FAILURE TO UPDATE METACOGNITIVE CONTROL IN RESPONSE TO EXPECTED RETENTION INTERVALS

BY

JOSHUA FIECHTER

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Arts in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Adviser:

Professor Aaron Benjamin

ABSTRACT

To effectively allocate encoding resources, learners should take into account how long they need to retain information before it will be needed. Four experiments investigated whether expected retention intervals affect subjects' encoding strategies. 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. On 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 expected retention interval was found for 1 minute versus 4 minutes (Experiment 1), 30 seconds versus 3 minutes (Experiment 2), for 30 seconds versus 10 minutes (Experiments 3 and 4), and even when subjects were given complete control over the pacing of study items (Experiment 4). Subjects completed a study strategy questionnaire after Experiments 3 and 4 that indicated that these null effects were not due to unsuccessful strategy implementation; subjects appear to have adopted nearly identical learning strategies for the two intervals. This set of results accords with much of the test-expectancy literature, in which subjects rarely make *qualitative* adjustments to their encoding strategies based on expected test features. A Bayesian analysis provided strong evidence to suggest that learners fail to compensate for anticipated forgetting with differential encoding.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

In an educational setting, students often inquire about various properties of an upcoming test. Many inquiries regard 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 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 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 other times they are far removed. That is, the retention interval (RI) between a student's study session and upcoming test may be of long or short duration. Experiments manipulating RI are plentiful, though little is known about students' responses to expected RIs. The present research sought to answer the question of whether students make changes to their study habits if they expect a long or 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 encoding strategies (Fiechter, Benjamin, & Unsworth, 2015). 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). By this view, an item is studied until it meets a goal level of learning, which they call the norm of study (Le Ny, Denhiere, & Taillanter, 1972). As an item is being studied, strategies are monitored and updated to ensure study is effective, suggesting that learners will apply the most effective strategies to the most difficult items (e.g., Benjamin & Bird, 2006; Toppino, Cohen, Davis, & Moors, 2009).

The discrepancy reduction hypothesis predicts the assignment of more effective strategies for more difficult items, but little research has investigated qualitative encoding differences between easy and difficult items. Instead, most research has investigated the amount of study time allocated to material. Study time is easy to measure and likely correlates with expended effort on a given item. Previous research on study time has shown that 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. (1972) presented subjects with paired three-digit numbers and letters (i.e. 446:Z). 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 towards study performed better than a control group that did not have control over study time. Subjects that failed to adopt a discrepancy reduction approach performed no better than 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 Ma'ayan, & Nussinson, 2006; Mazzoni & Cornoldi, 1993; Mazzoni, Cornoldi, & Marchitelli, 1990; Nelson & Leonesio, 1988; Nelson, Dunlosky, Graf, & Narens, 1994). For example, Metcalfe and Finn (2008) found that learners choose to restudy items that they erroneously perceive as 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. 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 two study sessions). Critically, items that received higher 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 as most difficult (Cull & Zechmeister, 1994). That is, both good and poor learners conform to 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 & Dunlosky (1999) manipulated subjects' performance goals. They presented subjects with 30 word pairs and gave them a goal of recalling either 6 items (easy goal) or 24 items (difficult goal) on an upcoming cued-recall test. After studying all the pairs, subjects were to select pairs that they would like to further re-study. 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; Experiment 1) presented subjects with eight biographical essays, each six pages in length. Subjects were given only 30 minutes to look through the 48 pages of essays before taking a fill-in-the-blank test. With such limited time, subjects chose to study 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 either testing in a way that conforms to their expectation or is opposite of 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 different from what the subject has been led to expect. Performance is 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 expectancy and recognition test expectancy. Given students' interest in the features of an upcoming test, it seems that expected test format would be very influential in how they conduct their study. So if a subject expects a recall test, they should study in a way that benefits performance on a recall test. Likewise, if a subject expects a recognition test, they should study such that their performance will be maximal on a test of recognition. But this result is almost never found; rather, the most common finding in this literature is for an expectation of a recall test to lead 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 subjects 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 tailors his or her study habits to encode information in a manner than is optimal for an expected test format. Convincing evidence for qualitative changes in study require a disordinal interaction, such that subjects expecting a given type of test perform better on that test format relative to subjects expecting another test format.

Finley and Benjamin (2012) found evidence of such qualitative changes in study. They gave subjects four study-test phases, with all four tests being either cued-recall or free-recall of target words. Stimuli were related and unrelated word pairs, distributed equally across all lists. Subjects then had a fifth study-test phase where the test format was either what they had been induced to expect or 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, 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 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 word pairs 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 their 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 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." The present research sought to investigate the effects of 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. 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 three 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 their expectations. This outcome would reflect a discrepancy reduction policy in which subjects reserve 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 their expectations. Overall superior performance on easier items would indicate that the task constraints pushed subjects towards 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 compared to words tested at an unexpected 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 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 RI were at odds.

CHAPTER 2

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 x 2 within-subjects design. Independent variables were actual RI (1 or 4 minutes) and expected RI (1 or 4 minutes). The dependent variable was performance on cued-recall test trials. Performance was collapsed across individual items to get a percent correct measurement for each subject.

Materials. ***Stimuli.*** Stimuli were 140 paired associates consisting of words from the Graduate Record Exam and a synonym. Of the 140 word pairs, 100 were tested. The remaining 40 were used in filler study trials so that subjects were occupied during RIs near the end of the list. Study trials 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 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 number 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 word pairs cued for 4 minutes were really tested in 1 minute and word pairs cued for 1 minute were really tested in 4 minutes. The switched cues provided the key manipulation of testing subjects either at the expected RI or at 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 RIs when all 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. 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 milliseconds, and then remained on the screen while the word pair was presented to the subject for 6000 milliseconds.

The screen then went blank for 2000 milliseconds. Test trials were subject paced. The experiment was programmed using MATLAB programming software.

Results

Findings from Experiment 1 are presented in Figure 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 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 alternative hypotheses. Thus, Bayesian analyses allow us to prove a null effect (Gallistel, 2009) if a preponderance of evidence supports it. Critically, NHST does not allow evidence to accumulate in favor of the null; it instead assesses the level of evidence for the alternative hypothesis under the assumption that the null effect is true. Thus, 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 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 odds in favor of the null.

The selection of priors is not a straightforward task and is left to the discretion of the analyst. For our analyses, we followed recommendations by Rouder, Speckman, Sun, Morey, and Iverson (2009), who proposed the Jeffreys-Zellner-Siow (JZS) 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 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 actual RI and two for the effect of expected RI at each actual RI. For actual RI, we obtained a $B_{01} < .03$, 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 minute RI (M = 31%, SD = 21%) than after the 4 minute RI (M = 18%, SD = 18%). At the 4 minute RI, subjects recalled 18% (SD = 16%) of items when expecting a 4 minute RI and they recalled 18% (SD = 19%) of items when expecting a 1 minute RI, $B_{01} = 7.16$. At the 1 minute RI, subjects recalled 31% (SD = 23%) of items when expecting a 4 minute RI and they recalled 31% of items (SD = 20%), $B_{01} = 7.42$. 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 where 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 on the 4 minute RI may have been impervious to changes in strategy because recalling items 4 minutes after study proved too challenging. To remedy these potential issues, Experiment 2 attempted to make

the intervals proportionally farther apart such that they would be more distinguishable and one would clearly be easier/harder than the other.

CHAPTER 3

EXPERIMENT 2

Experiment 2 was identical to Experiment 1 except for the RIs, which were now 30 seconds and 3 minutes. The longer interval was now six times the length of the short, which hopefully would stand out to the subject as being a distinctly harder amount of time to retain information in memory.

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 design and variables were the same as Experiment 1, with the exception of the different RIs.

Procedure. The procedure for Experiment 2 was identical to Experiment 1.

Results

The results are shown in Figure 2. For actual RI, we obtained a $B_{01} < .03$, 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 second RI (M = 32%, SD = 20%) than after the 3 minute RI (M = 21%, SD = 17%). At the 3 minute RI, subjects recalled 21% (SD = 14%) of items when expecting a 3 minute RI and they recalled 20% (SD = 17%) of items when expecting a 30 second RI, $B_{01} = 6.47$. At the 30 second RI, subjects recalled 32% (SD = 22%) of items when expecting a 3 minute RI and they recalled 32% of items (SD = 19%) when expecting a 30 second RI, $B_{01} = 7.51$. Bayes factos indicate approximately the same level of support in favor of the null hypothesis as seen in Experiment 1.

Discussion. As was the case with Experiment 1, subjects did not appear to differentiate between the two expected RIs. However, the difference in performance was approximately the same across the two RIs as in Experiment 1, so it is possible that the RIs were still not distinct enough for subjects to decide one was more difficult than the other. Experiment 3 attempted to separate the intervals even more drastically than did Experiment 2.

CHAPTER 4

EXPERIMENT 3

Experiment 3 was identical to Experiments 1 and 2 except for the RIs, which were now 30 seconds and 10 minutes. The long interval was now twenty times the length of the short interval. The great disparity in interval length was to encourage subjects to determine that recalling target words 10 minutes from study would clearly be more difficult than recalling target words 30 seconds 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 Experiment 1, with the exception of the different RI.

Procedure. The procedure for Experiment 3 was identical to Experiments 1 and 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

Findings from Experiment 3 are presented in Figure 3. For actual RI, we obtained a $B_{01} <$ .03, 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 second RI (M = 35%, SD = 19%) than after the 10 minute RI (M = 12%, SD = 12%). At the 10 minute RI, subjects recalled 12% (SD = 11%) of items when expecting a 10 minute RI and they recalled 13% (SD = 13%) of items when expecting a 30 second RI, B_{01} = 6.29. At the 30 second RI, subjects recalled 35% (SD =

22%) of items when expecting a 10 minute RI and they recalled 35% of items (*SD* = 16%) when expecting a 30 second RI, B_{01} = 7.08. Both B_{01} values in favor of the null were in the same range as the prior two experiments.

Questionnaire. Subjects responded to a questionnaire (adapted from Finley and Benjamin, 2012) regarding their study strategies. This questionnaire asked subjects to what extent they used 11 strategies. Subjects responded on a scale from 1 (not at all) to 7 (very frequently) how often they used a strategy. Subjects were shown the name of a strategy along with a short description (shown in Table 1). After providing their response, subjects were asked if they used a strategy more for word pairs expected in 10 minutes, for word pairs expected in 30 seconds, if they used the strategy equally among 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 strategies, subjects were asked two final questions. First, they were asked if they had tried harder on word pairs based on expected RI. This was to see if changes in effort, if not strategy, were occurring in response to the time cues. Second, subjects were asked if they noticed that the time cues were not always accurate. This was a simple manipulation check in order to see if subjects were perhaps not showing effects of expected RI because they noticed that the cues were not a reliable source of information.

***Questionnaire results*.** The key interest in administering the questionnaire was to see if subjects used strategies more for one expected RI than for another. To that end, the results predictably indicated that changes in strategy were rare. For instance, the encoding strategy most often used was rote rehearsal. Subjects overwhelmingly reported (47%) using rote rehearsal equally for word pairs with a 30 second or 10 minute expected RI. The second most popular response was “Not Sure” (25%), followed by word pairs cued for 30 seconds (22%), and then

word pairs cued for 10 minutes (7%). Thus, the difference in reported usage among word pairs cued for 30 seconds versus 10 minutes was overshadowed by nearly three-quarters of participants reporting that they used rote rehearsal the same amount or were uncertain for which word pairs they 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 seconds or for 10 minutes. 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 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 if they tried harder on word pairs based on provided time cues, 47% of subjects said they tried harder on word pairs cued for 30 seconds, 10% said they tried harder on word pairs cued for 10 minutes, 33% said they tried equally hard on all word pairs, and 10% reported being uncertain if they tried harder on certain word pairs versus 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 seconds. This variable will be considered further following the next experiment.

The final question was a manipulation check. Nearly half (47%) of all subjects reported noticing that the time cues were not always reliable, while 53% of all subjects reported not noticing the manipulation. This variable will also be considered further following the next experiment.

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 based on the expected RI. Responses on a study strategy questionnaire confirmed the behavioral evidence for subjects not changing encoding strategies based on expected RI.

Experiments 1–3 all restricted study time to a fixed 5-seconds-per-item pace. Although we hoped that a fixed study time would encourage subjects to adopt shifts in their encoding strategies or effort, the evidence strongly points to these differences 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 update their encoding strategies at their own pace; it would also allow us to observe whether subjects were inclined to study certain items for longer should they be reluctant to engage more sophisticated strategies.

CHAPTER 5

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 time to implement any strategies that were previously stymied by experimenter-paced study time, 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 Experiment 3.

Procedure. The procedure for Experiment 4 was identical to Experiments 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 self-paced study, we tested subjects on 50 rather than 100 items.

Results

Findings from Experiment 4 are presented in Figure 4. We calculated four Bayes Factors: one for the effect of actual RI, two for the effect of expected RI at each actual RI, and one for effect of expected RI on study time. For actual RI, we obtained a $B_{01} < .03$, 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 second RI (M = 52%, SD = 25%) than after the 10 minute RI (M = 25%, SD = 24%). At the 10 minute RI, subjects recalled 24% (SD = 23%) of items when expecting a 10 minute RI and they recalled 27% (SD = 25%) of items when expecting a 30 second RI, B_{01} = 3.99. At the 30 second RI, subjects recalled 51% (SD = 25%) of items when expecting a 10 minute RI and they recalled 52% of items (SD = 25%) when expecting a 30 second 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.

Subjects studied words for 11.17 seconds (SD = 7.22) if they expected a 10 minute RI and for 10.42 seconds (SD = 6.85) if they expected a 30 second RI, B_{01} = 2.29. We found weak evidence in favor of the null that subjects were not varying their study time based on expected RI.

***Questionnaire results*.** As with Experiment 3, we once again administered a questionnaire to see if subjects used strategies more for one expected RI than for another. The results were almost entirely in accordance with what was seen in Experiment 3. 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. Unlike 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 based on the time cues.

When asked if they tried harder on word pairs based on provided time cues, 28% of subjects said they tried harder on word pairs cued for 30 seconds, 25% said they tried harder on

word pairs cued for 10 minutes, 41% said they tried equally hard on all word pairs, and 7% reported being uncertain if they tried harder on certain word pairs versus others. As opposed to Experiment 3, where half of subjects tried harder on words cued for 30 seconds, the majority response for Experiment 4 was that subjects did not differentially allocate their encoding efforts.

Finally, subjects were also asked if they noticed the manipulation. Approximately one third (34%) of all subjects reported noticing that the time cues were not always reliable, while 66% of all subjects reported not noticing the manipulation.

CHAPTER 6

GENERAL DISCUSSION

In Experiments 1–4, no overall differences were found in performance based on expected RIs. The consistently null findings across all four experiments motivated an analysis on pooled data to assess the reliability of the observed null differences. The goal of the analysis was to assess the degree of confidence we could have in asserting that no real differences existed between groups. The pooled data are presented in Figure 5.

Data Analysis Across all Experiments

Confidence intervals. Collapsing data across all experiments resulted in a sample of 257 subjects. Because RIs differed in length across experiments, they were now generally referred to as the "long" RI and the "short" RI. Thus, the 2 x 2 design was actual RI (long or short) by expected RI (long or short). The primary goal was to assess the reliability of the null difference between the long and short expected RI groups. Reliability was assessed via a 95% confidence interval analysis. For the actually short and long RI groups, a difference score was obtained for each subject by subtracting the performance score when expecting a short RI from the performance score when expecting a long RI. The 95% confidence interval for differential performance by expectation on the actually short RI was [-2.43%, 1.79%]. The confidence interval reflects a miniscule range of differential performance on either side of the mean. Subjects' differential performance by expected RI was a reliably null finding. For the actually long RI, the 95% confidence interval for differential performance by expected RI was [-2.10%, 1.34%]. As with the actually short RI, differential performance based on expected RI was tightly around 0 for the actually long RI.

Power analysis. As another way to assess the reliability of the observed null findings, we conducted a power analysis assuming a small effect size between performance scores on the expected RIs. Cohen (1988) defined a small effect size as $d = .20$. We used an adjusted value based on the within-subjects design of $d = .41$.[1] The power to obtain a small effect size between the expected RI groups was $(1 - \beta) = 0.996$. Collapsing across experiments, the combined experiments would have detected a small effect size of expected RIs 99.6% of the time. It is highly unlikely that a difference of such a magnitude actually existed between the expected long and short RI groups given this finding.

Differences in performance based on expected RI. We also analyzed the pooled data for differences in performance between the expected RIs at each of the actual RIs. For the actually long RI, subjects recalled 19% (*SD* = 17%) of items when expecting a long RI and they recalled 19% (*SD* = 19%) of items when expecting a short RI, $B_{01} = 14.48$. For the actually short RI, subjects recalled 37% (*SD* = 24%) of items when expecting a long RI and they recalled 37% (*SD* = 22%) of items when expecting a short RI, $B_{01} = 13.90$. In both cases, using Jeffrey's (1961) interpretation of Bayes Factors, this constitutes strong evidence for the null hypothesis. Generally, the evidence strongly suggests that, within the parameters of our experiments, learners do not exhibit differential memory performance based on expected RI.

Comparing RI manipulation detectors to non-detectors. To see if performance varied as a function of whether subjects detected our RI manipulation, we collapsed the data from

[1] For the within-subjects case, *d* must be adjusted to reflect the increase in power afforded by repeated measures. That is, standard power analysis assumes two independent groups; therefore it assumes that the groups show no correlation. However, groups obtained through repeated measures can be expected to correlate to a moderate degree. Cohen's *d* must therefore be adjusted to make the value applicable to standard power analysis, where no correlation is assumed. The adjusting equation provided by Cohen (p. 49) is $d = d / \sqrt{(1 - r)}$. The performance scores between expected RIs had a correlation coefficient of $r = .76$. When applied to a small effect size ($d = .20$) for the purpose of power analysis, this equation (using $r = .76$, the correlation coefficient between performance scores when expecting long or short RIs) resulted in an adjusted value of $d = .41$.

Experiments 3 and 4 and compared the performance of those subjects who reported detecting the manipulation versus those who did not. At the actually long RI, subjects who detected the manipulation recalled 16% (*SD* = 18%) of items when expecting a long RI and 16% (*SD* = 20%) of items when expecting a short RI, B_{01} = 6.47. At the actually short RI, they recalled 44% (*SD* = 27%) of items when expecting a long RI and 43% (*SD* = 21%) of items when expecting a short RI, B_{01} = 6.09.

At the actually long RI, subjects who did not notice the manipulation recalled 20% (*SD* = 19%) of items when expecting a long RI and 22% (*SD* = 22%) of items when expecting a short RI, B_{01} = 3.06. At the actually short RI, they recalled 43% (*SD* = 24%) of items when expecting a long RI and 44% (*SD* = 23%) of items when expecting a short RI, B_{01} = 6.06. The evidence thus suggests that detecting the violation of expectations did not influence performance as a function of expectation.

Comparing rote rehearsers to non-rote rehearsers. To see if performance varied as a function of the extent to which subjects engaged in rote rehearsal, we again collapsed the data from Experiments 3 and 4 and split the data into high and low rote rehearsers. (The data were median split because there were not enough responses at each scale response to analyze performance for each level of reported rote rehearsal.) At the actually long RI, high rote rehearsers recalled 14% (*SD* = 12%) of items when expecting a long RI and they recalled 15% (*SD* = 18%) of items when expecting a short RI, B_{01} = 5.60. At the actually short RI, high rote rehearsers recalled 40% (*SD* = 24%) of items when expecting a long RI and they recalled 39% (*SD* = 19%) of items when expecting a short RI, B_{01} = 5.65.

At the actually long RI, low rote rehearsers recalled 21% (*SD* = 22%) of items when expecting a long RI and they recalled 23% (*SD* = 24%) of items when expecting a short RI, B_{01} =

4.43. At the actually short RI, low rote rehearsers recalled 45% (*SD* = 24%) of items when expecting a long RI and they recalled 46% (*SD* = 25%) of items when expecting a short RI, B_{01} = 6.24. Both groups reveal no effect of expectation.

Present Findings

We obtained strong evidence that learners do not account for anticipated RI when encoding to-be-learned items. Why might retention intervals be largely neglected in our metacognitive updating? It may be the case that learners fail to accurately predict the level of forgetting that occurs between study and test. Other research has demonstrated that students are poor predictors of their own forgetting. For instance, Koriat and Bjork (2005) had subjects study associated word pairs and then make JOLs for an upcoming cued recall test 48 hours after study. Subjects' JOLs were overly optimistic, and they recalled about half as many target words as predicted after study. When estimates of forgetting are collected between subjects, researchers have found that subjects will predict the same level of performance on an immediate test as on a test following a one year RI (Koriat, Bjork, Sheffer, & Bar, 2004). These findings suggest that participants' metacognitive monitoring is rather insensitive to an expected RI, a finding that the present results support.

The present experiments may also have had design qualities that fostered neglect for expected RI. For one, it may have been the case that the time cues provided with each word pair were insufficient in making the expected RI a salient enough feature of the study session so that it would alter subjects' study. Subjects had little time to perceive and appreciate how long they had until the current word pair would be tested. Also, merely informing subjects of the timing of an upcoming test may have been insufficient. Perhaps a blocked design, in which an entire set of word pairs would be studied and then tested after the same RI, would give subjects a fuller

appreciation of the expected RIs. This would be akin to other studies in the test expectancy paradigm that induced expectation of test formats by giving subjects practice study-test sessions before a critical study-test session.

Finally, the present experiment used stimuli that were neither as complex nor perhaps as interesting to our subjects as course learning materials. Studying word pairs is a task not frequently engaged in outside the laboratory. Perhaps if our subjects had studied excerpts from stories, or expository passages from textbooks, their interest in the materials would have been higher. Research has shown that subjects will choose to study materials that they deem interesting over materials they deem as uninteresting (i.e. Son & Metcalfe, 2000). Increased motivation may have led to differential test performance between expected RIs. The present experimental task may have cultivated a sense of apathy towards the provided time cues.

The present research found strong evidence that expected RI is given little consideration at encoding. These findings differ from results in the study-time allocation literature that suggest that students focus their efforts on cases that they expect to be more difficult, but are in general accordance with much of the test-expectancy literature indicating students often fail to effectively tailor their study methods to features of an upcoming test.

TABLES

Table 1

Average usage ratings of encoding strategies from questionnaire in Experiments 3 and 4

Strategy	Description	Average Rating, Exp. 3 (*SD*)	Average Rating, Exp. 4 (*SD*)
Cue-Target Association	*Made associations between the left hand and right hand word in a pair*	5.28 (1.64)	4.69 (2.23)
Inter-Item Association	*Made associations between multiple pairs across the list*	2.73 (1.62)	2.44 (1.61)
Inter-Item Narrative	*Used groups or pairs of words in a sentence, phrase, or story*	2.48 (1.84)	2.84 (2.08)
Intra-Item Narrative	*Used a single pair or word in a sentence, phrase, or story*	3.43 (2.10)	3.85 (2.29)
Mental Imagery	*Used mental imagery (formed a picture in your head)*	3.65 (1.67)	3.84 (2.02)
Observation	*Just read or looked at the words*	5.00 (1.64)	4.59 (1.82)
Personal Significance	*Related words to something personally significant*	4.28 (1.90)	3.93 (1.77)
Rote Rehearsal	*Repeated individual words or pairs over and over*	5.70 (1.55)	5.31 (1.62)
Target Focus	*Focused more on the right-hand words*	5.10 (1.65)	4.56 (1.81)
Target-Target Association	*Made associations between the right-hand words across multiple pairs*	2.33 (1.72)	2.21 (1.43)
Verbalization	*Spoke words out loud or under your breath*	5.03 (2.03)	4.77 (2.04)

Table 2

Proportion of subjects using encoding strategies based on time cues in Experiments 3 and 4. Strategies are listed from most to least highly rated.

Experiment 3				
Strategy	Words cued for 10 minutes	Words cued for 30 seconds	Same Amount	Not Sure
Rote Rehearsal	0.07	0.22	0.47	0.25
Cue-Target Association	0.13	0.10	0.45	0.32
Target Focus	0.05	0.05	0.55	0.35
Verbalization	0.10	0.03	0.48	0.38
Observation	0.17	0.10	0.48	0.25
Personal Significance	0.15	0.12	0.58	0.15
Mental Imagery	0.08	0.10	0.68	0.13
Intra-Item Narrative	0.03	0.25	0.67	0.05
Inter-Item Association	0.05	0.23	0.60	0.12
Inter-Item Narrative	0.10	0.07	0.32	0.52
Target-Target Association	0.10	0.13	0.65	0.12
Experiment 4				
Strategy	Words cued for 10 minutes	Words cued for 30 seconds	Same Amount	Not Sure
Rote Rehearsal	0.07	0.22	0.47	0.25
Verbalization	0.10	0.03	0.48	0.38
Cue-Target Association	0.13	0.10	0.45	0.32
Observation	0.17	0.10	0.48	0.25
Target Focus	0.05	0.05	0.55	0.35
Personal Significance	0.15	0.12	0.58	0.15
Intra-Item Narrative	0.03	0.25	0.67	0.05
Mental Imagery	0.08	0.10	0.68	0.13
Inter-Item Narrative	0.10	0.07	0.32	0.52

Table 2 (cont.)

Inter-Item Association	0.05	0.23	0.60	0.12
Target-Target Association	0.10	0.13	0.65	0.12

FIGURES

Figure 1. Mean cued-recall performance as a function of actual retention interval (1 minute vs. 4 minutes) and expected retention interval (1 minute vs. 4 minutes) in Experiment 1.

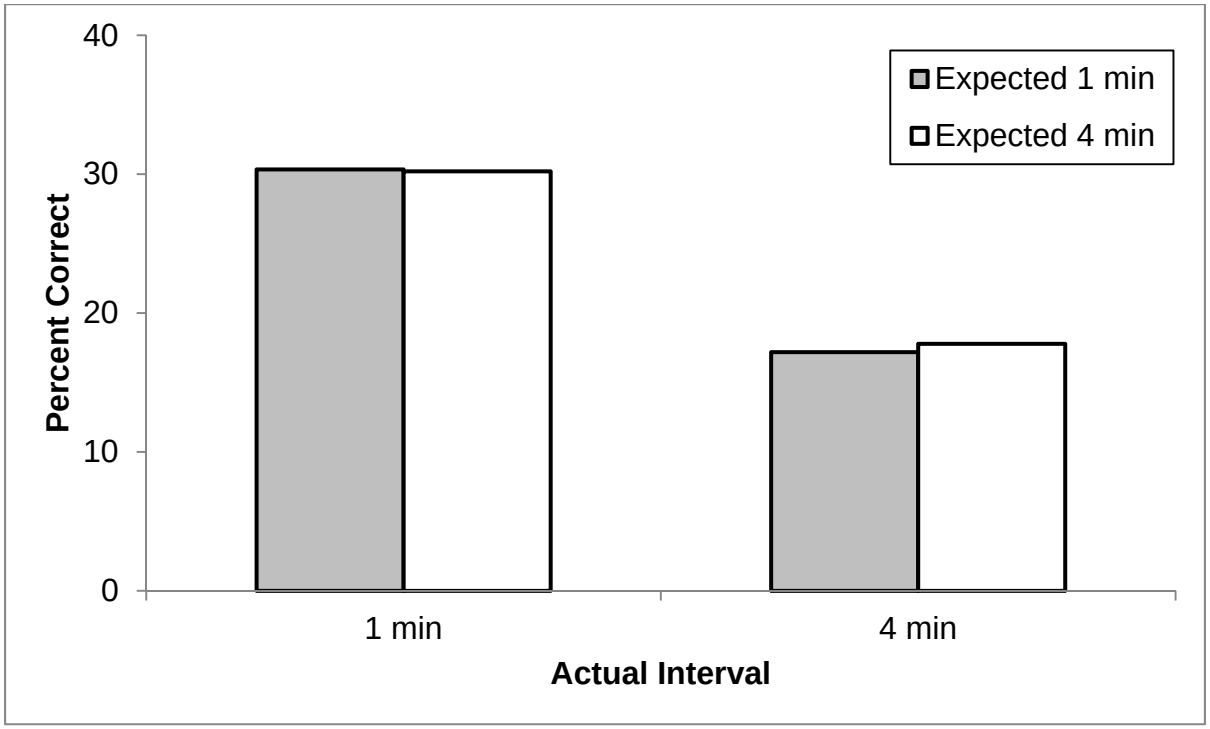


Figure 2. Mean cued-recall performance as a function of actual retention interval (30 seconds vs. 3 minutes) and expected retention interval (30 seconds vs. 3 minutes) in Experiment 2.

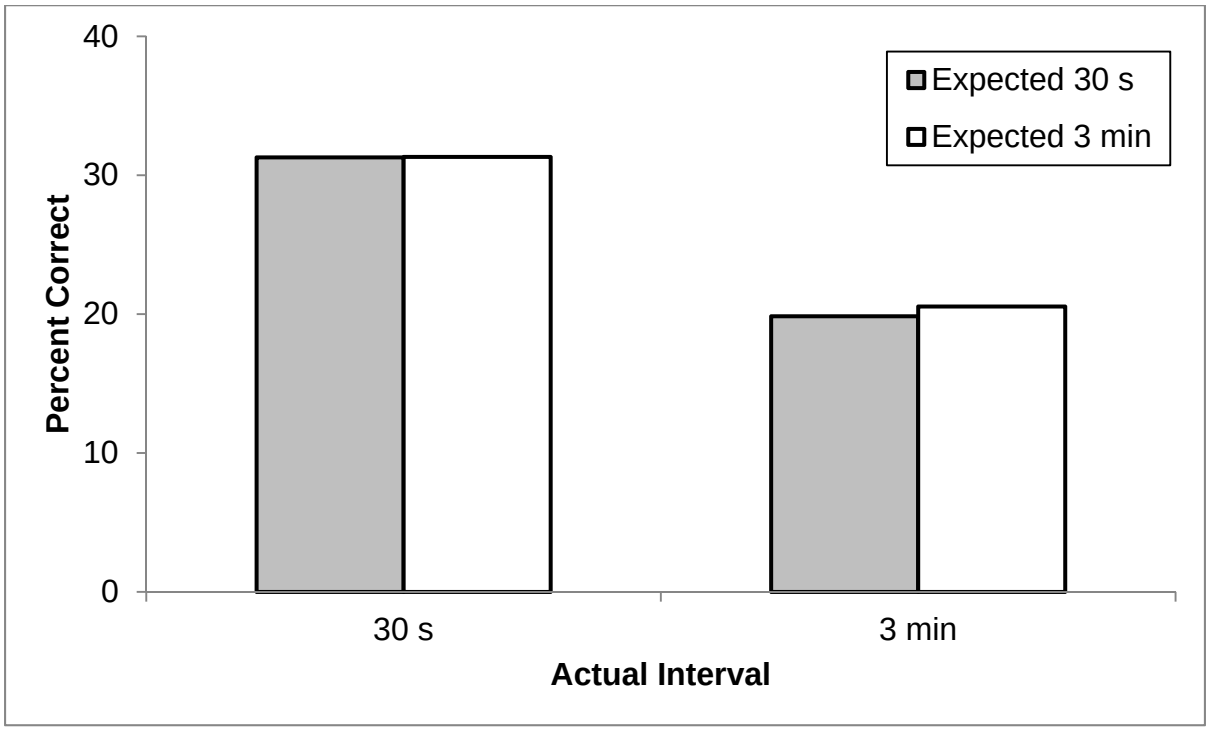


Figure 3. Mean cued-recall performance as a function of actual retention interval (30 seconds vs. 10 minutes) and expected retention interval (30 seconds vs. 10 minutes) in Experiment 3.

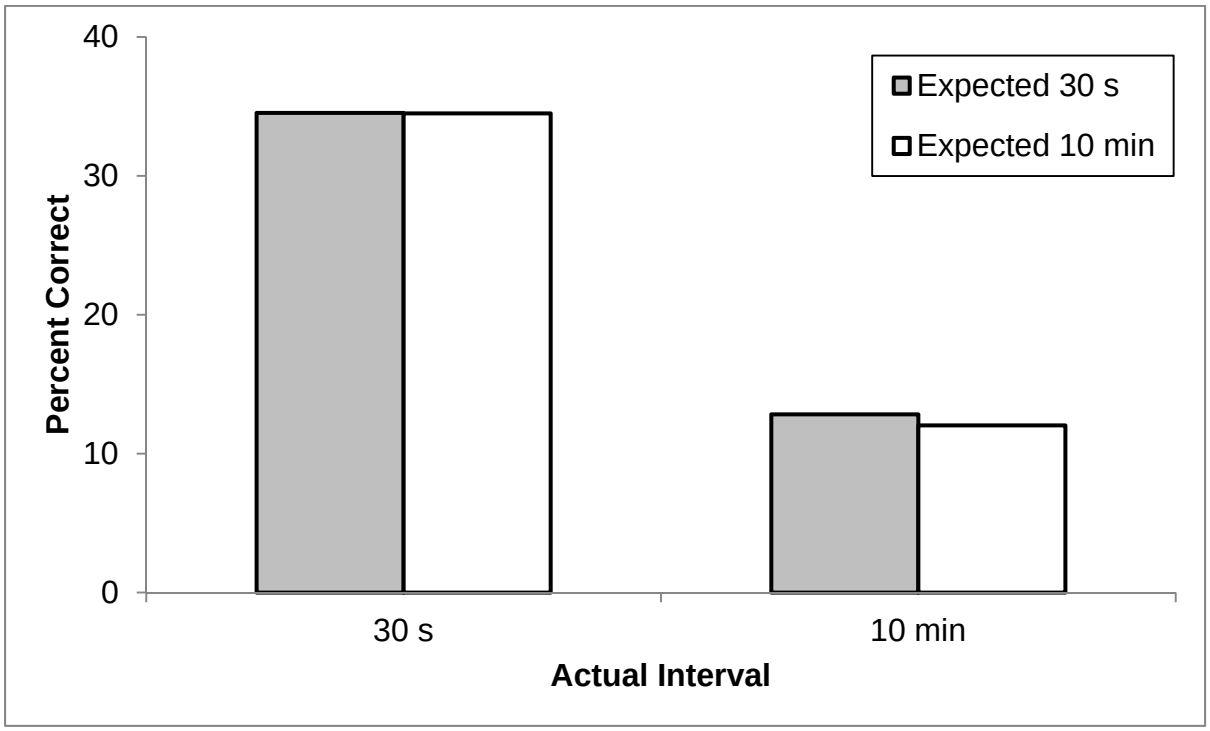


Figure 4. Mean cued-recall performance as a function of actual retention interval (30 seconds vs. 10 minutes) and expected retention interval (30 seconds vs. 10 minutes) in Experiment 4.

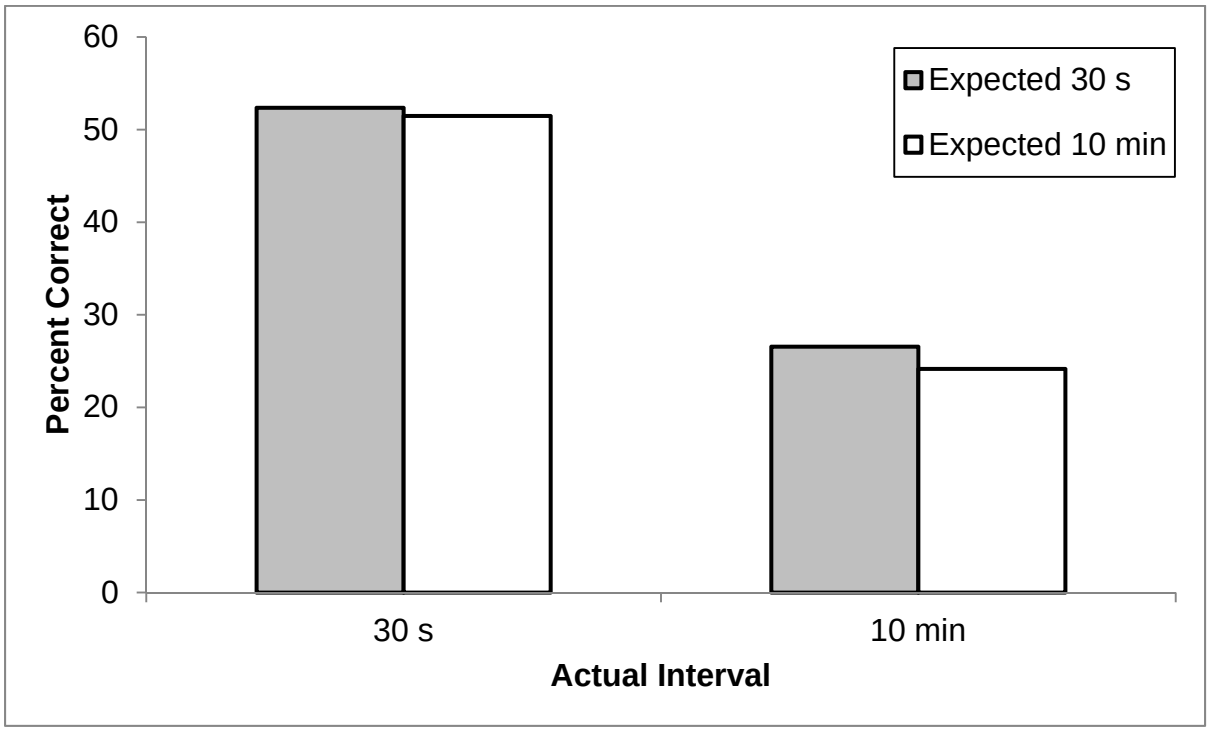
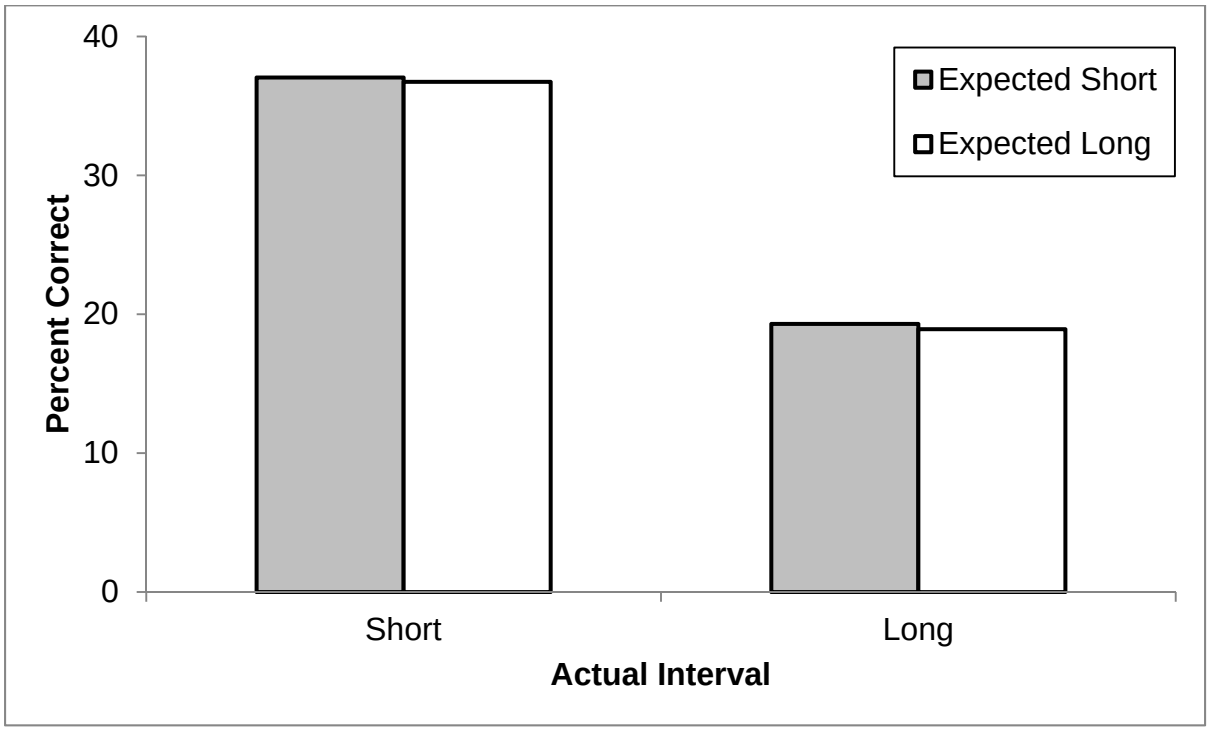


Figure 5. Mean cued-recall performance as a function of actual retention interval (short vs. long) and expected retention interval (short vs. long) across Experiments 1–4.



REFERENCES

Balota, D. A., & Neely, J. H. (1980). Test-expectancy and word-frequency effects in recall and recognition. *Journal of Experimental Psychology: Human Learning & Memory, 6*(5), 576-587.

Belmont, J. M., & Butterfield, E. C. (1971). Learning strategies as determinants of memory deficiencies. *Cognitive Psychology, 2*(4), 411-420.

Benjamin, A. S., & Bird, R. D. (2006). Metacognitive control of the spacing of study repetitions. *Journal of Memory & Language, 55*(1), 126-137.

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*(1), 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*(3), 354-380.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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*(2), 249-257.

Dufresne, A., & Kobasigawa, A. (1989). Children's spontaneous allocation of study time: Differential and sufficient aspects. *Journal of Experimental Child Psychology, 47*(2), 274-296.

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. (2015). The metacognitive foundations of effective remembering. In J. Dunlosky & S. K. Tauber (Eds.), *Oxford handbook of metamemory*. 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 & Cognition, 38*(3), 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: Springer.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-development theory. *American Psychologist, 34*, 906–911.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116*(2), 439-453.

Hall, J. W., Grossman, L. R., & Elwood, K. D. (1976). Differences in encoding for free recall vs. recognition. *Memory & Cognition, 4*(5), 507-513.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: 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*(2), 223.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31*(2), 187-194.

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*(4), 643-646.

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*(1), 36-69.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 32*(3), 609-622.

Le Ny, J., Denhiere, G., & Le Taillanter, D. (1972). Regulation of study-time and interstimulus similarity in self-paced learning conditions. *Acta Psychologica, 36*(4), 280-289.

Leonard, J. M., & Whitten II, W. B. (1983). Information stored when expecting recall or recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 9*(3), 440-455.

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*(2), 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*(1), 47-60.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory & Cognition, 18*(2), 196-204.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131*(3), 349-363.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin and Review, 15*(1), 174-179.

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*(4), 530-542.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory & Language, 52*(4), 463-477.

Neely, J. H., & Balota, D. A. (1981). Test-expectancy and semantic-organization effects in recall and recognition. *Memory & Cognition, 9*(3), 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*(4), 207-213.

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, & Cognition, 14*(4), 676-686.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In

G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 125-141). New York: Academic Press.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1-25). Cambridge, MA: MIT Press.

Rouder, J. N., Speckman, P. L., Sun, D., & Morey, R. D. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 26*(1), 204-221.

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 & Cognition, 25*(4), 1024-1037.

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, & Cognition, 35*(5), 1352-1358.

Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory & Language, 64*(2), 109-118.

Schmidt, S. R. (1988). Test-expectancy and individual-item versus relational processing. *The American Journal of Psychology, 101*(1), 59-71.

Thiede, K. W. (1996). The relative importance of anticipated test format and anticipated test difficulty on performance. *The Quarterly Journal of Experimental Psychology, 49*(4), 901-918.

von Wright, J. (1977). On the development of encoding in anticipation of various tests of

retention. *Scandinavian Journal of Psychology, 18*(2), 116-120.

von Wright, J., & Meretoja, M. (1975). Encoding in anticipation of various tests of retention. *Scandinavian Journal of Psychology, 16*(2), 108-112.