Timing of Quizzes during Learning: Effects on Motivation and Retention

Alice F. Healy, Matt Jones, Lakshmi Lalchandani, and Lindsay Anderson Tack

University of Colorado Boulder

Author Note

Alice F. Healy, Matt Jones, Lakshmi Lalchandani, and Lindsay Anderson Tack, Department of Psychology and Neuroscience, University of Colorado Boulder.

This research was supported in part by National Science Foundation Grant DRL1246588 to the University of Colorado. We thank the members of the Center for Research on Training at the University of Colorado for helpful suggestions about this research. We are especially grateful to Shaw Ketels for directing us to the cognitive engagement instrument developed by Rotgans and Schmidt and Mary Chapman for help testing participants in Experiment 2. Parts of Experiment 1 were presented at the 2014 meeting of the Psychonomic Society in Long Beach, California.

Correspondence concerning this manuscript should be addressed to Alice F. Healy, Department of Psychology and Neuroscience, Muenzinger Building, 345 UCB, University of Colorado, Boulder, CO 80309-0345. E-mail: alice.healy@colorado.edu

Abstract

This paper investigates how the timing of quizzes given during learning impacts retention of studied material. We investigate the hypothesis that interspersing quizzes among study blocks increases student engagement, thus improving learning. Participants learned 8 artificial facts about each of 8 plant categories, with the categories blocked during learning. Quizzes about 4 of the 8 facts from each category occurred either immediately after studying the facts for that category (standard) or after studying the facts from all 8 categories (postponed). In Experiment 1, subjects were given tests, shortly after learning and several days later, including both the initially quizzed and unquizzed facts. Test performance was better in the standard than in the postponed condition, especially for categories learned later in the sequence. This result held even for the facts not quizzed during learning, suggesting that the advantage cannot be due to any direct testing effects. Instead the results support the hypothesis that interrupting learning with quiz questions is beneficial because it can enhance learner engagement. Experiment 2 provided further support for this hypothesis, based on participants' retrospective ratings of their task engagement during the learning phase. These findings have practical implications for when to introduce quizzes in the classroom.

Timing of Quizzes during Learning: Effects on Motivation and Retention

Recent technological advances have radically altered how classroom instructors can interact with large groups of students. Universities and high schools are rapidly adopting student electronic hand-held response devices (“clickers”), which enable instructors to administer brief multiple-choice quizzes at any point during a lecture, with immediate feedback for both teachers and students (Smith et al., 2009). We refer to this as the *clicker technique*.

The potential benefits of the clicker technique are twofold. For the students, there is a learning benefit, derived from the opportunities to test their knowledge, from feedback following those tests, and from the additional engagement afforded by the interactive nature of the clicker procedure. For the instructor, there is an assessment benefit, derived from the ability to maintain an ongoing evaluation of students’ understanding and to predict their later retention of the material. There is also great usage flexibility within this technique, in terms of when to ask questions, what questions to ask, how to structure incorrect alternative answers, and what type of feedback to present to the class. Thus there are many important theoretical and practical questions about how to maximize the clicker technique’s benefits for learning and for assessment.

The experiments reported here focus on the timing of clicker questions. Specifically, we tested whether it is better to interpose clicker questions throughout the course of a lecture, or to present all questions at the end. The latter option might be superior because it does not interrupt the flow of the lecture, but the former option might benefit students’ motivation by keeping them engaged in the task of listening to the instructor, which can otherwise be considered monotonous (Kole, Healy, & Bourne, 2008).

The experiments used a laboratory paradigm in which participants studied artificial facts and were given quizzes meant to simulate clicker questions. Experiment 1 also included posttests to simulate exams, whereas Experiment 2 replaced posttests with a retrospective survey in which students rated their task engagement during the training period. To preview the results, mean posttest performance in Experiment 1 was reliably better when quiz questions were interleaved among study blocks than when quizzes were postponed until all study blocks were completed, supporting the hypothesis that interrupting a lecture with clicker questions aids student engagement. This result held even for unquizzed items, implicating a general motivational effect over any direct effects of retrieval and feedback on individual items during quizzing. Block-level analysis further supports this engagement hypothesis, in that the advantage from interleaved quizzes was greater for items presented later in the training phase. The engagement ratings provided in Experiment 2 yield more direct evidence for the engagement hypothesis: Self-reported engagement declined across blocks of trials, and this decline was smaller when the quizzes were interleaved rather than postponed. Thus, the results suggest that it is best to present clicker questions throughout the course of a lecture as opposed to at the end.

Experimental Approach

The goal of the experiments was to test the hypothesis that the clicker technique has motivational benefits, in that it keeps students more engaged during a lecture. This hypothesis is grounded in experimental findings supporting the *cognitive antidote* principle, whereby introducing an additional cognitive requirement to an otherwise tedious task can improve performance by increasing attention to the task (Kole et al., 2008). Likewise, Szpunar, Khan, and Schacter (2013) have argued that interrupting an extended task can reduce mind wandering and thereby improve performance on the task. Our general proposal is that, at least for a task that makes minimal cognitive demands on the participants, any aspect of the task environment that requires additional attention from the participant will increase his or her engagement with the task. This enhanced engagement will in turn increase the participant's motivation to perform well in the task, which should manifest in measures such as reduced mind wandering, increased note taking (Szpunar, Jing, & Schacter, 2014; Szpunar et al., 2013), and increased learning, retention, and transfer. One prediction that follows from the cognitive antidote principle, in an educational setting, is that students will learn more when quiz questions are interspersed during lecture or training than when they are all presented at the end.

We tested this prediction by comparing the *standard* procedure used in most classrooms, in which quizzes are presented interspersed throughout learning, to a *postponed* procedure, in which all quizzes are presented at the end of learning (see Figure 1). Specifically, participants studied 64 novel facts about plants, divided into eight plant-type categories. They were then quizzed on those facts, with correct answer feedback, either immediately after study of each category (standard condition) or after studying all of the categories (postponed condition). The quiz questions were formatted as multiple-choice to simulate clicker questions given during a classroom lecture. Following training in Experiment 1, participants were given two tests on all studied facts: a *posttest* approximately five minutes later (following a distractor task) and a *retention test* either two or seven days later. The primary prediction, based on the hypotheses laid out above, was that test performance would be superior for participants in the standard condition than for participants in the postponed condition. Additionally, the advantage for the standard condition at test was predicted to be greater for the later-studied categories because the benefit of the cognitive antidote should grow over the course of training (given that the potential for boredom is expected to grow). Following training in Experiment 2, participants rated several aspects of their engagement during each study block, in order to test the effects of the timing manipulation on motivation directly rather than via its consequences for learning. The primary prediction paralleled that for Experiment 1: greater reported engagement in the standard condition, with the difference between conditions being larger for later-studied categories.

One challenge in designing the experiments was that the quizzes themselves might influence learning, via what is known as the testing effect. That is, testing can be as beneficial to learning as study (or more so), likely reflecting the fact that testing provides an opportunity to practice retrieving and generating the correct answers (Karpicke & Roediger, 2007, 2008; Roediger & Butler, 2011; Roediger & Karpicke, 2006). Moreover, the learning benefits from the quizzes might differ between conditions due to the differences in their timing. Specifically, research on the spacing effect in learning has shown that spaced practice is better for long-term retention than is massed practice (e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). With spaced practice, the review opportunities for each to-be-learned item are temporally separated, either with empty lags (e.g., Tzeng, 1973) or with intervening items (e.g., Thios & D'Agostino, 1976). With massed practice, the review opportunities for any item occur closely together in time. In both cases, the review opportunities for each item—other than the first opportunity—can be in the form of additional study (i.e., the participant reads or hears the material again), or they can be in the form of tests (i.e., the participant is quizzed on the material). Thus, the review opportunities in the present Experiment 1 were the initial study of each item and the quiz of that item, and therefore the aspect of our paradigm that is relevant to the spacing effect is the time between the study and quiz. Together, the spacing and testing effects suggest that quizzes will provide more of a learning benefit if they occur after a delay, as in the postponed condition of the present experiment, than immediately after study, as in the standard condition.

To control for any item-specific effects of quizzing and of quiz timing on posttest and retention test performance, the quizzes given in both experiments covered only half of the studied facts, whereas the tests (given in Experiment 1) covered all facts. Our primary analyses of the tests were restricted to performance on the half of the questions not covered in the quizzes. The studied facts in the present experiments were all unrelated and, therefore, practice or quizzing on any one fact should not directly influence or facilitate knowledge of any other fact (see, e.g., Pan, Gopal, & Rickard, 2016). Any difference between conditions in test performance on these items must therefore be due to general effects of the study-quiz schedule, due to mechanisms such as engagement, rather than any direct testing effects from the specific material queried in quiz questions.

As one final manipulation, the questions on the posttest and retention test in Experiment 1 were divided into two forms, one matching the form given during training, and the other a novel form requiring a small degree of transfer (generalization) from the trained fact. This manipulation enabled us to test whether any effect found for the timing manipulation differed according to the level of transfer required.

Experiment 1

To review, in Experiment 1, participants studied the eight categories of plant facts, and were then quizzed on half of those facts, either immediately after each category was studied (standard condition) or after all eight categories were studied (postponed condition). Both 5 min and 2 or 7 days after training, participants were tested on all of the studied facts.

Method

Participants. Ninety-six University of Colorado Boulder introductory psychology undergraduate students participated in and completed this two-session experiment in exchange for partial course credit. An additional 20 participants were tested but excluded from all analyses due to failure to return for the second session of the experiment. The 96 analyzed participants were assigned to four conditions in a fixed rotating order. The first 32 participants tested had a 1-week delay between sessions, and the remaining 64 participants had a 2-day delay between sessions. This change in the retention interval was made solely due to logistical considerations.

Materials. The experiment used a fact-learning task adopted from Anderson, Healy, Kole, and Bourne (2013), consisting of 64 facts about eight different categories of plants (trees, herbs, vines, weeds, wildflowers, fungi, shrubs, vegetables). Each category contained eight exemplars. To control for prior knowledge, each true plant name was replaced by an orthographically regular nonword that had been generated from true plant names in its category. Each fact was presented for study as a sentence containing a plant category, a descriptor phrase, and a name (e.g., "A wildflower that is used in make-up products is the Shasty"). Two questions were created for each fact: a *specific form* using the study sentence verbatim, and a *general form* using a fact entailed by the studied fact (e.g., "A wildflower that is used for beauty is the Shasty"). For each question, the plant name was replaced by a blank underline, with four options below labeled A through D. The three distractor options for each question were other plant names from the same category, with each plant name used as a distractor in three different questions.

An irrelevant letter-detection task was used as a distractor task between the fact-learning training phase and the posttest. The letter-detection task contained two paragraphs each followed by one comprehension question. One paragraph was written in a uniform color (black or red). The other paragraph was written with letter colors alternating between black and red. Participants were to circle either the letter sequence *the* or the letter *h* in each paragraph. The content of the paragraphs did not relate to the plant facts in any manner.

Design. The experiment consisted of two sessions, separated by one week or two days. In Session 1, each participant was trained in one of two conditions: standard or postponed. In the standard condition, the quiz for each category was presented immediately after the study block for that category. In the postponed condition, quizzes for all eight categories were presented (in order) after all eight study blocks had been completed (Figure 1).

Training was blocked so that facts about all of the plants in a given category were trained together in a block. A fixed order of categories was used for all participants, with the order of categories during testing the same as that during training.

At training, all facts were studied and quizzed in their specific forms. The quiz for each category covered only half of the studied facts, counterbalanced between subjects. Half of the participants received the even questions, and half received the odd questions.

The posttest (in Session 1) and the retention test (Session 2) both covered all facts from all categories. For the posttest, for each participant, in each block half of the questions were in specific form and half were in general form, with the assignment of form to question counterbalanced across participants. For the retention test, the assignments were reversed, such that each participant saw each question in a different form on the two tests.

The assignment of correct answers to response letters (A, B, C, D) was counterbalanced such that each letter was correct exactly twice within each block of eight questions for each participant. These assignments were varied across tests for each participant, so that a participant could not answer a question on the posttest or retention test simply by recalling which letter was correct on the quiz or previous test.

In sum, the design for training (i.e., analyzing quiz performance as a dependent measure) was a 2 x 8 mixed factorial, including condition (standard vs. postponed) as a between-subjects variable and training block (i.e., fact category, 1-8) as a within-subject variable. The design for testing (i.e., analyzing test performance) was a 2 x 2 x 2 x 2 x 8 mixed factorial, with condition as a between-subjects variable, and test (posttest vs. retention test), question form (specific vs. general), quiz status (quizzed vs. unquizzed facts), and testing block (i.e., fact category, 1-8) as within-subject variables.

Procedure. Participants were tested in separate rooms on Apple *i*Mac computers. In Session 1, participants first read instructions on the computer screen, informing them that they would be viewing several sets of eight facts about different types of plants, and that they would be tested on their ability to remember those facts. Before the training phase began, they were given a specific example of the type of question they would be shown: "When you are tested, you will be presented with part of the fact (e.g., A flower that has thorns is the ___________?) and four possible answer choices, from which you will select the correct letter answer using the keyboard." During each study block, the eight facts for that category were presented individually in a fixed random order (the same for all participants) for 3 s each. During each quiz block, the four questions were presented individually in a new random order. The participant was given 9 s to respond, by typing A, B, C, or D, and then the complete fact with the correct answer filled in was displayed for 6 s, with the answer choices still visible. Between the training phase and the posttest, the participant completed the distractor letter-detection task, lasting about 5 min. The posttest presented questions about all 64 facts, in a blocked order (categories 1-8), with question order randomized within blocks (separately for each participant). Responding was as in the quizzes, and no feedback was given. Session 2 consisted of only the retention test, which followed the same procedure as the posttest with a new random ordering of questions within each block.

Results

Quiz performance. An initial analysis compared quiz performance between conditions, using a mixed factorial analysis of variance (ANOVA) with condition as a between-subjects variable and block as a within-subject variable. The analysis revealed a main effect of condition, $M_{\text{standard}} = .686$, $M_{\text{postponed}} = .441$, $F(1,94) = 58.87$, $MSE = .195$, $\eta^2 = .385$, $p < .001$. This difference was expected, because quizzes in the standard condition occurred immediately after study of the corresponding facts, whereas in the postponed condition studying and quizzing of each fact category were separated by studying or quizzing of all seven other categories.

There was also a main effect of block, $F(7, 658) = 5.07$, $MSE = .059$, $\eta^2 = .051$, $p < .001$, reflecting general improvement across blocks of training ($M_1 = .513$, $M_2 = .505$, $M_3 = .576$, $M_4 = .609$, $M_5 = .518$, $M_6 = .656$, $M_7 = .526$, $M_8 = .602$), and there was a significant linear contrast across blocks, $F(1, 658) = 7.45$, $\eta^2 = .011$, $p = .006$. However, the improvement is not monotonic across blocks, presumably because the questions in the different plant categories differed in intrinsic difficulty. The interaction of condition and block was not significant, $F(7, 658) < 1$, and the interaction of condition and the linear contrast across blocks was also not significant, $F(1, 658) < 1$.

Test performance. Figure 2 displays mean test performance as a function of condition and block, averaged over the posttest and retention test. The two tests are averaged in this figure to highlight the Condition × Block interaction, which did not differ between the two tests (i.e., there was no significant three-way interaction, as reported below). The descriptive results indicate an advantage for the standard condition that increases for later blocks, with little or no difference in the initial blocks (see the red difference curve in Figure 2). This result confirms both of the primary predictions from the cognitive antidote principle, that the standard condition was more engaged and that this engagement advantage grew over the course of training.

Overall. Formal analysis was carried out with a mixed factorial ANOVA, with a between-subjects variable of condition (standard vs. postponed) and within-subject variables of block (1-8), quiz status (quizzed vs. unquizzed), test (posttest vs. retention test), and question form (specific vs. general).

The analysis confirmed the main effect of condition, $M_{\text{standard}} = .432$, $M_{\text{postponed}} = .371$, $F(1, 94) = 6.34$, $MSE = .903$, $\eta^2 = .063$, $p = .014$. The main effect of block was significant, $F(7, 658) = 9.38$, $MSE = .168$, $\eta^2 = .091$, $p < .001$, and there was a significant linear contrast across blocks, $F(1, 658) = 10.85$, $\eta^2 = .016$, $p = .001$. More importantly, the interaction between block and condition was significant, such that the advantage for the standard condition was greater on the later blocks, $F(7, 658) = 2.28$, $MSE = .168$, $\eta^2 = .024$, $p = .027$. The interaction of condition and the linear contrast across blocks was also significant, $F(1, 658) = 11.86$, $\eta^2 = .018$, $p < .001$. This pattern held for both tests; the three-way interaction between block, condition, and test was not significant, $F(7, 658) = 1.55$, $MSE = .111$, $\eta^2 = .016$, $p = .149$. These results are consistent with the cognitive antidote principle and the assumption that the potential for boredom and task disengagement increases over the course of training.

The analysis also revealed a main effect of quiz status, $F(1, 94) = 47.41$, $MSE = .170$, $\eta^2 = .335$, $p < .001$. Participants performed better at test on the quizzed items than on the unquizzed items (M_{quizzed} = .438, $M_{\text{unquizzed}}$ = .365), in agreement with the testing effect (Roediger & Karpicke, 2006). This effect held for both the posttest and the retention test; the interaction of quiz status and test was not significant, $F(1, 94) < 1$. Most important is the observation that the advantage for interspersing quizzes occurred both for facts quizzed during training (M_{standard} = .473, $M_{\text{postponed}}$ = .402) and for facts not quizzed during training (M_{standard} = .391, $M_{\text{postponed}}$ = .340). In particular, the interaction of condition and quiz status was not significant, $F(1, 94) < 1$. Thus the advantage found for the standard condition cannot be attributed to any direct effects of quizzing individual items. This conclusion is further supported in the next subsection by an analysis restricted to the unquizzed items. The remainder of this subsection summarizes the other results of the full analysis on all test items.

Participants performed better on the posttest than on the retention test, reflecting forgetting across the delay between sessions, M_{posttest} = .446, $M_{\text{retention}}$ = .357, $F(1, 94) = 60.44$, $MSE = .200$, $\eta^2 = .391$, $p < .001$. There was also a significant interaction between test and condition, $F(1, 94) = 5.02$, $MSE = .200$, $\eta^2 = .051$, $p = .028$. The advantage of the standard condition was greater on the posttest (M_{standard} = .489, $M_{\text{postponed}}$ = .402) than on the retention test (M_{standard} = .375, $M_{\text{postponed}}$ = .339), in line with the overall forgetting from the first test to the second.

There was a main effect of question form, with participants performing better on specific-form questions (which matched the facts used during training) than on general-form questions, M_{specific} = .436, M_{general} = .367, $F(1, 94) = 52.39$, $MSE = .140$, $\eta^2 = .358$, $p < .001$. The interaction between question form and quiz status proved significant, $F(1, 94) = 17.09$, $MSE = .094$, $\eta^2 = .154$, $p < .001$. The pattern of the interaction indicates that the effects of question form and quiz status combine in a super-additive manner ($M_{\text{quizzed,specific}}$ = .488, $M_{\text{quizzed,general}}$ =

.387, $M_{\text{unquizzed,specific}}$ = .383, $M_{\text{unquizzed,general}}$ = .347). An interaction was also found between question form and test, $F(1, 94) = 16.76$, $MSE = .121$, $\eta^2 = .151$, $p < .001$. The decline from the posttest to the retention test was greater for questions in the specific form (M_{posttest} = .498, $M_{\text{retention}}$ = .373) than for questions in the general form (M_{posttest} = .393, $M_{\text{retention}}$ = .340).

A significant three-way interaction was found between block, question form, and condition, $F(7, 658) = 2.07$, $MSE = .088$, $\eta^2 = .022$, $p = .045$. The advantage of the standard over the postponed condition varied across blocks differently for the specific-form than for the general-form questions (see Figure 3). No other higher-order interactions were significant.

Unquizzed questions. A second ANOVA was performed restricted to the unquizzed items for each participant, to exclude any possible contributions of testing and spacing effects. Participants' experience with the unquizzed items was identical between the two conditions, and thus any differences in test performance on these items must be due to global effects of whether study and quizzing were separated or interleaved (although the correct answers to some of the unquizzed items did occur as lures for some of the quizzed items). This analysis replicated the significant effect of condition, $\text{M}_{\text{standard}}$ = .391, $\text{M}_{\text{postponed}}$ = .340, F(1, 94) = 4.12, MSE = .486, η^2 = .042, p = .045. The interaction between condition and block was marginally significant, F(7, 658) = 1.76, MSE = .133, η^2 = .018, p = .092, and the interaction of condition with the linear contrast for block was significant, F(1, 658) = 5.92, η^2 = .009, p = .015. As in the analysis above of all items, participants in the standard condition outperformed those in the postponed condition on the later blocks, with little or no difference in the initial blocks. Thus the critical results from above hold when restricted to unquizzed items.

Retention interval. A mixed factorial ANOVA was also conducted that included the between-subjects variable of retention interval (two days vs. one week), in addition to the other variables used in the previous analyses (condition, block, quiz status, test, and question form). All test items (viz., quizzed and unquizzed) were included in the analysis. Only a single effect involving retention interval was found. The interaction of test and retention interval was significant, $F(1, 92) = 12.67$, $MSE = .179$, $\eta^2 = .121$, $p < .001$, reflecting the fact that forgetting across the retention interval was less for the 2-day group ($M_{posttest} = .428$, $M_{retention} = .367$) than for the 1-week group ($M_{posttest} = .480$, $M_{retention} = .337$). This finding must be interpreted cautiously because of the lack of random assignment of retention intervals, but nevertheless it is a sensible and expected result.

Discussion

The results of Experiment 1 show a significant advantage for the standard condition over the postponed condition both during training and during testing. This advantage was predicted from the cognitive antidote principle (Kole et al., 2008) on the assumption that the interpolated quizzes serve a motivational function by dispelling boredom or reducing mind wandering (Szpunar et al., 2013), thereby keeping the participants engaged in the task. Additional support for this motivational interpretation comes from the pattern of performance across blocks. The data show large differences in performance levels across blocks, due in part to the fact that different item sets were used in the different blocks. However, these item effects cancel out when we look at the differences between the conditions (see the red curve in Figure 2). This difference curve shows a smooth, nearly linear effect of block number. That is, the advantage at test for the standard condition was stronger for fact categories that appeared later in training. This finding is naturally predicted by the cognitive antidote principle, because without the antidote of interspersed quizzes boredom should increase over time. Critically, the growing advantage for the standard over the postponed condition was found at test even for the questions that were not quizzed, so the advantage of interspersing cannot be due to direct effects of testing individual items. Also noteworthy is the finding that this pattern was observed for the general-form questions as well as for the specific-form questions that were shown during training, implying that the advantage from interspersed quizzes impacts mild transfer to related facts as well as retention of learned facts.

The fact that the advantage for the standard over the postponed condition was only observed for the later blocks provides an explanation for why the present study found a difference between the two conditions at test whereas no such difference was found in a recent study by Weinstein, Nunes, and Karpicke (2016), who also compared quizzes interspersed throughout learning to quizzes given only at the end of the learning period. For the quizzes themselves, Weinstein et al. found that performance was significantly better for the interspersed group than for the end-of-learning group. This is the expected result (and the same result obtained in the present experiment) because of the shorter delay between study and quiz for the former group. However, in contrast to the present findings, Weinstein et al. did not find any difference between the two quiz placement conditions during the test. One possible explanation is that they used only 10 questions in their learning and test phases, in contrast to the 64 questions used in the present experiment. After only 10 questions there was no difference between the two conditions in the present experiment as well (see Figure 2); as mentioned earlier, the difference between the two conditions increased across blocks of trials, presumably because of the increase in boredom as training trials progressed. Thus, it is possible that the finding by Weinstein et al. of no difference between the two conditions at test might be due at least in part to the relatively small number of questions they used so that boredom did not accumulate during learning to the same extent as it did in the present experiment.

Experiment 2

Although the findings of Experiment 1 are consistent with the cognitive antidote principle (Kole et al., 2008), neither that experiment nor any previous experiments supporting the cognitive antidote principle directly measured motivational variables, and such measurement needs to be done in order to map out more fully the psychological processes involved. Specifically, it would be useful to provide more direct evidence that boredom and task disengagement increased during training to a larger extent in the postponed condition than in the standard condition, in which the interpolated quizzes are assumed to serve a motivational function by dispelling boredom or increasing task engagement.

Experiment 2 thus attempted to measure the effect of quiz timing on motivational variables. Towards that end, we used a retrospective self-report engagement survey to assess the level of task engagement of the participants during every block of training. This survey was a modified and expanded version of an instrument developed by Rotgans and Schmidt (2011), who verified that their instrument was both reliable and valid for assessing situational cognitive engagement in an academic learning context. In order to use different verbal routes to elicit the participants' assessment of their engagement, the rating scale included six different descriptors, three positive (effort, engagement, motivation) and three negative (boredom, mind wandering, fatigue), in an attempt to obtain a single robust measure encompassing the types of motivational effects that might have arisen from the experimental manipulation. According to the cognitive antidote principle, engagement should decline across training blocks, and that decline should be larger in the postponed condition than in the standard condition. Because assessments of engagement interpolated during the training process itself would be likely to dispel boredom in the same way that interpolated quizzes would, the assessments were made instead after the training phase had concluded. As in Experiment 1, all of the facts about a given plant category occurred together in a single block of trials. Participants were told that they would be given the plant categories to rate in the order in which they were shown during training, which enabled them to rate their level of engagement for each block of trials whether or not they could recall the specific facts presented in the block.

Method

Participants. Ninety-six University of Colorado Boulder introductory psychology undergraduate students participated in and completed this single-session experiment in exchange for partial course credit. An additional 11 participants were tested but excluded from all analyses due to failure to complete all cells of the survey (6 participants), not following directions (1 participant), or experimenter error in not handing out the correct survey (4 participants). The 96 analyzed participants were assigned to four conditions in a fixed rotating order.

Materials. The fact-learning task used in Experiment 1 was used again, with no changes in Experiment 2.

Instead of the posttest an engagement survey was used in Experiment 2. The survey was a modified and expanded version of a cognitive engagement instrument developed by Rotgans and Schmidt (2011). It began with the following instructions, "You just studied 64 facts about plants. They were divided into 8 plant categories, with 8 plants in each category. Listed below are the plant categories in the order in which they were shown to you. Please rate to the best of your ability each of these categories in terms of each statement listed below." Six statements were listed below each category in the following order: (a) I put in a lot of effort to learn these facts, (b) I was bored when learning these facts, (c) I was engaged with learning these facts, (d) I was motivated to learn these facts, (e) My mind wandered when learning these facts, (f) I was fatigued when learning these facts. A five-point Likert scale was used with each point labeled in the following order: (a) Not true at all for me, (b) Not true for me, (c) Neutral, (d) True for me, (e) Very true for me. The first four categories were listed on the front side of the single-page survey (trees, herbs, vines, weeds), and the last four categories were listed on the back side of the survey (wildflowers, fungi, shrubs, vegetables), with each category including a 6 x 5 grid with the six statements labeling the rows and the five Likert scale points labeling the columns.

Design. The experiment consisted of a single session. As in Experiment 1, each participant was trained in one of two conditions: standard or postponed. These conditions were defined in the same way as in Experiment 1, and training was designed in the same way as in Experiment 1. Thus, the design for training (i.e., analyzing quiz performance as a dependent measure) was a 2 x 8 mixed factorial, including condition (standard vs. postponed) as a between-subjects variable and training block (i.e., fact category, 1-8) as a within-subject variable. The design for the engagement survey (i.e., analyzing engagement ratings) was also a 2 x 8 mixed factorial, with condition as a between-subjects variable and training block as a within-subject variable. The dependent variable used for the engagement survey was the average rating across the six statements used, with the three positive statements (put in effort, engaged, motivated) scored from 1 to 5 as on the Likert scale, and the three negative statements (bored, mind wandered, fatigued) reversed scored from 5 to 1.

Procedure. As in Experiment 1, participants were tested in separate rooms on Apple *i*Mac computers. The instructions for training, which were read on the computer screen, were the same as in Experiment 1. The training procedure, comprising the study and the quizzes, was also the same as in Experiment 1. Unlike Experiment 1, there was no distractor task or posttest.

Instead, immediately after training participants were given a two-sided single page survey, which they read on paper and filled out with a pen.

Results

Quiz performance. As in Experiment 1, an initial analysis compared quiz performance between conditions, using a mixed factorial ANOVA with condition as a between-subjects variable and block as a within-subject variable. The analysis revealed a main effect of condition, $M_{\text{standard}} = .703$, $M_{\text{postponed}} = .474$, $F(1,94) = 53.15$, $MSE = .190$, $\eta^2 = .361$, $p < .001$. Again, this difference was expected, because quizzes in the standard condition occurred immediately after study of the corresponding facts, whereas in the postponed condition studying and quizzing of each fact category were separated by studying or quizzing of all seven other categories. The main effect of block was not significant, $F(7, 658) = 1.53$, $MSE = .058$, $\eta^2 = .016$, $p = .155$, nor was the linear contrast across blocks, $F(1, 658) = 1.94$, $\eta^2 = .003$, $p = .165$. There was a significant interaction of condition and block, $F(7, 658) = 2.46$, $MSE = .058$, $\eta^2 = .026$, $p = .017$ (Standard: $M_1 = .693$, $M_2 = .615$, $M_3 = .724$, $M_4 = .698$, $M_5 = .714$, $M_6 = .698$, $M_7 = .745$, $M_8 = .740$; Postponed: $M_1 = .469$, $M_2 = .505$, $M_3 = .500$, $M_4 = .453$, $M_5 = .417$, $M_6 = .516$, $M_7 = .380$, $M_8 = .552$). Inspection of the means indicated that the interaction was due primarily to an especially small effect of condition on Block 2 relative to the other blocks. The effect of condition at Block 1 was at the median of the other six blocks, and the interaction of condition and the linear contrast across blocks was not significant, $F(1, 658) = 2.22$, $\eta^2 = .003$, $p = .137$.

Survey ratings. Figure 4 displays mean cognitive engagement ratings as a function of condition and block. The descriptive results indicate an advantage (i.e., higher ratings) for the standard over the postponed condition that increased for later blocks, with little or no difference in the initial blocks (see the red difference curve in Figure 4). This result is consistent with the prediction from the cognitive antidote principle that the standard condition would be more engaged and that this engagement advantage would grow over the course of study.

Formal analysis was conducted with a mixed factorial ANOVA, with a between-subjects variable of condition (standard vs. postponed) and a within-subject variable of block (1-8). The analysis yielded a marginally significant main effect of condition, $M_{\text{standard}} = 3.211$, $M_{\text{postponed}} = 2.984$, $F(1, 94) = 3.23$, $MSE = 3.085$, $\eta^2 = .033$, $p = .075$. The main effect of block was significant, $F(7, 658) = 12.84$, $MSE = 0.230$, $\eta^2 = .120$, $p < .001$, and there was a significant linear contrast across blocks, $F(1, 658) = 59.35$, $\eta^2 = .083$, $p < .001$. More importantly, the interaction between block and condition was significant, $F(7, 658) = 3.01$, $MSE = 0.230$, $\eta^2 = .031$, $p = .004$, such that the advantage for the standard condition was greater on the later blocks. The interaction of condition and the linear contrast across blocks was also significant, $F(1, 658) = 6.75$, $\eta^2 = .010$, $p = .010$. These results are consistent with the cognitive antidote principle and the hypothesis that task engagement decreases over the course of training especially when quizzes do not interrupt learning.

Discussion

The results for quiz performance in Experiment 2 were similar to those in Experiment 1. Notably, performance on the quiz was significantly better in the standard condition than in the postponed condition, presumably because of the shorter retention interval between studying and quizzing in the former condition.

The engagement survey results from Experiment 2 concord with the test performance results from Experiment 1, further supporting the cognitive antidote principle and providing direct evidence for a motivational mechanism underlying the effect. Specifically, participants' ratings of task engagement decreased across blocks of trials, and this decline was more marked in the postponed condition, where there were no interruptions between study blocks during training, than in the standard condition, where quizzes interrupted the series of study blocks during training. Although this finding more directly implicates a motivational explanation (i.e., a decline in effort, engagement, and motivation and an increase in boredom, mind wandering, and fatigue) than do the results of Experiment 1, the evidence is based solely on self-reported ratings. The pattern observed in these ratings might be due at least in part to demand characteristics of the experiment because participants might guess that the experimenter expected their engagement ratings to decline across blocks of trials. However, demand characteristics should be equivalent for the standard and postponed conditions, so the main effect of block (and the linear contrast across blocks) on engagement ratings could be explained on the basis of demand characteristics, but it would difficult to explain by demand characteristics the significant interaction of condition and block (or the interaction of condition and the linear contrast across blocks).

General Discussion

Both Experiments 1 and 2 showed an advantage for the standard over the postponed condition during training, which can be explained by the shorter retention interval between study and quiz in the standard than in the postponed condition. More important was the finding in Experiment 1 of an advantage for the standard over the postponed condition during testing, with that advantage increasing across blocks of trials and evident for unquizzed items as well as quizzed items. This advantage was explained in terms of the cognitive antidote principle (Kole et al., 2008), according to which the interspersed quizzes enhance motivation by reducing boredom and increasing task engagement. This explanation was supported by the results of the survey in Experiment 2, in which self-reported task engagement decreased across trial blocks to a larger extent for the postponed than for the standard condition. Our approach of analyzing test performance in Experiment 1 on previously unquizzed items avoids a number of problems that would otherwise confound interpretation of our results. In particular, the two conditions differed in the time elapsed between study and quizzing of each item. However, our primary analysis is restricted to test items that were not quizzed at all. Participants' experience with those unquizzed items was essentially identical in the two conditions. This sets the present study apart in an important way from previous studies examining similar manipulations (e.g., Szpunar et al., 2013; Weinstein et al., 2016; Wissman & Rawson, 2015), although unquizzed items have been examined in studies of the testing effect (see Cho, Neely, Crocco, & Vitrano, in press, for a recent example). In the previous studies with similar manipulations, the final tests covered the same material as the quizzes, and therefore the conditions differed not only in the overall schedule of studying and quizzing but also in the timeline of the participants' experience with the specific information tested. This previous design makes it difficult to disentangle effects due to motivation from effects due to mechanisms of learning and memory. According to research on the well-established spacing effect in learning (Cepeda et al., 2006), the postponed condition should yield better test performance because of the increased spacing between learning opportunities (i.e., between study and quiz). On the other hand, quiz performance will likely be superior in the standard condition (as it was in our study), which in turn could produce a difference in subsequent test performance between the conditions. Specifically, when participants are incorrect on a quiz, they receive a learning opportunity only through the feedback provided, whereas when they are correct on a quiz they receive learning opportunities both from their own response and from the feedback. This mechanism would predict superior test performance in the standard condition. Details aside, the important point is that considerations like these complicate interpretation of the data when the analysis of test performance is based on items that were part of the manipulation of quiz timing. None of these complications arise for unquizzed items in our design. The fact that our results showed an advantage of the standard condition even restricted to unquizzed items thus indicates that the effect of the experimental manipulation is due to general factors of cognitive or motivational state (e.g., reduced boredom and increased engagement) that are not specific to individual items.

The finding of the block-by-condition interaction is crucial for theoretical diagnosis of the mechanism underlying the advantage for interpolated quizzing. As noted, our interpretation of this advantage is that it represents a motivational effect, as captured by the cognitive antidote principle. However, there are a number of other possibilities that need to be considered. One possibility is that the advantage of the standard condition at test was mediated by confidence. Specifically, participants in the standard condition performed better on the quizzes than did participants in the postponed condition (recall that all participants were given feedback on their performance), and this difference should have led the former group to have greater confidence at the onset of the test in their ability to remember the material. Greater confidence could in turn benefit test performance by raising participants' expectations about their ability to succeed on the test. This explanation is similar to our cognitive antidote hypothesis in that both posit improved motivation from interspersed quizzes. However, the confidence explanation would not predict the observed block-by-condition interaction that was naturally explained by the cognitive antidote hypothesis. Because the testing of all blocks occurred after the study-quiz phase was complete, any difference between conditions in participants' level of confidence at the end of the study-quiz phase should affect all components of the test equally.

The block-by-condition interaction also provides some evidence against cognitive explanations for the advantage of the standard over the postponed condition. A number of past studies have demonstrated a mnemonic advantage from interpolated testing (Bäuml & Kliegl, 2013; Jang & Huber, 2008; Pastötter, Schicker, Niedernhuber, & Bäuml, 2011; Szpunar, McDermott, & Roediger, 2008). The explanations offered in these papers all come down to the proposal that interspersed quizzing increases contextual separation between lists or blocks, thereby facilitating retrieval at test. This cognitive mechanism cannot explain our observed interaction with block, because the benefit of contextual separation should be symmetric across the eight blocks (i.e., should be as great for the beginning blocks as for the ending blocks). Much of the work supporting the contextual separation hypothesis has focused on reductions in proactive interference (PI; Bäuml & Kliegl, 2013; Szpunar et al., 2008; Weinstein, Gilmore, Szpunar, & McDermott, 2014), which would indeed selectively benefit later blocks. However, contextual separation should produce an equal reduction in retroactive interference (RI) on the final test, and thus the overall prediction should be symmetric. Unlike the present design, the design of these previous studies did not permit a test of a block-by-condition interaction and, thus, effects on PI and RI could not be compared.

An alternative hypothesis that is consistent with the block by condition interaction is that, during the initial blocks of the quiz, participants learned about the type of question to expect (i.e., a question in which the participants must select the correct plant name for a given fact) and that in the standard condition, but not in the postponed condition, participants could use that information to study more effectively in later study blocks. This hypothesis would predict a specific interaction between condition and block in subjects' quiz performance, such that the difference between performance in the standard and postponed conditions should increase across blocks in the quiz, as it did in the test. Experiment 1 showed no interaction at all. Although there was a significant interaction of condition and block in quiz performance in Experiment 2, the form of the interaction did not conform to expectations based on this alternative hypothesis; the interaction of condition and the linear contrast across blocks was not significant in either Experiment 2 or Experiment 1. This alternative hypothesis is also weakened by the fact that at the beginning of both experiments participants were given a specific example of a question that would be quizzed, so all participants knew the question format before any studying began, regardless of their condition.

Another potential explanation, which is also consistent with the block-by-condition interaction, involves changes in students' metacognitive expectations about performance (e.g., Szpunar et al., 2014). Students who are initially overconfident in their expectations about their performance might become aware through quizzing that their level of performance does not meet their expectations and thereby increase their effort in subsequent study blocks (e.g., they might improve their method of encoding the subsequent information to be learned). This hypothesis would predict an advantage for the standard condition that is greater for later blocks. It resembles the cognitive antidote hypothesis because the change in effort that is entailed could be viewed as reflecting a change in motivation. However, the metacognitive calibration hypothesis differs from the cognitive antidote hypothesis because they rely on different mechanisms. The cognitive antidote principle postulates that increased motivation is due to increased engagement, or lack of boredom. In contrast, the calibration hypothesis is based on increased effort, following a discrepancy between expectations and performance. These two motivational mechanisms, which are not necessarily mutually exclusive, cannot be distinguished on the basis of either the test data in the present Experiment 1 or the earlier data provided by Szpunar et al. (2014), because although that earlier study found that interpolated quizzing improved test performance, it did not demonstrate any change in performance expectations due to interpolated quizzing. The survey data from the present Experiment 2 also do not discriminate between these two mechanisms; both mechanisms are consistent with the observed pattern of engagement as a function of condition and block. However, the pattern of results on the quizzes seems to favor the cognitive antidote explanation over the metacognitive calibration hypothesis because, just as in the case of the previous alternative hypothesis, the metacognitive calibration hypothesis should predict a specific condition by block interaction in quiz performance such that the difference between performance in the standard and postponed conditions should increase across blocks in the quiz because participants should learn from the quizzes and adjust their effort and study strategies accordingly. This specific interaction was not observed for quiz performance in either experiment.

The cognitive antidote principle might be viewed as a special case of the principle of desirable difficulties, whereby manipulations that increase the cognitive challenge during training can improve later test performance (e.g., Bjork, 1994). In the standard condition of the present experiments the interspersed quizzes could be viewed as a desirable difficulty because they introduce an additional cognitive requirement that interrupts study. On the other hand, quiz performance was worse in the postponed condition than in the standard condition (because of the greater delay), and thus one could think of the postponed quizzes as the desirable difficulty. Under this alternative interpretation, the theory of desirable difficulties would make the wrong prediction for the test results of Experiment 1. Because of these opposing applications, the theory of desirable difficulties does not provide guidance in analyzing our paradigm. Moreover, it does not yield predictions regarding the motivational measures in Experiment 2.

We believe the correct theoretical characterization is that the cognitive antidote is a separate type of mechanism from that of desirable difficulties. Indeed, McDaniel and his colleagues (McDaniel & Butler, 2011; McDaniel & Einstein, 2005) have shown that difficulties are desirable during learning only when they cause learners to apply task-relevant cognitive processes to the learning material (e.g., they might generate answers from fragments) that otherwise would not be engaged. In contrast, the cognitive antidote effect has been observed from the addition of task-irrelevant processes during learning (Kole et al., 2008). For example, in one experiment an alternating keystroke requirement was added to the task of entering four-digit numbers into the computer: Instead of typing a single constant keystroke to conclude every trial (serving an "enter" function), participants were required to alternate the concluding keystroke between the plus and minus keys. That simple but irrelevant requirement served to improve data-entry accuracy overall and to eliminate the decline in accuracy across trials, presumably due to its decreasing task disengagement or boredom. Therefore, the cognitive antidote appears to be a distinct mechanism from those described under the principle of desirable difficulties, in that it affects motivational processes rather than cognitive or learning processes.

Although our work to date on the cognitive antidote principle does not fully pin down the mechanisms involved, our findings do suggest a rich interplay among attention, motivation, and learning that has been largely neglected in modern cognitive psychology. The present manipulation involved only the timing of study and quizzing events, without any manipulation of reward or other overt motivational variables. Nevertheless, as evident from the survey results in Experiment 2, this manipulation of the learning schedule appears to have affected participants' level of engagement in the task, which in turn enabled those in the standard condition to learn more in the later blocks of study, as evident from the test results in Experiment 1. Although we view this theoretical approach as novel and promising, we have thus far measured motivational variables only using a self-report survey, and more objective measurement needs to be made in future research to verify more fully the interactive dynamics involved. For example, it might be informative to assess mind wandering with direct probes (e.g., Smallwood & Schooler, 2006; Szpunar et al., 2013) to investigate whether it serves as a mediator of the cognitive antidote effect, with the reduction in mind wandering responsible for increasing task engagement. We hope that continued work in this vein will lead to a mature theory about the interaction of learning and motivation in human knowledge acquisition.

This type of interactive dynamic between learning and motivation has important practical implications for educational settings. Timing of quizzes and other classroom activities, though they might seem inconsequential, can have a large impact on students' engagement in the classroom and, consequently, on their learning. More specifically, our findings can be directly translated into prescriptions for classroom practice using the clicker technique: Clicker questions should be used to interrupt a lecture, rather than saving them to the end. Our recent research has begun to test these issues in real classrooms (Ketels, Jones, Healy, & Martichuski, 2013).

References

Anderson, L. S., Healy, A. F., Kole, J. A., & Bourne, L. E., Jr. (2013). The clicker technique: Cultivating efficient teaching and successful learning. *Applied Cognitive Psychology, 27,* 222-234. doi: 10.1002/acp.2899

Bäuml, K.-H. T., & Kliegl, O. (2013). The critical role of retrieval processes in release from proactive interference. *Journal of Memory and Language, 68,* 39-53. doi: 10.1016/j.jml.2012.07.006

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge, MA: MIT Press.

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

Cho, K. W., Neely, J. H., Crocco, S., & Vitrano, D. (in press). Testing enhances both encoding and retrieval for both tested and untested items. *Quarterly Journal of Experimental Psychology*. Advance online publication. doi: 10.1080/17470218.2016.1175485

Jang, Y. & Huber, E. (2008). Context change in free recall: Recalling long-term memory drives list isolation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34, 112-127.* doi: 10.1037/0278-7393.34.1.112

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

Karpicke, J. D., & Roediger, H. L., III (2008). The critical importance of retrieval for learning. *Science, 319,* 966-968. doi: 10.1126/science.1152408

Ketels, S., Jones, M., Healy, A. F., & Martichuski, D. (2013, November). *When should clicker questions be presented during a lecture? Effects on exam performance*. Poster presented at the 54th Annual Meeting of the Psychonomic Society, Toronto, Canada.

Kole, J. A., Healy, A. F., & Bourne, L. E., Jr. (2008). Cognitive complications moderate the speed-accuracy tradeoff in data entry: A cognitive antidote to inhibition. *Applied Cognitive Psychology, 22*, 917-937. doi: 10.1002/acp.1401

McDaniel, M. A., & Butler, A. C. (2011). In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork* (pp. 175-198). New York: Psychology Press.

McDaniel, M. A., & Einstein, G. O. (2005). Material appropriate difficulty: A framework for determining when difficulty is desirable for improving learning. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications* (pp. 73-85). Washington, DC: American Psychological Association. doi: 10.1037/10895-006

Pan, S. C., Gopal, A., & Rickard, T. C. (2016). Testing with feedback yields potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology, 108,* 563-575. doi: 10.1037/edu0000074

Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K.-H. T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 287-297. doi: 10.1037/a0021801

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

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

Rotgans, J. I., & Schmidt, H. G. (2011). Cognitive engagement in the problem-based learning classroom. *Advances in Health Sciences Education, 16,* 465-479. doi: 10.1007/s10459-011-9272-9

Smallwood, J., & Schooler, J. W. (2006). The restless mind. *Psychological Bulletin, 132,* 946-958. doi: 10.1037/0033-2909.132.6.946

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science, 323,* 122-124. doi: 10.1126/science.1165919

Szpunar, K. K., Jing, H. G., & Schacter, D. L. (2014). Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education. *Journal of Applied Research in Memory and Cognition, 3,* 161-164. doi: 10.1016/j.jarmac.2014.02.001

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

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

Thios, S. J., & D'Agostino, P. R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning and Verbal Behavior, 15,* 529-536. doi: 10.1016/0022-5371(76)90047-5

Tzeng, O. J. (1973). Stimulus meaningfulness, encoding variability, and the spacing effect. *Journal of Experimental Psychology, 99,* 162-166. doi: 10.1037/h0034642

Weinstein, Y., Gilmore, A. W., Szpunar, K. K., & McDermott, K. B. (2014). The role of test expectancy in the build-up of proactive interference in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1039-1048. doi: 10.1037/a0036164

Weinstein, Y., Nunes, L., & Karpicke, J. D. (2016). On the placement of practice questions during study. *Journal of Experimental Psychology: Applied, 22,* 72-84. doi: 10.1037/xap0000071

Wissman, K. T., & Rawson, K. (2015). Grain size of recall practice for lengthy text material: Fragile and mysterious effects on memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 439-455. doi: 10.1016/j.jml.2015.05.003

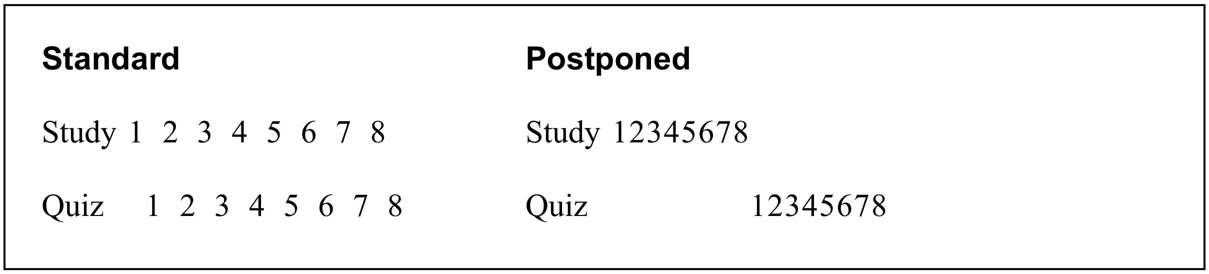

Figure 1. Ordering of study blocks and quizzes during the training phase in each experimental condition. Each number (1-8) refers to a different category of eight facts.

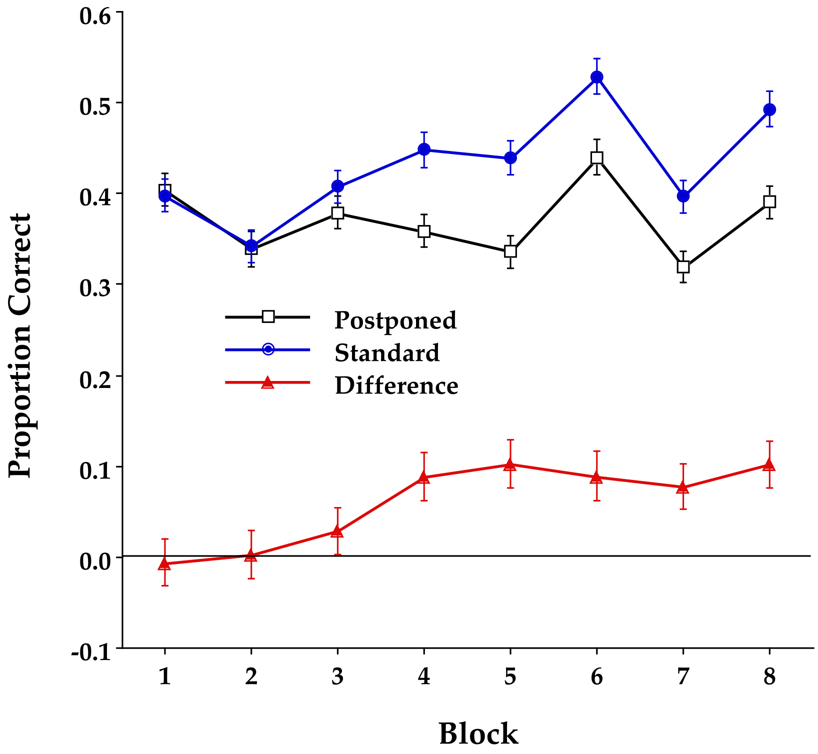

Figure 2. Proportion of correct responses at test as a function of block and condition, in Experiment 1. The difference between the two conditions at each block is also shown. Error bars indicate standard errors of the mean.

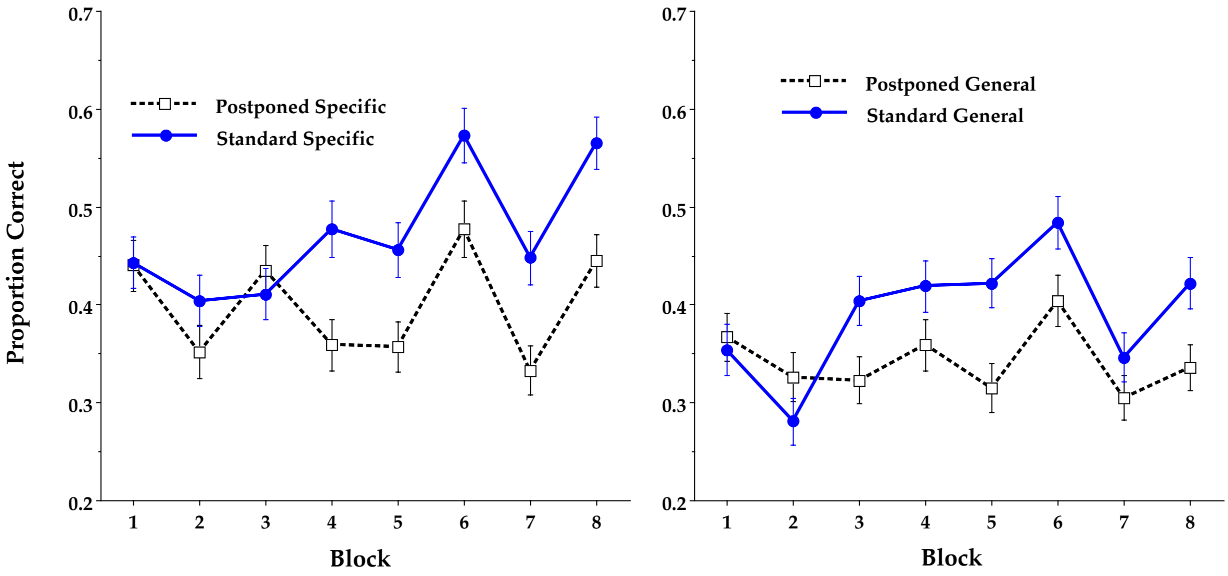

Figure 3. Proportion of correct responses at test as a function of block, question form, and condition, in Experiment 1. Error bars indicate standard errors of the mean.

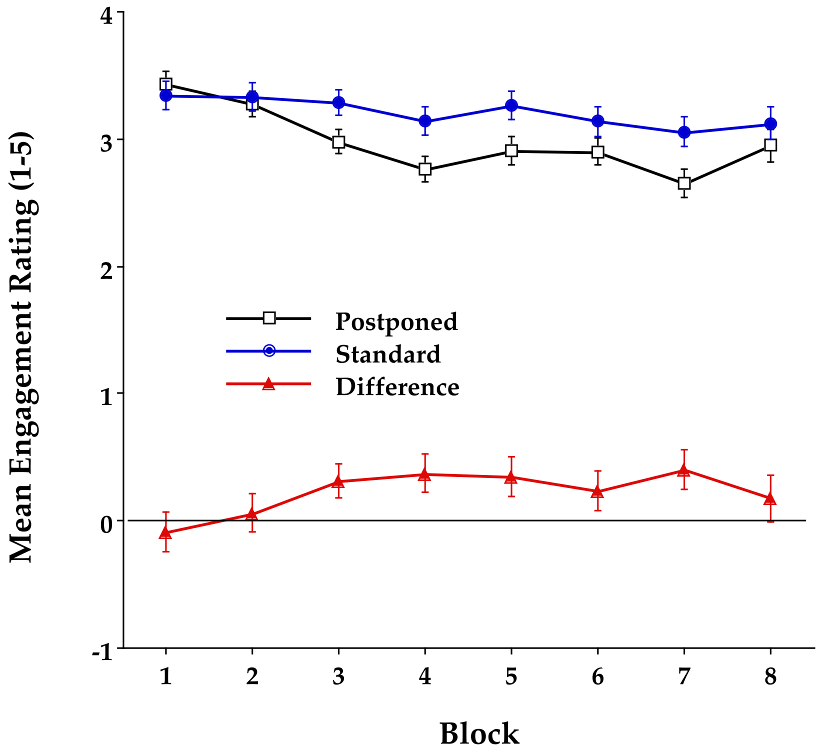

Figure 4. Mean engagement rating as a function of block and condition, in Experiment 2. The difference between the two conditions at each block is also shown. Error bars indicate standard errors of the mean.