi.org/10.1037/a0015928.

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 940–945. http://dx.doi.org/10.1037/a0029199.

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language, 52*, 566–577. http://dx.doi.org/10.1016/j.jml.2005.01.012.

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition, 31*, 297–305. http://dx.doi.org/10.3758/BF03194388.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*, 55–68. http://dx.doi.org/10.1037/0096-3445.127.1.55.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*, 201–210. http://dx.doi.org/10.3758/BF03193441.

Bjork, E. L., DeWinstanley, P. A., & Storm, B. C. (2007). Learning how to learn: Can experiencing the outcome of different encoding strategies enhance subsequent encoding? *Psychonomic Bulletin & Review, 14*, 207–211. http://dx.doi.org/10.3758/BF03194053.

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative regency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII. Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444. http://dx.doi.org/10.1146/annurev-psych-113011-143823.

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*, 279–283. http://dx.doi.org/10.1177/0963721412452728.

Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition, 36*, 429–437. http://dx.doi.org/10.3758/MC.36.2.429.

Castel, A. D., McCabe, D. P., & Roediger, H. L. III, (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning. *Psychonomic Bulletin & Review, 14*, 107–111. http://dx.doi.org/10.3758/BF03194036.

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61*, 153–170. http://dx.doi.org/10.1016/j.jml.2009.04.004.

Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory, 18*, 49–57. http://dx.doi.org/10.1080/09658210903405737.

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. III, (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*, 553–571. http://dx.doi.org/10.1037/0096-3445.135.4.553.

DeWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition, 32*, 945–955. http://dx.doi.org/10.3758/BF03196872.

Dunlosky, J., & Ariel, R. (2011). Self-regulated learning and the allocation of study time. In B. Ross (Ed.). *Psychology of learning and motivation* (Vol. 54, pp. 103–140). San Diego, CA US: Elsevier Academic Press.

Dunlosky, J., & Hertzog, C. (1998). Training programs to improve learning in later adulthood. Helping older adults educate themselves. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in education theory and practice* (pp. 249–275). Mahwah, NJ: Erlbaum.

Dunlosky, J., & Thiede, K. W. (1998). What makes people study more? An evaluation of factors that affect self-paced study. *Acta Psychologica, 98*, 37–56. http://dx.doi.org/10.1016/S0001-6918(97)00051-6.

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 238–244. http://dx.doi.org/10.1037/0278-7393.33.1.238.

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition, 40*, 505–513. http://dx.doi.org/10.3758/s13421-011-0174-0.

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

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 290–296. http://dx.doi.org/10.1037/a0028468.

Hertzog, C., & Dunlosky, J. (2004). Aging, metacognition, and cognitive control. *The psychology of learning and motivation: Advances in research and theory* (Vol. 45, pp. 215–251). San Diego, CA US: Elsevier Academic Press.

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 22–34. http://dx.doi.org/10.1037/0278-7393.29.1.22.

Izawa, C. (1966). Reinforcement-test sequences in paired-associate learning. *Psychological Reports, 18*, 879–919. http://dx.doi.org/10.2466/pr0.1966.18.3879.

Izawa, C. (1968). Effects of reinforcement, neutral and test trials upon paired-associate acquisition and retention. *Psychological Reports, 23*, 947–959. http://dx.doi.org/10.2466/pr0.1968.23.3.947.

Izawa, C. (1970). Optimal potentiation effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83*, 340–344. http://dx.doi.org/10.1037/h0028541.

Izawa, C. (1971). The test trial potentiating model. *Journal of Mathematical Psychology, 8*, 200–224. http://dx.doi.org/10.1016/0022-2496(71)90012-5.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *The American Journal of Psychology, 93*, 329–343. http://dx.doi.org/10.2307/1422236.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*, 469–486. http://dx.doi.org/10.1037/a0017341.

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

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

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 187–194. http://dx.doi.org/10.1037/0278-7393.31.2.187.

Koriat, A., & Bjork, R. A. (2006a). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition, 34*, 959–972. http://dx.doi.org/10.3758/BF03193244.

Koriat, A., & Bjork, R. A. (2006b). Mending metacognitive illusions: A comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 1133–1145. http://dx.doi.org/10.1037/0278-7393.32.5.1133.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*, 147–162. http://dx.doi.org/10.1037/0096-3445.131.2.147.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133*, 643–656. http://dx.doi.org/10.1037/0096-3445.133.4.643.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 609–622. http://dx.doi.org/10.1037/0278-7393.32.3.609.

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17*, 493–501, DOI: 10.1080/09658210902832915.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 989–998. http://dx.doi.org/10.1037/a0015729.

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science, 23*, 1337–1344. http://dx.doi.org/10.1177/0956797612443370.

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review, 15*, 237–255. http://dx.doi.org/10.3758/PBR.15.2.237.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131*, 349–363. http://dx.doi.org/10.1037/0096-3445.131.3.349.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*, 463–477. http://dx.doi.org/10.1016/j.jml.2004.12.001.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. Available at http://w3.usf.edu/FreeAssociation/.

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*, 267–270. http://dx.doi.org/10.1111/j.1467-9280.1991.tb00147.x.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect". *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 676–686. http://dx.doi.org/10.1037/0278-7393.14.4.676.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and some new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 125–173). New York: Academic Press.

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science, 5*, 207–213. http://dx.doi.org/10.1111/j.1467-9280.1994.tb00502.x.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*, 335. http://dx.doi.org/10.1126/science.1191465.

Pyc, M. A., & Rawson, K. A. (2012). Why is test-restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 737–746. http://dx.doi.org/10.1037/a0026166.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying Judgments of Learning (JOLs) on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137*, 131–148. http://dx.doi.org/10.1037/a0021705.

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

Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mestre & B. Ross (Eds.), *Psychology of learning and motivation: Cognition in education* (pp. 1–36). Oxford: Elsevier.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 592–604. http://dx.doi.org/10.1037/0278-7393.4.6.592.

Soderstrom, N. C., & McCabe, D. P. (2011). The interplay between value and relatedness as bases for metacognitive monitoring and control: Evidence for agenda-based monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 1236–1242. http://dx.doi.org/10.1037/a0023548.

Son, L. K., & Kornell, N. (2008). Research on the allocation of study time: Key studies from 1890 to the present (and beyond). In J. Dunlosky & R. A. Bjork (Eds.), *A handbook of memory and metamemory* (pp. 333–351). Hillsdale, NJ: Psychology Press.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 204–221. http://dx.doi.org/10.1037/0278-7393.26.1.204.

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences USA, 110*, 6313–6317.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. III, (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1392–1399. http://dx.doi.org/10.1037/a0013082.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review, 18*, 1140–1147. http://dx.doi.org/10.3758/s13423-011-0140-7.

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society, 15*, 41–44.