Mem Cogn
DOI 10.3758/s13421-016-0664-1

Updating metacognitive control in response to expected retention intervals

Joshua L. Fiechter[1] · **Aaron S. Benjamin**[1]

Abstract In five experiments, we investigated whether expected retention intervals affect subjects' encoding strategies. In the first four experiments, our subjects studied paired associates consisting of words from the Graduate Record Exam and a synonym. They were told to expect a test on a word pair after either a short or a longer interval. Subjects were tested on most pairs after the expected retention interval. For some pairs, however, subjects were tested after the other retention interval, allowing for a comparison of performance at a given retention interval conditional upon the expected retention interval. No effect of the expected retention interval was found for 1 min versus 4 min (Exp. 1), 30 s versus 3 min (Exp. 2), and 30 s versus 10 min (Exps. 3 and 4), even when subjects were given complete control over the pacing of study items (Exp. 4). However, when the difference between the expected retention intervals was increased massively (10 min vs. 24 h; Exp. 5), subjects remembered more items that they expected to be tested sooner, an effect consistent with the idea that they traded off efforts to remember items for the later test versus items that were about to be tested. Overall, this set of results accords with much of the test-expectancy literature, revealing that subjects are often reluctant to adjust encoding strategies on an item-by-item basis, and when they do, they usually make *quantitative*, rather than *qualitative*, adjustments.

Keywords Metamemory · Memory · Decision making

✉ Joshua L. Fiechter
fiechte2@illinois.edu

[1] University of Illinois at Urbana-Champaign, Urbana, IL, USA

Published online: 21 October 2016

In an educational setting, students often inquire about various properties of an upcoming test. Many inquiries regard the test format. That is, students want to know if an exam will contain multiple-choice questions, fill-in-the-blank questions, or essay questions. A related consideration that has received no attention in the metacognitive literature is the anticipated timing of a test. If a student is studying for a test scheduled for tomorrow or for one week from now, do they prepare differently?

Ideally, students would attempt to space their study over the days leading up to a test. The benefits of distributed learning over massed learning are firmly established (Benjamin & Tullis, 2010; Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). But competing agendas—and perhaps also a lack of planning—often impede the implementation of ideal study plans. The final weeks of a semester are especially fraught with difficulty, when the demands from multiple classes must be met in a short period of time and students cannot afford the time to schedule multiple study opportunities for a given class. That is, they must pre-allocate study time for a certain day, while other obligations fill up the remaining days on their schedule. Sometimes these study periods are in close proximity to the test day, and at other times they are far removed. That is, the retention interval (RI) between a student's study session and the upcoming test may be of long or short duration. Experiments manipulating RI are plentiful, though little is known about learners' responses to expected RIs. In the present research, we sought to answer the question of whether learners make changes to their study habits if they expect a long or a short RI.

Metacognitive monitoring and control

Learners exert multiple forms of metacognitive control over self-guided learning, often with success (Benjamin, 2008; Finley, Tullis, & Benjamin, 2010; Koriat, Ma'ayan, &

Nussinson, 2006; Kornell & Metcalfe, 2006; Mazzoni & Cornoldi, 1993; Tullis & Benjamin, 2011), but they are generally reluctant to make wholesale changes to their encoding strategies (Fiechter, Benjamin, & Unsworth, 2016). They are, however, willing to spend more effort, resources, or time on materials that they deem to be more difficult, a principle called *discrepancy reduction* (Dunlosky & Hertzog, 1998). According to this view, an item is studied until it meets a goal level of learning, which is called the *norm of study* (Le Ny, Denhiere, & Taillanter, 1972). As an item is being studied, strategies are monitored and updated to ensure that the study is effective, suggesting that learners will expend more resources on the most difficult items. Discrepancy reduction is easy to see in paradigms in which study time is controlled by the learner, because greater study time is typically accorded to items or conditions known to be more difficult (Son & Metcalfe, 2000). However, it is also apparent in paradigms with more elaborate encoding strategies. For example, studies have assessed strategy use in a massed-versus-spaced learning paradigm and have shown that learners use the superior strategy—spacing—for more difficult words (e.g., Benjamin & Bird, 2006; Toppino, Cohen, Davis, & Moors, 2009).

Although some studies have examined the implications of discrepancy reduction for strategy use, most research has investigated the amount of study time allocated to material. Study time is easy to measure and likely correlates with the effort expended on a given item. As we noted above, subjects will generally choose to spend more time on difficult items, just as the discrepancy reduction hypothesis predicts (Belmont & Butterfield, 1971; Bisanz, Vesonder, & Voss, 1978; Dufresne & Kobasigawa, 1989; Kobasigawa & Metcalf-Haggert, 1993; Le Ny et al., 1972; Masur, McIntyre, & Flavell, 1973). For instance, Le Ny et al. presented subjects with paired three-digit numbers and letters (i.e., 446:Z). The difficulty of the items was manipulated by having the three-digit numbers appear very similar to one another, somewhat similar, or not similar at all. Subjects dedicated the most study time to stimuli that were very similar, and the least study time to stimuli that were not similar. Tullis and Benjamin (2011) qualified the traditional discrepancy reduction findings. They found that not all subjects chose to focus on more difficult items. However, only those subjects that adopted a discrepancy reduction approach toward study performed better than a control group that did not have control over their study time. Subjects that failed to adopt a discrepancy reduction approach performed no better than the subjects in the control group. This result suggests that more-effective learners choose to focus on more-difficult items.

Learners also spend more study time on items that they deem to be more difficult than others, regardless of normative difficulty (Cull & Zechmeister, 1994; Koriat et al., 2006; Mazzoni & Cornoldi, 1993; Mazzoni, Cornoldi, & Marchitelli, 1990; Nelson, Dunlosky, Graf, & Narens, 1994; Nelson & Leonesio, 1988). For example, Metcalfe and Finn (2008) found that learners choose to restudy items that they erroneously perceive as being more difficult than other items. Over two study sessions, they had subjects study word pairs either one time and then three times, or three times and then one time. A cued-recall test followed each study session. The items studied three times in the first session were given higher judgments of learning (JOLs), even though recall performance on the second test was equal across all items (subjects saw all items four times over the course of the two study sessions). Critically, the items that received lower JOLs in the second study session were also more frequently selected for (hypothetical) restudy. Subjects' study choices were influenced by their misguided JOLs and not by their actual learning. Also of interest is the finding that learners across the spectrum have been shown to study items they deem to be the most difficult (Cull & Zechmeister, 1994). That is, both good and poor learners conform to the predictions of the discrepancy reduction hypothesis, in spite of their differences in memory performance.

Task constraints Not all findings on study time allocation are explained by the discrepancy reduction hypothesis. For instance, Thiede and Dunlosky (1999) manipulated subjects' performance goals. They presented subjects with 30 word pairs and gave them a goal of recalling either six items (easy goal) or 24 items (difficult goal) on an upcoming cued-recall test. After studying all the pairs, the subjects were to select those that they would like to further restudy. Subjects chose to restudy difficult items if they had a difficult goal, or to restudy easy items if they had an easy goal. Similarly, Son and Metcalfe (2000, Exp. 1) presented subjects with eight biographical essays, each six pages in length. The subjects were given only 30 min to look through the 48 pages of essays before taking a fill-in-the-blank test. With such limited time, subjects chose to study the essays that they had judged to be the easiest to learn. This result was also in contrast to what the discrepancy reduction hypothesis predicts. These findings indicate that performance goals, as dictated by task constraints, provide an important boundary condition to the application of a discrepancy reduction strategy.

Test expectancy

Returning to our earlier example, students also utilize knowledge of their upcoming assessment to determine how and how much they study. The experimental analogue for such a situation is the *test-expectancy* paradigm, a methodology of creating expectations in the subject and then testing in a way that either conforms or is opposite to their expectations. Generally, subjects are led to expect a given test attribute in one of two ways, either through simple instruction or by a series of study–

test sessions in which each test possesses some critical attribute. A test then follows that is the same or is different from what the subject had been led to expect. Performance is then compared on tests that possess the same attribute, conditionalized on what the subject was expecting prior to taking the test.

Much of the research in the test-expectancy paradigm has focused on the differences in recall test and recognition test expectancy. Given students' interest in the features of an upcoming test, it seems that the expected test format would be very influential in how they conducted their study. If a subject expects a recall test, he or she should study in a way that benefits performance on a recall test. Likewise, if a subject expects a recognition test, he or she should study such that performance would be maximal on a test of recognition. But this result is almost never found; rather, the most common finding in this literature is that expectation of a recall test leads to better performance on both a recall test and a recognition test (Balota & Neely, 1980; Hall, Grossman, & Elwood, 1976; Leonard & Whitten, 1983; Neely & Balota, 1981; Schmidt, 1988; Thiede, 1996; von Wright, 1977; von Wright & Meretoja, 1975).

Although this result provides evidence that subjects are attending to and taking into account differences in test format, it suggests that those expecting a recall test may be trying harder to learn material, because a recall test is presumed to be more demanding than a recognition test. This change can be considered a *quantitative* shift, since the subject merely appears to be applying more of the same strategy to the more difficult material. A quantitative change suggests that students do not fully use information about an upcoming test to their advantage. If students were using information about a test to optimize their study, they would be making *qualitative* changes in their encoding. Here, the subject would tailor his or her study habits to encode information in a manner that was optimal for the expected test format. We consider both quantitative and qualitative updates to encoding to be strategies; the former requires only differential effort or time, whereas the latter requires a change in approach.

Convincing evidence for qualitative changes in study would require a disordinal interaction, such that subjects expecting a given type of test would perform better on that test format than would subjects expecting another test format. Finley and Benjamin (2012) did find evidence of such qualitative changes in study. They gave subjects four study–test phases, with all four tests being either the cued or the free recall of target words. The stimuli were related and unrelated word pairs, distributed equally across all lists. Subjects then had a fifth study–test phase, in which the test format either was what they had been induced to expect or had been switched. This fifth test was the critical assessment. Performance on the cued-recall test was better for subjects expecting a cued-recall test than for subjects expecting a free-recall test. Likewise, performance on the free-recall test was better for subjects expecting a free-recall test than for subjects expecting a cued-recall test. Furthermore, the subjects expecting a free-recall test allocated study time less on the basis of the relatedness of the word pair than did subjects expecting a cued-recall test. That is, as subjects experienced a free-recall test format, they learned that the association of word pairs was not a helpful feature of the stimuli to focus on; by the third study–test phase, they were spending equal amounts of study time on both related and unrelated word pairs.

Expected retention intervals

The timing of a test is an obvious feature that may influence study strategies. A standard (but now woefully outdated) classroom example for introducing the concept of short-term memory involves repeating a phone number as one makes one's way from a phone book to a phone across the room. In the metacognitive framework laid out by Nelson and Narens (1990), the authors make specific mention of the anticipated RI: "When a delay is expected to occur between acquisition and the retention test, then the person's *theory of retention* . . . is used to modulate how well each item would have to be mastered now, in order for it to still be remembered on the retention test" (pp. 129–130). In the present research, we sought to investigate the effects of the expected RI on subjects' study strategies. To that end, the methodology was similar to past research in the test-expectancy paradigm. The only difference was different RIs rather than different test formats. The predictions for the present research can be drawn in a straightforward way from the discrepancy reduction hypothesis, with the idea that what will make one item more "difficult" than another is the expected RI for a given item.

We anticipated four potential outcomes in subjects' performance patterns. The first was that they would perform uniformly better on words expected to be tested at a long RI, independent of the actual RI. This outcome would reflect a discrepancy reduction policy in which subjects reserved their most effective strategies or the most resources for the items that will be the most difficult at test. The second was that subjects would perform better on words expected to be tested at a short RI, independent of the actual RI. Overall superior performance on easier items would indicate that the task constraints pushed subjects toward the adoption of a strategy in which they reserved their resources for the easiest items (cf. Metcalfe & Kornell, 2003). The third outcome of interest was that subjects would perform best on words that were tested at the expected RI, as compared to words tested at an unexpected RI, particularly for the short RI (cf. Finley & Benjamin, 2012). This pattern would indicate a more sophisticated qualitative shift in encoding strategy across conditions; here, subjects

would be tailoring their study to meet the anticipated RI demands. That is, they would not be focusing on some items to the detriment of others, but rather would be implementing different strategies tuned for different RIs. Thus, a violation of their expectations would result in diminished performance when the expected and veridical RIs were at odds. In this scenario, performance would be more likely to differ at the short than at the long RI, since very few study methods are effective for a long interval but not effective for a short interval. Finally, we considered that subjects might not do anything in response to the retention intervals, and consequently we would not observe any differences in performance based on the expected RI.

In selecting our RIs, we sought to create an environment that would foster qualitative strategy shifts for our participants. Thus, we wanted a short RI that participants felt could be managed with relatively shallow processing (e.g., repeating the item over and over), but the long RI to be sufficiently long to encourage more elaborate processing (e.g., relating the items to something of personal significance). Our initial choice of RIs was 1 versus 4 min. We chose these intervals as our starting point since our initial set of stimuli (Graduate Record Exam [GRE] words paired with synonyms) were expected to be difficult for our participants to remember; consequently, we expected substantially more forgetting 4 min after study than 1 min after study, even though the absolute duration of either interval was fairly short. Put another way, our choice of RIs was dictated by the anticipated timescale of participants' forgetting. We chose increasingly disparate RIs, in terms of both proportionality (30 s vs. 3 min) and absolute difference (30 s vs. 10 min, 10 min vs. 24 h) over the course of our five experiments, to make the different durations even more salient to our participants.

Experiment 1

Method

Subjects

Sixty-eight students enrolled in an introductory psychology course at the University of Illinois at Urbana-Champaign participated for partial course credit.

Design

The experiment used a 2 × 2 within-subjects design. The independent variables were actual RI (1 or 4 min) and expected RI (1 or 4 min). Because test trials were self-paced and also interleaved with study trials, the RI length varied between items. On average, the short RI lasted 1 min 18 s, and the long RI lasted 5 min. The dependent variable was performance on the cued-recall test trials. Performance was collapsed across individual items to get a percent correct measurement for each subject.

Materials

Stimuli The stimuli were 140 paired associates consisting of words from the GRE and a synonym. All of the pairs were similarly difficult. Of the 140 word pairs, 100 were tested. The remaining 40 items were used in filler study trials so that subjects were occupied during the RIs near the end of the list. The study and test trials were interleaved such that the RI between study and test for a given word pair was filled with study and test trials for other word pairs.

Study and test trial schedule One list was compiled that randomly determined the order of the study and test trials. This list was a combination of two lists, each consisting of 70 study trials and 50 test trials, ensuring that the same numbers of study and test trials ensued over the two halves of the experiment. Furthermore, the list was constructed such that no more than five test trials occurred consecutively. Each participant received this randomly compiled list of study and test trials with word pairs randomly selected to appear on a given study trial. Thus, all participants studied and were tested at identical points in the experiment, but the word pairs that they studied and tested on differed at random.

Procedure

Subjects participated individually, in small rooms containing a single desktop computer. They were told that they would be studying word pairs for an upcoming memory test. They were informed that they would see two kinds of screens over the course of the experiment: one would be a word pair, and one would be a word placed over an empty box. Subjects were then instructed that when they saw a word pair, they were to study the pair for a later test, and when they saw a word placed above an empty box, they were to type the word that was paired with the provided word. The typed response appeared in the empty box.

Subjects were told that along with each word pair they would receive "hints" that would let them know how long they had until a word pair would be tested. These hints were time cues that indicated "1 Minute" or "4 Minutes" until a word pair would be studied. Unbeknownst to the subject, these cues were switched 20 % of the time, such that the word pairs cued for 4 min would then really be tested in 1 min, and the word pairs cued for 1 min would really be tested in 4 min. The switched cues provided the key manipulation of testing subjects at either the expected RI or the unexpected RI. Also unbeknownst to the subjects was that some of the studied pairs were not going to be tested. These untested pairs were used for

filler study trials during the RIs when all of the to-be-tested pairs had been presented.

All text displayed during the experiment was in 50-point Arial font. Time cues were presented at the top center of the computer screen, and word pairs were presented in the middle of the computer screen. During test trials, the cue word was presented in the middle of the screen with an empty black box below it. Subjects typed the target word into the black box. Their typed response was displayed on screen in the black box.

Each RI cue preceded its accompanying word pair by 500 ms, and then remained on the screen while the word pair was presented to the subject for 6,000 ms. The screen then went blank for 2,000 ms. The test trials were subject-paced, and the experiment was programmed using the MATLAB programming software.

Results

The findings from Experiment 1 are presented in Fig. 1. Because the data appear to support a null effect—and because the null effect is both meaningful and interpretable—we analyzed these data and the data from all of the experiments discussed here using Bayesian analyses. One advantage of Bayesian analyses versus traditional null hypothesis significance testing (NHST) is that Bayesian statistics evaluate how closely the data fit both the null and the alternative hypotheses. Thus, Bayesian analyses allow us to demonstrate support for a null effect if a preponderance of the evidence supports it (Gallistel, 2009; Rouder, Speckman, Sun, Morey, & Iverson, 2009). Critically, NHST does not allow evidence to accumulate in favor of the null; it instead assumes that the null is true and then assesses the probability of observing the data on the basis of that assumption. Thus, in cases in which $p > .05$, NHST only allows for conclusions of "failing to reject" the null hypothesis, rather than evaluating the null's veracity.

One form of Bayesian analysis returns a Bayes factor (B_{01}), which is a ratio of the marginal likelihoods for the null and alternative hypotheses. Although there are no critical values of B_{01} that indicate when we should declare the existence—or lack of—an effect, Jeffreys (1961) provided guidelines for interpreting B_{01}: A value greater than 3 indicates *some evidence*, a value greater than 10 indicates *strong evidence*, and a value greater than 30 indicates *very strong evidence*. All reported Bayes factors will be reported in terms of the odds in favor of the null.

Bayesian analyses depend on priors, which may be expressed as probability distributions of the anticipated null and alternative outcomes. The selection of priors is not a straightforward task and is left to the discretion of the analyst. For our analyses, we followed the recommendations by Rouder and colleagues (2009), who proposed the Jeffreys–Zellner–Siow (JZS) alternative prior. The JZS prior uses a Cauchy-distributed range of standardized effect sizes (scaled by a factor of $\sqrt{2}/2$) for the alternative hypothesis. This alternative prior is objective, meaning that it relies on minimal assumptions about the distribution of effect sizes under the alternative hypothesis. Objective alternative priors are desirable because of their greater potential for generalizability across forms of the alternative hypothesis, yet still place more probability on large effect sizes for the alternative than for the null prior.

We calculated three Bayes factors: one for the effect of the actual RI, and one for the effect of expected RI at each actual RI. For actual RI, we obtained a $B_{01} = 2.22 \times 10^{-9}$, indicating very strong evidence in favor of the alternative. As expected, subjects recalled a higher percentage of the target words when tested after the 1-min RI ($M = 31$ %, $SD = 19$ %) than after the 4-min RI ($M = 18$ %, $SD = 17$ %). At the 4-min RI, subjects recalled 18 % ($SD = 16$ %) of items for which they expected a 4-min RI, and they recalled 18 % ($SD = 19$ %) of items for which they expected a 1-min RI, $B_{01} = 7.16$. At the 1-min RI, subjects recalled 31 % ($SD = 23$ %) of items for which they expected a 4-min RI, and they recalled 31 % of items ($SD = 20$ %) for which they expected a 1-min RI, $B_{01} = 7.42$. Both B_{01} values in favor of the null indicate the higher end of "some evidence" in favor of expectation having no effect on performance.

Discussion

Only the actual RI had an effect on subjects' performance. Clearly subjects were not making adaptive changes to their metacognitive control in response to the expected RIs. Perhaps the two expected intervals did not seem distinct to the point at which subjects would alter their encoding strategies. Or perhaps the long interval was too challenging for performance to be amenable to any changes in strategy. That is, performance at the 4-min RI may have been impervious to changes in strategy because recalling items 4 min after study proved too challenging. To remedy these potential issues, in Experiment 2 we used RIs that were shorter, yet proportionally more disparate, than those used in Experiment 1.

Experiment 2

Method

The design and variables were the same as in Experiment 1, with the exception of new RIs: 30 s and 3 min. On average, the short RI lasted 46 s, and the long RI lasted 4 min 43 s. Sixty-eight students enrolled in an introductory psychology course

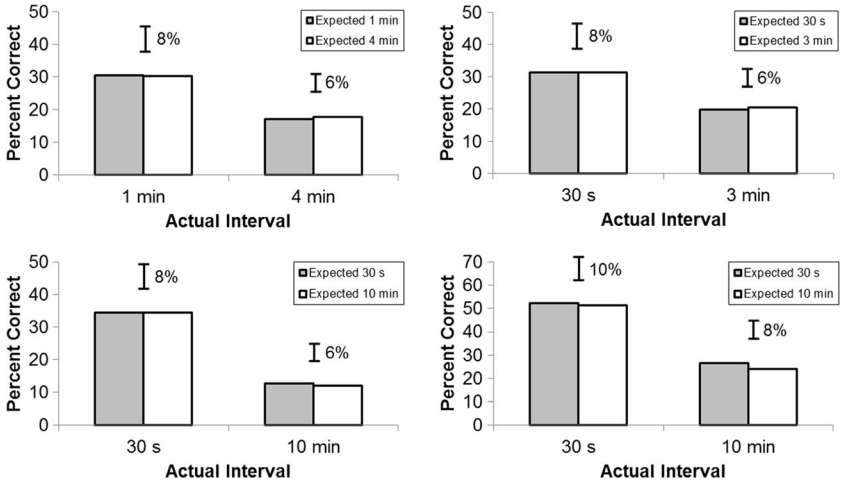

Fig. 1 Mean cued-recall performance as a function of the actual and expected retention intervals in Experiment 1 (upper left), Experiment 2 (upper right), Experiment 3 (lower left), and Experiment 4 (lower right). Error bars and values show the widths of the 95 % confidence intervals for the difference in performance between the expected RIs at each actual RI

at the University of Illinois at Urbana-Champaign participated for partial course credit.

Results

The results are shown in Fig. 1. Subjects recalled a higher percentage of the target words when tested after the 30-s RI ($M = 32$ %, $SD = 19$ %) than after the 3-min RI ($M = 21$ %, $SD = 14$ %), $B_{01} = 2.73 \times 10^{-8}$. At the 3-min RI, subjects recalled 21 % ($SD = 14$ %) of items for which they expected a 3-min RI, and they recalled 20 % ($SD = 17$ %) of items for which they expected a 30-s RI, $B_{01} = 6.47$. At the 30-s RI, subjects recalled 32 % ($SD = 22$ %) of items for which they expected a 3-min RI, and they recalled 32 % of items ($SD = 19$ %) for which they expected a 30-s RI, $B_{01} = 7.51$. As in Experiment 1, the subjects did not appear to differentiate between the two expected RIs.

Experiment 3

In Experiment 3, we used the same short RI as in Experiment 2 (30 s), but the long RI was now 10 min. The long interval was now 20 times the length of the short interval. The great disparity in interval length was created to encourage subjects to encode items differently, since recalling target words 10 min from study would hopefully be perceived as more difficult than recalling target words 30 s after study.

Method

Subjects Sixty students enrolled in an introductory psychology course at the University of Illinois at Urbana-Champaign participated for partial course credit.

Design The design and variables were the same as in Experiment 2, with the exception of the different RIs. On average, the short RI lasted 35 s, and the long RI lasted 11 min and 53 s.

Procedure The procedure for Experiment 3 was identical to that of Experiment 2, with one exception. After the study and test trials were completed, subjects completed a questionnaire on their study habits. The questionnaire will be described in greater detail below.

Results

The findings from Experiment 3 are presented in Fig. 1. For actual RI, we obtained a $B_{01} = 3.61 \times 10^{-17}$, indicating very strong evidence in favor of the alternative. As expected, subjects recalled a higher percentage of the target words when tested after the 30-s RI ($M = 35$ %, $SD = 17$ %) than after the 10-min RI ($M = 12$ %, $SD = 10$ %). At the 10-min RI, subjects recalled 12 % ($SD = 11$ %) of items for which they expected a 10-min RI, and they recalled 13 % ($SD = 13$ %) of items for which they expected a 30-s RI, $B_{01} = 6.29$. At the 30-s RI, subjects recalled 35% ($SD = 22$ %) of items for which they expected a 10-min RI, and they recalled 35 % of items ($SD = 16$ %) for which they expected a 30-s RI, $B_{01} = 7.08$. Both B_{01} values in favor of the null were in the same range as in the prior two experiments.

Questionnaire Subjects responded to a questionnaire (adapted from Finley & Benjamin, 2012) regarding their study strategies. This questionnaire asked the subjects to what extent they had used 11 strategies. Subjects responded on a scale from 1 (*not at all*) to 7 (*very frequently*) how often they had used each strategy. Subjects were shown the name of a strategy along with a short description (shown in Table 1). After

Table 1 Average usage ratings of encoding strategies, from the questionnaires in Experiments 3, 4, and 5

Strategy	Description	Average Rating, Exp. 3 (*SD*)	Average Rating, Exp. 4 (*SD*)	Average Rating, Exp. 5 (*SD*)
Cue–target association	*Made associations between the left-hand and right-hand words in a pair*	5.28 (1.64)	4.69 (2.23)	4.60 (1.47)
Interitem association	*Made associations between multiple pairs across the list*	2.73 (1.62)	2.44 (1.61)	3.30 (1.59)
Interitem narrative	*Used groups or pairs of words in a sentence, phrase, or story*	2.48 (1.84)	2.84 (2.08)	2.53 (1.80)
Intra-item narrative	*Used a single pair or word in a sentence, phrase, or story*	3.43 (2.10)	3.85 (2.29)	2.59 (1.62)
Mental imagery	*Used mental imagery (formed a picture in your head)*	3.65 (1.67)	3.84 (2.02)	3.04 (1.82)
Observation	*Just read or looked at the words*	5.00 (1.64)	4.59 (1.82)	5.91 (1.24)
Personal significance	*Related words to something personally significant*	4.28 (1.90)	3.93 (1.77)	3.22 (1.66)
Rote rehearsal	*Repeated individual words or pairs over and over*	5.70 (1.55)	5.31 (1.62)	4.84 (1.81)
Target focus	*Focused more on the right-hand words*	5.10 (1.65)	4.56 (1.81)	4.16 (1.73)
Target–target association	*Made associations between the right-hand words across multiple pairs*	2.33 (1.72)	2.21 (1.43)	3.16 (1.70)
Verbalization	*Spoke words out loud or under your breath*	5.03 (2.03)	4.77 (2.04)	3.67 (2.25)
Massing	*Studied the same set of terms over and over*	–	–	4.76 (1.80)
Spacing	*Avoided studying the same set of terms over and over*	–	–	2.44 (1.79)
Testing	*Guessed what term was being defined before revealing it*	–	–	4.09 (1.95)

providing their response, subjects were asked if they had used a strategy more for word pairs expected in 10 min, for word pairs expected in 30 s, if they had used the strategy equally between the expected RIs, or if they were not sure. Subjects were also allowed to report any strategies that they utilized that were not present on the questionnaire. After reporting their strategies, subjects were asked two final questions. First, they were asked if they had tried harder on word pairs on the basis of expected RI. This was to see whether changes in effort, if not strategy, were occurring in response to the time cues. Second, they were asked whether they had noticed that the time cues were not always accurate. This was a simple manipulation check to see whether subjects were perhaps not showing effects of expected RI because they had noticed that the cues were not a reliable source of information.

Questionnaire results The key interest in administering the questionnaire was to see whether the subjects used strategies more for one expected RI than for another. The results predictably indicated that changes in strategy were rare. For instance, the encoding strategy used most often was rote rehearsal. Subjects overwhelmingly reported (47 %) using rote rehearsal equally for word pairs with a 30-s or a 10-min expected RI. The second most popular response was "Not Sure" (25 %) when they had used rote rehearsal, followed by using it for word pairs cued "30 s" (22 %), and then for word pairs cued "10 min" (7 %). Thus, the difference in reported usage among word pairs cued for 30 s versus 10 min was overshadowed by nearly three-quarters of participants reporting that they had used rote rehearsal the same amount or were uncertain for which word pairs they had used it. This pattern held up among nearly all of the strategies. As might be inferred from the null effect of expectation on test performance, subjects were not trying to utilize different strategies for word pairs if they were cued for 30 s or for 10 min. The complete results from the strategy portion of the questionnaire are reported in Tables 1 and 2.

Subjects were invited to report strategies not included on the questionnaire. Few novel strategies were reported, however, and subjects did not consistently make mention of whether they had used strategies for word pairs based on the time cue. Thus, the results from this portion of the questionnaire are not considered further.

When asked whether they had tried harder at studying word pairs on the basis of the provided time cues, 47 % of subjects said they had tried harder on word pairs cued for 30 s, 10 % said they had tried harder on word pairs cued for 10 min, 33 % said they had tried equally hard on all word pairs, and 10 % reported being uncertain whether they had tried harder on certain word pairs than on others. This pattern indicates that changes in effort may have been fairly common, with nearly half of all subjects expending more effort on word pairs cued for 30 s. However, follow-up analyses failed to reveal any differences in performance based on participants' reported effort expenditure.

The final question was a manipulation check. Nearly half (47 %) of all subjects reported noticing that the time cues were not always reliable, whereas 53 % of all subjects reported not noticing the manipulation. However, follow-up analyses

Table 2 Proportions of subjects using encoding strategies based on the time cues in Experiments 3, 4, and 5

Strategy	Words Cued for 10 min	Words Cued for 30 s	Same Amount	Not Sure
Experiment 3				
Rote rehearsal	.07	.22	.47	.25
Cue–target association	.13	.10	.45	.32
Target focus	.05	.05	.55	.35
Verbalization	.10	.03	.48	.38
Observation	.17	.10	.48	.25
Personal significance	.15	.12	.58	.15
Mental imagery	.08	.10	.68	.13
Intra-item narrative	.03	.25	.67	.05
Interitem association	.05	.23	.60	.12
Interitem narrative	.10	.07	.32	.52
Target–target association	.10	.13	.65	.12
Experiment 4				
Rote rehearsal	.07	.22	.47	.25
Verbalization	.10	.03	.48	.38
Cue–target association	.13	.10	.45	.32
Observation	.17	.10	.48	.25
Target focus	.05	.05	.55	.35
Personal significance	.15	.12	.58	.15
Intra-item narrative	.03	.25	.67	.05
Mental imagery	.08	.10	.68	.13
Interitem narrative	.10	.07	.32	.52
Interitem association	.05	.23	.60	.12
Target–target association	.10	.13	.65	.12
Strategy	Words Cued for 24 h	Words Cued for 10 min	Same Amount	Not Sure
Experiment 5				
Observation	.04	.07	.84	.04
Rote rehearsal	.00	.16	.78	.07
Massing	.07	.24	.56	.13
Cue–target association	.09	.24	.53	.13
Target focus	.09	.04	.69	.18
Testing	.04	.02	.80	.13
Verbalization	.04	.11	.62	.22
Interitem association	.02	.07	.47	.42
Personal significance	.11	.07	.47	.36
Target–target association	.09	.02	.60	.27
Mental imagery	.11	.09	.56	.24
Interitem narrative	.13	.13	.36	.36
Intra-item narrative	.11	.09	.44	.36
Spacing	.16	.07	.49	.29

Strategies are listed from the most to the least highly rated

revealed no differences in performance between noticers and non-noticers.

Discussion

Once again, subjects did not appear to differentiate between the two expected RIs. As with the previous two experiments, only the actual RI affected subjects' cued-recall performance; performance was not different on the basis of the expected RI. The responses on a study strategy questionnaire confirmed the behavioral evidence that subjects did not change their encoding strategies on the basis of the expected RI.

Experiments 1 and 2 both restricted the study time to a fixed, 6-s-per-item pace. Although we hoped that a fixed study

time would encourage subjects to adopt qualitative shifts in their encoding strategies, the evidence strongly points to these shifts being largely absent. However, we may have created untenable encoding conditions, wherein subjects wanted to update their strategies for various items but were not given a sufficient amount of time to implement their strategic approach. Thus, in Experiment 4 we allowed subjects to self-pace their study. Self-paced study would give subjects a chance to qualitatively update their encoding strategies at their own pace; it would also allow us to observe whether subjects were inclined to make quantitative changes (i.e., to study certain items for longer) should they be reluctant to engage more sophisticated strategies.

Experiment 4

Experiment 4 was identical to Experiment 3, with the exception that study time was now self-paced rather than fixed. We hoped that self-paced study time would provide subjects with ample freedom to implement strategies that the previous procedure had prevented, and it also allowed us to observe whether subjects were deciding to study some items for longer.

Method

Subjects Sixty-one students enrolled in an introductory psychology course at the University of Illinois at Urbana-Champaign participated for partial course credit.

Design The design and variables were the same as in Experiment 3. On average, the short RI lasted 44 s, and the long RI lasted 10 min and 23 s.

Procedure The procedure for Experiment 4 was identical to that of Experiment 3, with two exceptions. First, subjects were instructed that they could study each word pair for as long as they wanted, and that they could press the space bar on their keyboard to proceed through the items. If a subject was studying a word pair when it was time for another word pair to be tested, the test was delayed until their study of the current word pair ceased. Second, to accommodate potential delays brought on by the self-paced study, we tested subjects on 50 rather than 100 items.

Results

The findings from Experiment 4 are presented in Fig. 1. We calculated four Bayes factors: one for the effect of actual RI, two for the effects of expected RI at each actual RI, and one for the effect of expected RI on study time. For actual RI, we obtained $B_{01} = 9.14 \times 10^{-17}$, indicating very strong evidence in favor of the alternative. As expected, subjects recalled a higher percentage of the target words when tested after the 30-s RI (M = 52 %, SD = 23 %) than after the 10-min RI (M = 25 %, SD = 22 %). At the 10-min RI, subjects recalled 24 % (SD = 23 %) of items for which they expected a 10-min RI, and they recalled 27 % (SD = 25 %) of items for which they expected a 30-s RI, B_{01} = 3.99. At the 30-s RI, subjects recalled 51 % (SD = 25 %) of items for which they expected a 10-min RI, and they recalled 52 % of items (SD = 25 %) for which they expected a 30-s RI, B_{01} = 6.76. Both B_{01} values in favor of the null were between 3 and 10; again, we found "some evidence" in favor of expectation having no effect on performance.

Study time data are presented in Fig. 2. Subjects studied words for 11.17 s (SD = 7.22) if they expected a 10-min RI, and for 10.42 s (SD = 6.85) if they expected a 30-s RI, B_{01} = 2.29. We found weak evidence in favor of the null that subjects were not varying their study time on the basis of the expected RI.

Questionnaire results As in Experiment 3, we once again administered a questionnaire to see whether subjects had used strategies more for one expected RI than for another. The results were almost entirely in accord with what we had seen in Experiment 3. The complete results from the strategy portion of the questionnaire are reported in Tables 1 and 2.

Subjects were also invited to report strategies that were not included on the questionnaire. Unlike in Experiment 3, a small number (n = 6) of subjects reported differential means of encoding based on the time cues. Most of these responses indicated a discrepancy reduction approach (i.e., "For thirty-second pairs I tended to just repeat [them] over and over in my head. For ten-minute pairs I tried to put [them] into a memorable sentence or story.") Still, most subjects reported not doing anything differently on the basis of the time cues.

When asked whether they had tried harder on word pairs on the basis of the provided time cues, 28 % of subjects said they had tried harder on the word pairs cued for 30 s, 25 % said they had tried harder on the word pairs cued for 10 min, 41 %

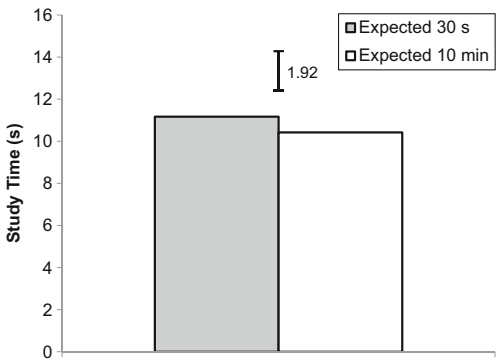

Fig. 2 Study time as a function of the expected retention interval in Experiment 4. The error bar and value show the 95 % confidence interval for the difference in study times

said they had tried equally hard on all word pairs, and 7 % reported being uncertain whether they had tried harder on certain word pairs than on others. As opposed to Experiment 3, in which half of the subjects had tried harder on words cued for 30 s, the majority response for Experiment 4 was that subjects did not differentially allocate their encoding efforts. As had been the case for Experiment 3, follow-up analyses failed to reveal any differences in performance based on participants' reported effort expenditures.

Finally, subjects were also asked whether they had noticed the manipulation. Approximately one-third (34 %) of all subjects reported noticing that the time cues were not always reliable, whereas 66 % of all subjects reported not noticing the manipulation. Once again, follow-up analyses found no differences in performance between noticers and non-noticers.

Discussion

The evidence from Experiments 1–4 suggested that learners do very little—quantitatively or qualitatively—in response to the expected RI. Collapsing across experiments, when subjects were tested at the long RI they recalled 19 % (*SD* = 17 %) of items for which they expected a long RI, and they recalled 19 % (*SD* = 19 %) of items for which they expected a short RI, B_{01} = 14.48. When tested at the short RI, subjects recalled 37 % (*SD* = 24 %) of items for which they expected a long RI, and they recalled 37 % (*SD* = 22 %) of items for which they expected a short RI, B_{01} = 13.90.

However, some features of our methodology proved worrisome when assessing our findings. First, our continuous study–test paradigm may have been too challenging for our subjects. Having the RIs interleaved meant that subjects had to continuously update their encoding strategies, and the challenge of doing so may have impeded subjects' attempts to differentially encode. Second, the RIs selected to this point had all been relatively close together; maybe a difference in RIs on the order of minutes does not convince learners that the RIs are meaningfully different. Third, our stimuli—word–synonym pairs—were not deeply complex or interesting, and might have engendered a general lack of motivation during encoding. Any single one these concerns might have influenced our findings; the goal of Experiment 5 was to see whether learners might update their metacognitive control in a modified paradigm that addressed the potential issues listed above.

Experiment 5

In Experiment 5, we used a vastly different paradigm from that used in Experiments 1–4. Our objective was to increase awareness of the RIs and to facilitate differential encoding of items, and to that end we made several changes. First, we used longer and more distinguishable RIs: The short interval was 10 min, and the long interval was 24 h. Second, we abandoned the continuous study–test trials, and instead had a designated study session followed by two testing sessions (i.e., one test after each RI). Third, we changed the stimuli to flashcards containing definition–term pairs from an introductory psychology textbook. The change in the stimuli was meant to reflect materials that a student might actually study, as well as to provide richer encoding opportunities relative to the word pairs we had used in the previous experiments. Fourth, to ensure that our subjects were engaged and motivated to do well, the experiment was completed by students from an introductory psychology course roughly three weeks before the students would be tested on some of the items appearing in the experiment.

Method

Subjects Forty-eight students enrolled in an introductory psychology course at the University of Illinois at Urbana-Champaign participated for partial course credit. Three of these participants were excluded from the subsequent data analyses: two who reported writing down the terms during the study phase, and a third who reported attempting the experiment multiple times due to technical difficulties.

Design As had been the case for the first four experiments, for Experiment 5 we used a 2 × 2 within-subjects design. The independent variables were expected RI (10 min or 24 h) and actual RI (10 min or 24 h). The dependent variable was performance on the tests following each RI.

Materials The stimuli were 50 definition–term pairs from an introductory psychology textbook. Half of the terms were randomly assigned to be expected for testing in 10 min, and the other half for 24 h. Of the 25 terms expected to be tested at each RI, 20 were randomly selected to be tested at the indicated interval, and five were to be tested at the contraindicated interval.

Procedure All subjects were recruited from an introductory psychology course. They completed the experiment online at their convenience over a one-week period. They were instructed that they would be reviewing online flashcards containing terms from their textbook that they could be tested on as part of their upcoming final exam. Furthermore, subjects were told that the terms were broken down into two sets, one that would be tested 10 min after they were done studying, and another that would be tested the next day. Subjects were told that each set contained 25 terms and that they would have 25 min to study both sets of terms.

Upon beginning the study portion of the experiment, subjects saw a display in the upper right corner of the screen that indicated how much time they had spent studying. (This

display was always visible throughout the study session.) Subjects were presented with two buttons, one for each set of terms to be tested at each RI. They could click the button corresponding to either set; upon clicking, they were shown the flashcards, one at a time, belonging to that group. For a given flashcard, subjects initially saw a definition with an empty box above it, and after they had clicked a button, the term being defined appeared in the formerly empty box. Study was entirely self-paced. Once a subject had reviewed all of the terms for a given set, he or she was once again given the option of choosing which set of terms to study. This routine went on for 25 min. This new setup allowed the subjects a larger degree of flexibility in their encoding—and consequently more opportunity to update strategies—than had the previous experiments. For instance, they could now place more importance on a certain set of terms by dedicating more rounds of study to that set (i.e., a quantitative shift). They could also decide to space or mass one of the sets of terms (i.e., a qualitative shift). Also, because the stimuli were presented as flashcards, the subjects could test themselves more easily than in the previous experiments—perhaps also encouraging differential application of self-testing based on expected RI (i.e., a qualitative shift).

After 25 min, subjects' study sessions immediately ended, even if they were in the middle of a set of terms. Subjects then completed a distractor task for 10 min before beginning the first test. The first test consisted of 20 items from the set of terms that was expected to be tested in 10 min, and five items from the set of terms expected to be tested in 24 h. Unlike in the continuous study–test paradigm used in our previous experiments, here we were concerned that the testing at contraindicated intervals might be more conspicuous in the current paradigm, in which study and test were clearly separated. To avoid arousing suspicions on the first test, we explicitly told subjects that they would see some items from the 24-h set because we were interested in assessing differences in memory for those items between the first and second tests. During the test trials, the subjects were provided with a definition and asked to provide the corresponding term.

On the second day of the experiment, subjects completed the second test. They were once again tested on 20 items from the expected set of terms and five items from the unexpected set. After completing the second test, subjects completed a study strategy questionnaire, as they had in Experiments 3 and 4. Because the paradigm in Experiment 5 lent itself to the implementation of strategies not as easily employed in the previous experiments, we added three new items to the questionnaire to assess spacing, massing, and testing.

Results

Lenient scoring Some of the terms that subjects were asked to produce at test contained characters that we anticipated might not be included in subjects' responses, yet would have little bearing on whether a response was accurate. For instance, some items included hyphens ("door-in-the-face technique"), and others were pluralized ("collective delusions"). Failing to report these features may have affected the accuracy scores of some subjects in unimportant ways. Thus, before analyzing the data in Experiment 5, we wrote a lenient scoring algorithm that determined the proportion of characters matching between a target term and a subject's response, and, on the basis of a criterion proportion, decided whether a response was correct. We ultimately decided on a 90 %-match criterion; a 90 % match accounted for many of the cases highlighted above, without erroneously determining that clearly incorrect responses were correct. Of the 1,614 incorrect test responses gathered, only 113 were counted as correct after being leniently scored.

Test performance The findings from Experiment 5 are presented in Fig. 3. We calculated eight Bayes factors: one for the effect of actual RI, two for the effects of expected RI at each actual RI, one to test for an interaction between the expected and actual RIs, three for the effects of expected RIs on study time, and one for the number of practice rounds initiated for each set of flashcards. For actual RI, we obtained a $B_{01} = 3.36 \times 10^{-4}$, indicating very strong evidence against the null. As expected, subjects recalled a higher percentage of the terms tested after the 10-min RI (M = 38 %, SD = 27 %) than after the 24-h RI (M = 25 %, SD = 25 %). At the 10-min RI, subjects recalled 45 % (SD = 24 %) of items for which they expected a 10-min RI, and they recalled 32 % (SD = 29 %) of items for which they expected a 24-h RI, $B_{01} = 0.01$. At the 24-h RI, subjects recalled 32 % (SD = 26 %) of items for which they expected a 10-min RI, and they recalled 21 % of items (SD = 24 %) for which they expected a 24-h RI, $B_{01} = 0.01$. In contrast to the previous experiments, we found "very strong evidence" against the null hypothesis at both the 10-

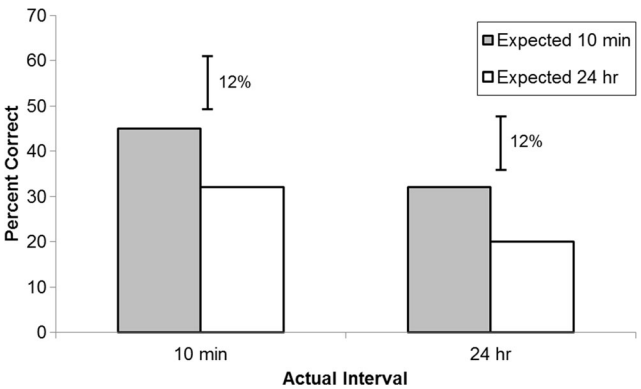

Fig. 3 Mean cued-recall performance as a function of the actual retention interval (actual RI: 10 min vs. 24 h) and the expected RI (10 min vs. 24 h) in Experiment 5. The error bars and values show the width of 95 % confidence interval for the difference in performance between the expected RIs at each actual RI

min RI and the 24-h RI. In both cases, subjects performed better on the terms expected to be tested in 10 min. To evaluate whether the expected RI was interacting with the actual RI, we compared the differences in performance, on the basis of expected RI, at each interval. At the 10-min RI, subjects got 13 % (*SD* = 21 %) more items correct if they expected a 10-min RI; at the 24-h RI, they got 12 % (*SD* = 20 %) more items correct if they expected a 10-min RI, B_{01} = 6.04. We obtained "some evidence" in favor of the null, suggesting that subjects were not effectively tailoring their study for the two RIs.

Study time allocation The study time data are presented in Fig. 4. We split these data into the time spent with just the definition (which we called *definition study*) and time spent with both the definition and the term (which we called *pair study*). Partitioning the study time on the basis of the terms' presence or absence allowed us to get a potential measure of self-testing (definition study) as well as of time spent making associations between the definitions and their terms (pair study). Subjects' definition study lasted for 7.13 s (*SD* = 8.66) if they expected a 10-min RI, and for 6.26 s (*SD* = 8.96) if they expected a 24-h RI, B_{01} = 5.65. We found "some evidence" in favor of the null that subjects were not varying their definition study time on the basis of the expected RI. (Inasmuch as definition study time was considered a proxy for self-testing, the questionnaire results—discussed later—corroborated the behavioral data.) Furthermore, subjects' pair study lasted for 3.93 s (*SD* = 3.23) if they expected a 10-min RI, and for 3.01 s (*SD* = 2.88) if they expected a 24-h RI, B_{01} = 2.54. This constitutes weak evidence in favor of the null for pair study time. Combining definition study with pair study, the subjects studied a given flashcard for 11.06 s (*SD* = 10.37) if they expected a 10-min RI, and for 9.28 s (*SD* = 9.62) if they expected a 24-h RI, B_{01} = 4.54. We found "some evidence" in favor of the null for overall study time.

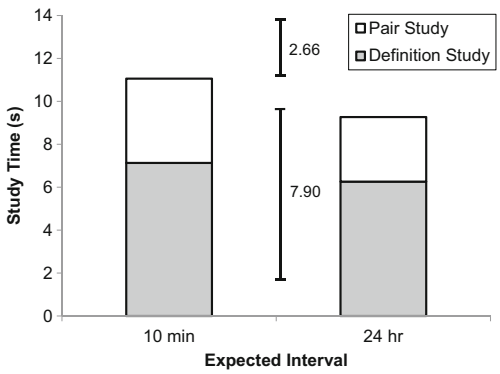

Fig. 4 Study time as a function of the expected retention interval in Experiment 5. The error bars and values show the 95 % confidence intervals for the differences in definition study (lower bar) and pair study (upper bar)

Selection of study items Finally, because subjects were also in control of how many rounds of study they could dedicate to each set of flashcards, we looked at the amount of rounds that they *initiated* for each set (keeping in mind that a round may have been prematurely terminated by the 25-min time limit during the study phase). These data are presented in Fig. 5. Subjects initiated 3.31 rounds of practice (*SD* = 1.89) for the set of terms to be tested after the 10-min RI, and 2.16 rounds of practice (*SD* = 1.62) for the set of terms to be tested after the 24-h RI, B_{01} = 0.06. We found "strong evidence" against the null, indicating that subjects were initiating more practice rounds for the 10-min than for the 24-h set.

Questionnaire results As in Experiments 3 and 4, we once again administered a questionnaire to see whether subjects had used strategies more for one expected RI than for the other. The items were slightly reworded to be more applicable to the new paradigm (i.e., using "definition" and "term" instead of "cue" and "target"). Once again, the subjects overwhelmingly reported using the strategies equally for the two RIs. Of the three new items, massing was the third highest-rated strategy, followed by testing (6th), and then spacing (14th—the lowest-rated strategy). The complete results from the strategy portion of the questionnaire are reported in Tables 1 and 2.

Subjects were also invited to report strategies not included on the questionnaire. Few subjects reported anything novel; however, on the basis of the test performance data, one subject may have articulated the majority mindset by writing, "I tried to switch out from learning one set of cards to another, but in the end I just stuck with the cards I was going to be tested on the earliest because they seemed more relevant at the time."

When asked whether they tried harder on word pairs on the basis of the provided time cues, 44 % of the subjects said they had tried harder when they anticipated a 10-min RI, 16 % said they had tried harder on the 24-h RI, 36 % said they had tried equally hard on all definition–term pairs, and 4 % reported being uncertain whether they had tried harder on one set or the other. As the performance data suggested, the most frequent

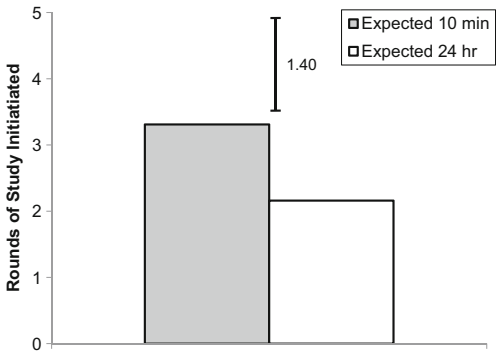

Fig. 5 Rounds of study initiated as a function of the expected retention interval in Experiment 5. The error bar and value show the 95 % confidence interval for the difference in rounds initiated

response was to allocate more effort to the 10-min set of flashcards.

Finally, subjects were asked whether they had expected to be tested on definitions from the 10-min set at the 24-h RI. Approximately one-half (51 %) of all subjects said that they had anticipated a test on some of the 10-min flashcards at the 24-h RI, whereas 49 % of the subjects reported no such expectation. Among the anticipating subjects, at the 24-h RI, they recalled a higher percentage of the 10-min items (M = 31 %, SD = 26 %) than of the 24-h items (M = 19 %, SD = 23 %), B_{01} = 0.06. For the subjects who anticipated the 10-min items, we found strong evidence against the null. Among nonanticipating subjects, they also recalled a higher percentage of the 10-min items (M = 33 %, SD = 27 %) than of the 24-h items (M = 21 %, SD = 26 %), B_{01} = 0.49. For the nonanticipating subjects, we found weak evidence against the null. Although the weak evidence for the nonanticipators might suggest that these subjects were subpar in differentially encoding the items, we still had behavioral measures of differential encoding—specifically, the number of practice rounds initiated—that strongly suggested that subjects were responding to the expected RIs.

Discussion

In contrast to the previous four experiments, the subjects in Experiment 5 showed an effect of RI expectation on performance. On the basis of performance, behavioral study measures, and, to a lesser extent, questionnaire data, subjects placed more emphasis on encoding terms that were to be tested in 10 min. As had been the case in Experiment 4, subjects did not differentially allocate study time on the basis of expectation. And as had been the case for the previous questionnaires, subjects rarely reported differentially engaging their encoding strategies. It appears that subjects placed more emphasis on the 10-min terms by electing to study them more often, and not by studying them for longer or by strategically altering their encoding.

General discussion

To recap, our first four experiments suggested that learners do not account for the anticipated RI when encoding to-be-learned items. Data collected via behavioral measures and study strategy questionnaires showed that subjects were taking virtually the same tacks for words expected to be tested at different RIs. In Experiment 5, we altered our methodology to potentially heighten subjects' RI sensitivity and to give them more freedom in their study choices: We used considerably different RIs of 10 min and 24 h, and subjects were allowed to self-pace their study trials as well as to self-select which materials they would study. This new setup did in fact make subjects more responsive to the different RIs; specifically, they appeared to prioritize items to be tested in 10 min. Nonetheless, there was no evidence that they adopted qualitatively different strategies for learning the materials that they expected to be tested on sooner.

Experiment 5 supported our second prediction, that learners choose to allocate more resources to the shorter—and therefore easier—RI rather than to the longer RI (cf. Metcalfe & Kornell, 2003). Emphasizing one set of items over another reflects a quantitative shift in encoding—learners are administering different *amounts* rather than different *kinds* of processing. Quantitative shifts are common in the test-expectancy literature (particularly when learners expect a recall vs. a recognition test), but learners usually shift toward the more difficult test, whereas we found a shift toward the easier test.

Our experimental design could explain this discrepancy in shift: We used a within-subjects design, whereas most of the test-expectancy literature has used a between-subjects design. Had we used a between-subjects manipulation, the group expecting a 24-h RI might have consistently outperformed the group expecting a 10-min RI simply because they were preparing for a more difficult test (though between-subject manipulations of RI rarely result in accurate predictions of forgetting; Koriat, Bjork, Sheffer, & Bar, 2004). It could also be that the perceived magnitudes of difficulty are dissimilar when comparing test format to test timing: Learners might feel that they can adequately prepare for a recall test by studying harder, but encoding information to be retrieved in 24 h might feel too daunting and not worth diverting resources away from a shorter RI. Finally, the time constraints in our experiments might have made subjects' study habits more conservative. The average overall performance was rather low (33 %), suggesting that subjects had insufficient study time to master the materials, and perhaps adopted a strategy that prioritized easier material (Son & Metcalfe, 2000).

Given our concerns about potential time constraints, future work should investigate whether our findings would persist in less-constraining study conditions (i.e., with fewer materials or longer total study time), since most students presumably allow themselves more time for greater mastery than our subjects were given. Future work should also use longer expected RIs. The immediacy of the 10-min test in Experiment 5 may have been enough to persuade learners to focus their efforts on those items, independent of any considerations of forgetting associated with the RIs. Learners would ideally be able to differentially prepare for two RIs when forestalling future forgetting is of primary concern and neither test introduces immediate performance demands (e.g., 24 vs. 48 h).

In the present research, we found that, under the right conditions, learners will differentially encode items on the basis of the expected RI, but they have no particular strategies beyond allocating more effort to one RI than to another. Our findings

agree with much of the work in the test-expectancy literature that has shown that changes in amount of effort prevail over changes in kind: Subjects chose to study one set of items more than another, and reported few instances of differential processing based on expected RI. Our findings also support work on effort allocation that has suggested that learners tend to focus on easier items at encoding, with subjects preferring to study items expected after shorter RIs. Given that the role of expected RI in metacognitive control is relatively unexplored, future work will hopefully build on the results presented here and provide a fuller understanding of learners' metacognitive updating at encoding.

References

Balota, D. A., & Neely, J. H. (1980). Test-expectancy and word-frequency effects in recall and recognition. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 576–587. doi:10.1037/0278-7393.6.5.576

Belmont, J. M., & Butterfield, E. C. (1971). Learning strategies as determinants of memory deficiencies. *Cognitive Psychology, 2,* 411–420.

Benjamin, A. S. (2008). Memory is more than just remembering: Strategic control of encoding, accessing memory, and making decisions. In A. S. Benjamin & B. H. Ross (Eds.), *The psychology of learning and motivation: Skill and strategy in memory use* (Vol. 48, pp. 175–223). London, UK: Academic Press.

Benjamin, A. S., & Bird, R. D. (2006). Metacognitive control of the spacing of study repetitions. *Journal of Memory and Language, 55,* 126–137. doi:10.1016/j.jml.2006.02.003

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology, 61,* 228–247. doi:10.1016/j.cogpsych.2010.05.004

Bisanz, G. L., Vesonder, G. T., & Voss, J. F. (1978). Knowledge of one's own responding and the relation of such knowledge to learning. A developmental study. *Journal of Experimental Child Psychology, 25,* 116–128.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132,* 354–380. doi:10.1037/0033-2909.132.3.354

Cull, W. L., & Zechmeister, E. B. (1994). The learning ability paradox in adult metamemory research: Where are the metamemory differences between good and poor learners? *Memory & Cognition, 22,* 249–257.

Dufresne, A., & Kobasigawa, A. (1989). Children's spontaneous allocation of study time: Differential and sufficient aspects. *Journal of Experimental Child Psychology, 47,* 274–296. doi:10.1016/0022-0965(89)90033-7

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 educational theory and practice* (pp. 249–276). Mahwah, NJ: Erlbaum.

Fiechter, J. L., Benjamin, A. S., & Unsworth, N. (2016). The metacognitive foundations of effective remembering. In J. Dunlosky & S. K. Tauber (Eds.), *Oxford handbook of metamemory* (pp. 301–324). New York, NY: Oxford University Press.

Finley, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: Evidence from the test-expectancy paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 632–652.

Finley, J. R., Tullis, J. G., & Benjamin, A. S. (2010). Metacognitive control of learning and remembering. In M. S. Khine & I. M. Saleh (Eds.), *New science of learning: Cognition, computers, and collaboration in education* (pp. 108–132). New York, NY: Springer.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116,* 439–453. doi:10.1037/a0015251

Hall, J. W., Grossman, L. R., & Elwood, K. D. (1976). Differences in encoding for free recall vs. recognition. *Memory & Cognition, 4,* 507–513.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press, Clarendon Press.

Kobasigawa, A., & Metcalf-Haggert, A. (1993). Spontaneous allocation of study time by first- and third-grade children in a simple memory task. *Journal of Genetic Psychology, 154,* 223–235.

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,* 643–646. doi:10.1037/0096-3445.133.4.643

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135,* 36–69. doi:10.1037/0096-3445.135.1.36

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

Le Ny, J., Denhiere, G., & Le Taillanter, D. (1972). Regulation of study-time and interstimulus similarity in self-paced learning conditions. *Acta Psychologica, 36,* 280–289.

Leonard, J. M., & Whitten, W. B., II. (1983). Information stored when expecting recall or recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 440–455. doi:10.1037/0278-7393.9.3.440

Masur, E. F., McIntyre, C. W., & Flavell, J. H. (1973). Developmental changes in apportionment of study time among items in a multitrial free recall task. *Journal of Experimental Child Psychology, 15,* 237–246.

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective? *Journal of Experimental Psychology: General, 122,* 47–60. doi:10.1037/0096-3445.122.1.47

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory & Cognition, 18,* 196–204.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15,* 174–179. doi:10.3758/PBR.15.1.174

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General, 132,* 530–542. doi:10.1037/0096-3445.132.4.530

Neely, J. H., & Balota, D. A. (1981). Test-expectancy and semantic-organization effects in recall and recognition. *Memory & Cognition, 9,* 283–300.

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. doi:10.1111/j.1467-9280.1994.tb00502.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. doi:10.1037/0278-7393.14.4.676

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 26, pp. 125–141). San Diego, CA: Academic Press.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. doi:10.3758/PBR.16.2.225

Schmidt, S. R. (1988). Test-expectancy and individual-item versus relational processing. *American Journal of Psychology, 101,* 59–71.

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. doi:10.1037/0278-7393.26.1.204

Thiede, K. W. (1996). The relative importance of anticipated test format and anticipated test difficulty on performance. *Quarterly Journal of Experimental Psychology, 49,* 901–918.

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 25,* 1024–1037. doi:10.1037/0278-7393.25.4.1024

Toppino, T. C., Cohen, M. S., Davis, M. L., & Moors, A. C. (2009). Metacognitive control over the distribution of practice: When is spacing preferred? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1352–1358.

Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language, 64,* 109–118.

von Wright, J. (1977). On the development of encoding in anticipation of various tests of retention. *Scandinavian Journal of Psychology, 18,* 116–120.

von Wright, J., & Meretoja, M. (1975). Encoding in anticipation of various tests of retention. *Scandinavian Journal of Psychology, 16,* 108–112.