On the Placement of Practice Questions During Study

Yana Weinstein[1], Ludmila D. Nunes [2,3], and Jeffrey D. Karpicke[2]

[1]University of Massachusetts – Lowell

[2]Purdue University

[3]University of Lisbon

Journal of Experimental Psychology: Applied, in press

Author Note

Yana Weinstein, University of Massachusetts – Lowell; Ludmila D. Nunes, Department of Psychological Sciences, Purdue University and University of Lisbon; and Jeffrey D. Karpicke, Department of Psychological Sciences, Purdue University.

Address correspondence to: Yana Weinstein, PhD, Assistant Professor, UML Psychology Department, 113 Wilder Street, Suite 375, Lowell, MA 01854-3059, Tel: 1-978-934-3917

Acknowledgments

Andrea Dottolo, Mary Duell, Alice Frye, and Richard Siegel produced the materials used in Experiments 1 and 2. Rishi Vangapalli and Sean McCaffery scored Experiment 1 data. Fabian Jones processed and Kelsey Gilbert scored Experiment 2 data. Mary Ejaife and Gabriele Bard scored Experiment 3 data. Ludmila D. Nunes was, in part, supported by the Portuguese Foundation for Science and Technology Fellowship SFRH / BPD / 86253 / 2012.

Abstract

Retrieval practice improves retention of information on later tests. A question remains: when should retrieval occur during learning – interspersed throughout study, or at the end of each study period? In a lab experiment, an online experiment, and a classroom study, we aimed to determine the ideal placement (interspersed vs. at-the-end) of retrieval practice questions. In the lab experiment, 64 subjects viewed slides about APA style and answered short-answer practice questions about the content or restudied the slides (restudy condition). The practice questions either appeared one every 1-2 slides (interspersed condition), or all at the end of the presentation (at-the-end condition). One week later, subjects returned and answered the same questions on a final test. In the online experiment, 175 subjects completed the same procedure. In the classroom study, 62 undergraduate students took quizzes as part of class lectures. Short-answer practice questions appeared either throughout the lectures (interspersed condition) or at the end of the lectures (at-the-end condition). Nineteen days after the last quiz, students were given a surprise final test. Results from the three experiments converge in demonstrating an advantage for interspersing practice questions on the initial tests, but an absence of this advantage on the final test.

Keywords: testing effect, retrieval practice, massing, spacing, interleaving

The benefits of retrieval practice, also known as quizzing or testing, were first empirically demonstrated in the early 20th century (e.g., Abbott, 1909). Over the past few decades, cognitive psychologists have resumed interest in this phenomenon (e.g., Roediger & Karpicke, 2006). Retrieval practice, as compared with re-studying materials, has been shown to produce better long-term retention in the lab (e.g., Darley & Murdock, 1971) and in the classroom, including in university (e.g., McDaniel, Anderson, Derbish, & Morrisette, 2007) and middle-school settings (e.g., McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011). Taken together, these studies suggest that retrieval opportunities should be introduced into the classroom to aid learning. Some educators have adopted this technique, though mostly in Science, Engineering, Technology, and Math (STEM) disciplines such as Biology or Chemistry (e.g., Kay & LeSage, 2009).

Many studies have been dedicated to exploring theoretical hypotheses about the mechanisms behind the testing effect (see Rowland, 2014, for a meta-analysis and Karpicke, Lehman, & Aue, 2014, for a review) and trying to convince educators and students that they should take advantage of this effect (e.g., Mayer et al., 2009; Roediger, Agarwal, McDaniel, & McDermott, 2011). Meanwhile, fewer studies have been dedicated to determining the ideal placement of practice questions during learning. In classroom studies, retrieval practice typically occurs at the end of each study session (e.g., at the end of a class as in McDaniel et al., 2011). However, recent research using an "interpolated testing" paradigm (Szpunar, McDermott, & Roediger, 2008) suggests that interspersing quiz questions throughout learning may maintain encoding (Pastötter, Schicker, Niedernhuber, & Bäuml, 2011), possibly by boosting test expectancy (Weinstein, Gilmore, Szpunar, & McDermott, 2014).

Wissman and Rawson (2015) compared the mnemonic effects of recalling a text divided into sections (recall after each small text section) to a single recall opportunity of the full text. In their seven laboratory experiments, subjects studied prose passages and were either tested after every section ("small grain size"), or on the whole passage ("large grain size"). Those tested on every section did better initially, but on the final test, the groups performed similarly. Wissman and Rawson consistently showed that recall of the small sections was more successful initially than recall of the full text, but on a final free recall test (delayed by 20 min or 2 days), there were no recall differences between the small and large grain size conditions.

Wissman and Rawson's results provide initial evidence that freely recalling during reading may promote the same amount of learning as recalling after reading an entire prose passage. Interspersed retrieval with short answer quiz questions, however, has never been directly compared to the typical at-the-end quizzing that is usually employed in classroom settings. The present paper reports three experiments that directly compared interspersed and at-the-end short-answer question conditions in laboratory and classroom settings.

Which of the two quiz placement conditions – interspersed or at-the-end – did we expect to be the most effective, both in terms of initial and final test performance? Due to the proximity of the interspersed practice questions to the studied information, we expected initial performance to be higher in the interspersed condition (e.g., Carpenter & DeLosh, 2005; Karpicke & Roediger, 2007). However, the prediction was less clear for the later test. One might argue that interspersed retrieval should promote greater retrieval-based learning due to greater initial retrieval success. On the other hand, if retrieval success is equivalent in interspersed and at-the-end conditions, then questions at the end of the materials might afford more effortful or difficult retrieval, thereby leading to better retention than the relatively easier retrieval that occurs for

interspersed quiz questions answered immediately after study (see Bjork, 1994, 2013; Pyc & Rawson, 2009). Wissman and Rawson (2015) originally set out to test the "grain size hypothesis", which generated the prediction that interspersed retrieval should foster greater learning due to increased initial retrieval success relative to at-the-end testing. No evidence for this hypothesis was found in their studies.

The difference between interspersed and at-the-end questioning resembles the difference between massed and spaced retrieval practice (Cepeda, Vul, Pashler, Wixted, & Rohrer, 2006; Dempster, 1987). Interspersed questions occur shortly after material has been presented. Retrieval success is likely to be high on these questions, but they may not promote long-term learning for the same reasons that massed practice fails to do so. When initial retrieval occurs immediately after an item is studied, retrieval is unlikely to involve much effort or difficulty, and the context between study and initial retrieval will have changed very little (Delaney, Verkoeijen, & Spirgal, 2010; Karpicke, Lehman, & Aue, 2014; Siegel & Kahana, 2014). Consequently, massed retrieval practice often produces little or no benefit for subsequent retention (e.g., Carpenter & DeLosh, 2005; Karpicke & Roediger, 2007).

Because our at-the-end questions occur after the entire set of material has been studied, on the other hand, the spacing or lag between the presentation of the material and the questions is greater than it is for interspersed questions. In this case, retrieval will be more effortful and difficult, and the context between study and initial retrieval will have shifted and changed over time (Delaney et al., 2010; Karpicke et al., 2014; Siegel & Kahana, 2014). Thus, while levels of retrieval success are likely to be lower because of greater initial spacing, spaced retrieval produces greater long-term retention than does massed retrieval. In the present experiments, if retrieval success during practice is more or less equivalent, at-the-end practice questions may

lead to better long-term retention because these questions are spaced relative to when information was originally presented. However, if interleaving the questions throughout study promotes more initial retrieval success than presenting the questions at the end of study, then the benefits of spacing and initial retrieval success might cancel each other out. In this case, interspersed and at-the-end practice questions would equally benefit learning, which would be similar to Wissman and Rawson's (2015) findings with free recall of texts.

Experiment 1

Method

Subjects. The sample consisted of 64 General Psychology students at the University of Massachusetts Lowell who participated in the experiment for course credit. Subjects were recruited through Sona Systems and participated in two in-person sessions exactly 1 week apart. Subjects were randomly assigned to one of three between-subjects question placement conditions, resulting in 22 subjects in the interspersed condition, 20 subjects in the at-the-end condition, and 22 subjects in the none (restudy) condition. An additional 13 subjects started the experiment but were lost due to attrition or programming error. These subjects belonged equally to each of the three conditions, with 4 subjects each lost from the interspersed and at-the-end conditions, and 5 subjects lost from the restudy condition. Also, performance during the learning phase for the 8 lost subjects from the interspersed and at-the end conditions was almost identical to performance for the 42 retained subjects from those conditions (.77 vs. 78 respectively; performance during the learning test for restudy subjects could not be compared since there were no tests during learning in that condition).

Design. We used a between-subjects design with question placement (interspersed vs. at-the-end vs. none) as the independent variable. The dependent measures of interest were

performance on 10 short-answer practice questions during the learning phase (interspersed and at-the-end groups only), and performance on the same 10 short-answer questions on the delayed test for all groups, measured as a percentage of the total possible score.

Procedure. The experiment consisted of a learning phase and a delayed phase test. Prior to the learning phase, subjects were asked to indicate how well they would say they knew APA style (5-point Likert scale from "Not at all" to "Very well"); how much APA style they had covered in their class (5-point Likert scale from "None" to "Very much"); whether they had had to write any assignments in APA style for a class (yes/no); and to describe their experience with APA style in an open-response question. Subjects were told that they would read some information about APA style, and that at some point they may be tested on the information presented in the slides (but they were not specifically told about the delayed test). Subjects were told that they did not need to know anything about APA style to participate in the experiment.

In the learning phase, subjects viewed one of two Powerpoint presentations, relating to either References or In-Text Citations (exactly half of the subjects in each condition viewed one, and the other half viewed the other). These slides had been developed by University of Massachusetts Lowell Psychology faculty as an internal resource for students in the major. Within the presentation, each slide was presented on the screen for 45 seconds. An example slide is shown in Appendix 1.

Depending on condition, subjects either answered 10 short-answer practice questions interspersed throughout the 18 slides, or answered the same 10 practice questions presented at the end of the 18 slides (please see Appendices 2 and 3 for the exact question placement in each of the two presentations), or did not answer any practice questions but instead had a chance to restudy the presentation immediately after initial encoding.

In the interspersed condition, practice questions always appeared immediately after the slide that contained the relevant information. Each practice question was presented on screen for 45 seconds, during which time subjects could attempt to respond. Immediately afterwards, feedback in the form of the correct answer was presented for 15 seconds.

In the at-the-end condition, practice questions all appeared one after the other with one question per screen, after all slides of the presentation had been displayed. The timings were the same as for the interspersed condition, with feedback in the form of the correct answer presented immediately after each question.

In the restudy condition, subjects first viewed the entire presentation with fixed timings (45 seconds per slide), and then were able to freely review the slides immediately after the presentation finished, for the same amount of time that it took subjects in the interspersed and at-the-end conditions to answer the 10 practice questions. The slides were presented in one webpage so subjects could scroll up or down through the entire presentation at their own pace. Since the experiment took place in the lab, subjects remained on the page with the slides for the duration of the restudy time. The total time of the learning phase including slide presentation and retrieval practice was approximately 25 minutes in all three conditions.

Subjects returned one week later to answer the same 10 questions that those in the interspersed and at-the-end conditions had answered with feedback during the learning phase. On the delayed test, no feedback was provided and subjects again had 45 seconds to attempt each question. Test questions were presented in the same order as they had appeared for retrieval practice in the learning phase. For both the learning phase and the delayed test, subjects typed in their response in a text box underneath the question. In addition to completing the questions analyzed here, subjects also answered questions on the slides that they had not studied (i.e., on

the in-text citation slides if they had studied the references slides, or vice versa) as an assessment of their baseline knowledge of APA style. These baseline data were collected for departmental purposes and were not of interest to our study. The order of the in-text citation and references quizzes was randomized, such that sometimes subjects answered the criterion test for our study before this baseline test, and sometimes after. In addition, at the start of session two, subjects were asked if they had had any additional exposure to APA style outside of the experiment (e.g., in a class), and after the final delayed test, subjects completed an exercise in which they looked for APA style errors in a mock paper, also not of interest to our study.

Results

Scoring was performed by two research assistants blind to the condition assignment. A scoring rubric was used for grading both the learning phase and delayed test responses (see example in Appendix 1). Scores on each set of slides could range from 0 to 18 points with 1-3 possible points to be gained per question, and were converted to percentages. Inter-rater reliability was calculated separately for each of the two phases by correlating the scores of the two raters. The *Pearson's r* was .86 ($p = .002$) for the learning phase and .94 ($p < .001$) in the delayed test, showing reasonable to excellent agreement between the two raters. Scores given by the two raters were averaged for the purpose of analyses. The presentation about in-text citations resulted in better performance than the presentation about references ($M = .84$, $SD = .10$ and $M = .72$, $SD = .17$ respectively in the first session; and $M = .73$, $SD = .15$ and $M = .61$, $SD = .17$ respectively in the second session), but this variable did not interact with question placement[1] and thus was not included in the analyses.

[1] $F(1, 38) = 1.03$, partial $\eta^2 = .03$ for the first session and $F(1, 58) = 0.88$, partial $\eta^2 = .03$ for the second session for the interaction between materials and question placement.

Figure 1 presents the accuracy data for the learning phase in the interspersed and at-the-end conditions, and for the delayed test in all three conditions. The figure demonstrates a clear separation between the two testing conditions during the learning phase, with subjects performing considerably better in the interspersed condition (M = .84) than in the at-the-end condition, (M = .71), $t(40) = 2.95$, $p < .05$,[2] d = 0.91 [0.27, 1.54]. On the delayed test, however, there was no such difference between question placement conditions, but both conditions outperformed the restudy condition. There was a somewhat unreliable advantage of the interspersed condition (M = .70) relative to the restudy condition (M = .61), $t(42) = 1.79$, p = .08, d = 0.52 [-0.08, 1.12]. There was also a similar advantage of at-the-end condition (M = .70) compared to restudy, $t(42) = 1.82$, p = .08, d = 0.55 [-0.06, 1.15]. The difference between the interspersed and at-the-end conditions, on the other hand, was close to zero, $t(40) = 0.196$, p = .85, d = 0.06 [-0.55, 0.67].

To check for the effect of prior knowledge, experience with APA style – which subjects described in an open-response question – was coded on a scale from 0 to 3. One point was awarded for each of the following, which were the most common experiences subjects listed: having written a paper in APA style, studying APA style in class, and looking up information about APA style online. Data from the binary and Likert-scale questions (knowledge of APA style, how much APA style had been covered in class, and whether subjects had written a paper in APA style) were not included in this variable because they were redundant with the descriptions subjects provided in the open-ended question, e.g.: “I had to write a paper for an Anthropology class. It's the only time I've had to use it. We went over it a little bit in my current psychology class and I've read about it a little on Purdue Owl (in text citation, basic layout, cited

[2] We have reported p-values throughout the manuscript for informational purposes, but our interpretations are based on confidence intervals.

works, etc.)" The data were reanalyzed with experience of APA style as a covariate, but this variable had no effect on performance and did not alter the effect of question placement.

Discussion

In the first experiment, we demonstrated in the lab that interleaving questions throughout study produced better initial performance compared to answering questions at the end of the material, but that no such advantage remained after one week. Subjects in the interspersed condition were able to answer questions more accurately during study because of the close proximity of the practice questions to the presented material, since practice questions in this condition appeared immediately after relevant information was presented. In the at-the-end condition, on the other hand, subjects had to retrieve information that they had studied up to 10 minutes earlier, leading to poorer performance during the learning phase. One week later, however, it appeared that subjects in the at-the-end condition suffered from no forgetting of information from the initial learning phase to the delayed test, whereas subjects in the interspersed condition lost the entire advantage that they demonstrated during the learning phase. Both conditions did, however, perform better than a restudy control – demonstrating a typical testing effect (Roediger & Karpicke, 2006).

These results are consistent with Wissman and Rawson's (2015) findings with free recall of texts and extend the results to short answer quiz questions. The results also expand upon their initial findings by examining the effects relative to a baseline control condition that did not perform any retrieval activity. As expected, the interspersed and at-the-end retrieval activities resulted in better performance than restudying, consistent with a wealth of prior work on retrieval practice effects (e.g., Abbott, 1909; Darley & Murdock, 1971; McDaniel et al., 2007; McDaniel et al., 2011; and Roediger & Karpicke, 2006). However, the small sample sizes in

Experiment 1 left us with quite a bit of uncertainty about these effects. Thus, Experiment 2 was a replication of Experiment 1 with a substantially larger sample size.

Experiment 2

In Experiment 2 we attempted to replicate the results with a larger sample size in order to increase power and the precision of our effect size estimates. In addition, we wanted to extend our results from the lab to an online sample with greater diversity of experiences with APA. The procedure and analyses were very similar to those of Experiment 1, with mainly the environment, population, and sample size differing between the two experiments.

Method

Subjects. Subjects were recruited online via a Human Intelligence Task (HIT) posted on Amazon Mechanical Turk. Subjects were restricted to people who were located in the United States, had a 95% HIT acceptance rate, and had completed at least 1000 HITs. The sample consisted of 175 subjects who completed the experiment for financial compensation. Subjects participated in two online sessions 1 week apart. They received $2 for completing session one, which lasted approximately 25 minutes, and $2 for completing session two, which lasted approximately 8 minutes. Subjects were randomly assigned at the beginning of the first session to one of three between-subjects question placement conditions, resulting in 69 subjects in the interspersed condition (34 studying the slides about in-text citations and 35 studying the slides about references), 57 subjects in the at-the-end condition (26 studying the in-text citation slides and 31 studying the references slides), and 49 subjects in the none (restudy) condition (26 studying the in-text citation slides and 23 studying the references slides). Fifty-three additional subjects completed session 1 but did not return for session 2. These subjects were evenly distributed with respect to question placement conditions (19 interspersed, 18 at-the-end, and 16

restudy subjects) and presentation topic (24 studied the in-text citations presentation and 29 studied the references presentation). Comparisons of the non-returners' data from the learning phase to those subjects who completed the study are presented below in a footnote. Also, two subjects who completed both sessions were excluded because there was a malfunction in the automated slide-timing mechanism of the online study. Our final sample included 82 females, 1 genderqueer person, and 92 males. Six subjects indicated that they had a Psychology degree. The age range of our final sample was 20 to 63 years (M = 36.8, SD = 10.5).

Design and Procedure. The design was exactly the same as that of Experiment 1. The procedure was also very similar, with the following changes. Subjects completed the experiment entirely online at the time and location of their choosing, with the constraint that the second session had to take place one week after the first session. All subjects were told that there would be a second session while signing up for the first. Subjects also received an email via the Amazon Mechanical Turk system inviting them to complete the second session when the second session HIT was posted. Demographic questions (age, open-ended gender, and domain of education) were answered at the start of the survey. The questions about APA style were altered from Experiment 1 to account for the fact that subjects were not necessarily university students; question asked how well subjects knew APA style (Likert scale); how much APA style they had covered in their education (Likert scale); and whether they had ever had to write a paper in APA style (yes/no), as well as the open-ended question. Additionally – since we had less control online than in the lab with respect to subjects' ability to engage in other activities while completing the experiment – we asked subjects in the restudy condition to report how they had spent the restudy time. That is, at the end of the restudy phase subjects indicated whether they had spent the whole time studying the slides, had studied the slides as well as doing something

else, or had not studied the slides at all (all retained subjects in the restudy condition chose the first or second option). As in Experiment 1, at the start of session 2, we asked subjects if they had had any additional exposure to APA style outside of the experiment (e.g., in a class). Finally, at the end of session 1, we also asked subjects whether they had taken any notes.

Results

Scoring was performed by one research assistant blind to the condition assignment, using the same scoring rubric as for Experiment 1. One hundred of the 175 subjects (57%) had no experience with APA style prior to the experiment, whereas the other subjects had a variety of experience ranging from coming across it in one class, to using it extensively in a graduate psychology degree. Thus, experience with APA style was coded as a binary variable (0 or 1). However, an initial analysis suggested that performance in the first and second session was not related to prior experience with APA style, and this variable did not interact with any of the effects presented below, so it was not included in the main analyses. As in Experiment 1, the presentation about in-text citations resulted in better performance than the references presentation (M = .82, SD = .15 and M = .77, SD = .20 respectively in the first session; and M = .72, SD = .20 and M = .67, SD = .20 respectively in the second session), but this variable once again did not interact with question placement,[3] and thus was not included in the analyses.

Figure 2 presents the accuracy data for the learning phase in the interspersed and at-the-end conditions, and for the delayed test in all three conditions. The data exactly replicate the patterns observed in Experiment 1. As in Experiment 1, in the learning phase, there was a significant difference between question placement conditions with interspersed questions

[3] $F(1, 122) = 1.63$, partial $\eta^2 = .01$ for the first session and $F(1, 169) = 1.02$, partial $\eta^2 = .01$ for the second session for the interaction between materials and question placement.

producing better performance, $t(108.2) = 5.65$ (corrected for unequal variances), $p < .05$, $d = 1.01$ [0.64, 1.38].[4] On the delayed test, exactly as in Experiment 1, there was an advantage of the interspersed condition ($M = .73$) relative to the restudy condition ($M = .57$), $t(116) = 4.00$, $p < .05$, $d = 0.75$ [0.37, 1.12]; and also an advantage of at-the-end condition ($M = .75$) compared to restudy, $t(85.2) = 4.62$, $d = 0.90$, $p < .05$, [0.50, 1.30]. Most importantly, there was no advantage of interspersing over the at-the-end condition on the delayed test, $t(124) = 0.77$, $p = .44$, $d = 0.14$ [-0.21, 0.49]. All the above analyses were also performed excluding the 18 subjects who took notes during study and/or had additional exposure to APA between the first and second session, and all patterns and conclusions remained unchanged.

Discussion

In Experiment 2 we directly replicated the results of Experiment 1, this time in a much larger online sample. As in Experiment 1, subjects initially performed better during learning when questions were interspersed during study rather than massed, but then performed at the same level at a one-week delay. In this experiment we also obtained the expected effects of retrieval practice on long-term retention in both quizzing conditions compared to the restudy control condition. The larger sample size in Experiment 2 reduced the widths of the confidence intervals from 1.2 in Experiment 1 to under 0.8 in Experiment 2, improving the precision of the

---

[4] In an attrition analysis, we included the 37 subjects in the interspersed and at-the-end conditions who did not return to complete the delayed test in a 2 x 2 ANOVA with question placement and experiment completion as between-subjects variables. Those who did not return for the delayed test performed worse in the learning phase than those who did complete the study, $F(1, 159) = 4.03$, partial $\eta^2 = .03$, but completion did not interact with question placement, $F(1, 159) < 1$.

effect size estimation by about 33%. Finally, in this experiment we were able to confirm that the effect could be generalized to a sample that had a greater variety of experience with APA style.

Experiment 3

In Experiment 3 we attempted to extend the findings of the previous two experiments to the classroom. In addition to the change from lab to classroom, Experiment 3 also differed from Experiments 1 and 2 in a number of ways: Materials used in the learning phase were live lectures instead of a presentation that subjects read on their own; feedback was provided after a delay instead of immediately; question placement was manipulated within-subjects instead of between-subjects; and the delay between learning phase and delayed test was on average 41 days. These additional differences allowed us to not only observe a replication of the results we found in Experiments 1 and 2, but also to determine whether the effect was robust enough to withstand methodological differences.

Method

Subjects. Subjects were drawn from a population of 79 undergraduate students enrolled in two sections of a 200-level Cognitive Psychology class at the University of Massachusetts Lowell, both taught by the first author. Two students withdrew from the class prior to the delayed test and were excluded from the study. One subject was not naïve to the design due to being a research assistant in the first author's lab, and was excluded from the analyses for this reason. Finally, only students who attended all four lectures in each condition (interspersed and at-the-end) as well as the delayed test were included in the analyses; this criterion excluded an additional 31 students. The final sample consisted of 45 students; 26 enrolled in section 1, and 19 in section 2 of the course.

Design and Procedure. We used a within-subjects design with question placement (interspersed vs. at-the-end) as the independent variable. The dependent measures of interest were performance on 20 short-answer questions per question placement condition across 4 lectures during the learning phase, and performance on the same 20 short-answer questions during the delayed test, measured as a percentage of the total possible score.

In each of 8 different lectures, students answered 5 questions that occurred either at the end of the lecture presentation (at-the-end condition), or interspersed throughout the presentation (interspersed condition). On any given day of class, one section of the class received 5 questions interspersed throughout the lecture, whereas the other section received all 5 questions at the end of the lecture. To avoid students gleaning a pattern and anticipating question placement from lecture to lecture, the counterbalancing was performed in a quasi-random order, as described in Table 1, which also includes information on the retention intervals between each lecture and the final delayed test.

Regardless of whether questions were interspersed or presented at the end of the lecture, all questions appeared as a PowerPoint slide within the lecture. Questions were identified by the slide having a red background. Each question appeared on the screen for 90 seconds; at the end of this time period, either the lecture continued for the interspersed condition, or the next question was presented for the at-the-end condition. In the at-the-end condition, to alert students to the next question in case they were looking away from the screen, the lecturer announced "next question" when the slide transitioned to each subsequent question after the first.

Questions tested only material that had been presented in the given lecture. When questions were interspersed, they appeared directly following the slide or slides that described the information necessary for answering the practice question. There were no instructions to

students regarding the use of notes to help answer the questions. However, in most cases, questions were written in such a way that the information from preceding slides had to be applied to a novel situation, so that merely noting down the information from preceding slides would not directly translate to an ideal answer on any question. Questions were short-answer format, designed to be answered in one or two sentences (see Appendix 4 for sample question and rubric). The two exceptions to this were one question that involved matching (Quiz 1, Question 1), and one question that involved identifying the order of various items (Quiz 8, Question 1). Subjects wrote their responses to the questions on paper, and these papers were collected at the end of each lecture. In the class immediately following each lecture (either 2 days later or 5 days later, depending on whether the quiz had been given on Tuesday or Thursday respectively), students received their graded responses along with the rubric that was used for grading purposes.

Approximately 10 weeks into the semester, students were presented with a surprise test for extra credit. The test was presented to students as a way of identifying which topics they needed to focus on to prepare for the final. This test itself was not a course requirement listed on the syllabus, thus any points students achieved on this surprise test were added to their course grades as extra credit. The delayed test questions were presented on paper, with all questions identical to those encountered originally in the lectures, listed in the order that they were presented, with lecture name headings. The delayed test consisted of a total of 50 questions, of which 40 questions (5 per lecture) were of interest to the current study and 10 questions were from 2 lectures that were not included in this study. Students had approximately 1h 10 minutes to attempt all 50 questions.

Results

Scoring was performed by two research assistants blind to the condition assignment and lecture content; only the rubrics were used for grading. Questions were scored such that a given response could either get 0 points, 0.5 points (minimal criteria met), or 1 point (ideal answer), and these scores were converted to percentages. Inter-rater reliability was calculated by taking an average of scores given across all questions by each of the two raters for each student, and these average scores for each student were correlated between the two raters, separately for the learning phase questions and the delayed test questions. The *Pearson's r* was = .85 ($p < .001$) for the learning phase and .95 ($p < .001$) for the delayed test, showing reasonable to excellent agreement between the two raters.

Mean score by question placement (interspersed vs. at-the-end) was calculated for each student both for the learning phase by taking the mean of the two raters' scores across the four lectures in each of the two question placement conditions, and for the delayed test by computing the mean of the same questions that pertained to each question placement condition on the delayed test. Figure 2 presents the data for performance in the learning phase (lectures) and in the delayed test in the interspersed and at-the-end conditions. The figure demonstrates that students performed significantly better on the lecture questions when they were interspersed throughout the lecture, than when they were all presented at the end of the lecture. On the other hand, there was no difference between the interspersed and at-the-end conditions on the delayed test, although both performed much worse than in the learning phase. In the learning phase, there was a reliable difference between question placement conditions with interspersed questions producing better performance, $t(44) = 4.62$, $p < .05$, $d = 0.69$ [0.36, 1.01]. On the delayed test, there was no such advantage for interleaving $t(44) = .40$, $d = 0.06$, $p = .69$, [-0.23, 0.35]. These analyses were all repeated without excluding the 24 students who had missed any of the lectures.

No differences were observed in this analysis compared to the reduced sample, and performance of the students who missed at least one lecture did not differ reliably from those who had attended all lectures.

Discussion

In this classroom experiment, students attended four lectures in each of two conditions: interspersed, where 5 questions appeared immediately after studied material throughout the lecture; and at-the-end, where those same 5 questions appeared at the end of the lecture. Results showed that in the learning phase, the interspersed condition produced superior performance to the at-the-end condition. However, on the delayed test taken 19 days after the last lecture, there was no such difference between conditions. These results complement those of Experiments 1 and 2, where in a lab study and in an online study respectively, an initial advantage of interspersing was found along with no such advantage one week later. Contrary to Experiments 1 and 2, however, in Experiment 3 both groups performed much worse on the delayed test than during learning, indicating that both might have suffered a great deal of forgetting from the initial learning phase to the final test. We carried out this experiment using a very realistic set-up, or what Dunlosky, Bottiroli, and Hartwig (2009) call a "highly representative design" (That is, we performed the experiment with students in a real classroom responding to quizzes that counted towards their grade. At the same time, a repeated-measures, counterbalanced design was implemented in order to ensure experimental controls. Where possible, all other variables were held constant between conditions, including question presentation time and content of the lecture slides, striking a balance between representative design and experimental control.

General Discussion

In one lab study (Experiment 1), one online study (Experiment 2), and one classroom study (Experiment 3), we compared the relative mnemonic benefits of retrieval practice with interspersed versus at-the-end practice questions. That is, in all three experiments, in one condition subjects answered questions that appeared throughout the learning phase immediately after the relevant information was studied (interspersed condition), and in the other condition subjects answered questions that appeared at the end of the learning phase (at-the-end condition). Subsequently, both groups were tested with the same questions after a delay. In all three experiments, we found that performance during the learning phase was higher in the interspersed condition than in the at-the-end condition. The obvious explanation for this effect is the proximity of the interspersed practice questions to the studied information, with a smaller retention interval reducing forgetting (e.g., Slamecka & McElree, 1983). Another potential benefit of interspersing questions could be maintenance of effective encoding by interpolated testing (Pastötter et al., 2011), possibly due to increased test expectancy (Weinstein et al., 2014), although we did not specifically examine this possibility in our experiments. However, despite the initial advantage of the interspersed condition during learning, the interspersed and at-the-end conditions performed at the same level on the delayed test in all three experiments. We first look at the specific findings of each experiment, and then explore theories that may account for the observed effect.

In Experiments 1 and 2, subjects studied a presentation on APA style either with a short-answer question appearing after every few slides (interspersed condition), or all questions appearing at the end of the slideshow (at-the-end condition). During learning, higher accuracy was achieved by subjects in the interspersed condition. A week later, however, there was no

difference in accuracy between the at-the-end and interspersed conditions. Furthermore, the at-the-end condition maintained identical performance from retrieval practice to final test one week later. In Experiment 3, placement of quiz questions was manipulated between-subjects in a Cognitive Psychology class: half of the lectures included five questions at the end of class, whereas in the other half of the lectures, the 5 questions were distributed throughout the lecture with each question appearing immediately after relevant information was studied. Similarly to Experiments 1 and 2, the interspersed condition produced better performance during learning. Also similarly to Experiments 1 and 2, there was no difference between the interspersed and at-the-end conditions on the delayed test. In contrast to Experiments 1 and 2, though, both the at-the-end and interspersed condition dropped considerably in performance from initial learning to final test. However, the pattern was clearly consistent across all three experiments: interspersing questions throughout study resulted in better performance during learning, but no better performance on a delayed test.

Bjork (1994) proposed the desirable difficulties framework, according to which factors that introduce effort and reduce accuracy during learning lead to better long-term retention. Relevant to our data is the finding that retrieval practice after a longer retention interval (spaced retrieval) produces poorer performance during practice but better performance on a later test than retrieval practice after a shorter retention interval (massed retrieval; for a review see Cepeda et al., 2006). This is known as the spacing effect, whereby spaced retrieval practice appears to be worse in the short run but is better in the long run than massed retrieval practice. In the current study, the "interspersed" condition (not to be confused with typical the more common type of interleaving that is usually contrasted favorably with blocking) actually represents massing with respect to initial learning, with a short retention interval between studied material and retrieval

practice. On the other hand, the at-the-end condition creates spacing with respect to initial learning. Although retrieval practice occurs in one section at the end of study, this provides for a greater retention interval between studied material and retrieval practice, thus mirroring typical spaced retrieval practice. However, we did not find the reversal between performance during retrieval practice (massed better than spaced) and the final delayed test (spaced better than massed) that usually occurs in spacing effect paradigms. We did find better performance during retrieval practice in the interspersed (shorter retention interval and thus more similar to massed) condition, in line with the usual spacing effect, but we found equivalent performance in the two conditions on the final delayed test. Further research is needed to determine whether adding a delay in the interspersed condition (i.e., presenting quiz questions not immediately after the slide in which the information is presented, but several slides later) could combine the initial benefits of interspersing with the more long-lasting benefits of spacing.

Why did the at-the-end condition produce the same level of final test performance as the interspersed condition? Based on the general idea that more effortful or difficult retrieval enhances learning (Bjork, 1994; Bjork & Bjork, 2011), retrieving the answers to questions at the end of the materials ought to have increased learning relative to retrieving the answers to questions interspersed throughout the materials. On the other hand, in all three experiments, there was a significant difference in initial retrieval success favoring the interspersed condition. Although the questions placed at the end of the materials required more effortful retrieval (Bjork, 1994) and presumably more context reinstatement (Karpicke et al., 2014; Siegel & Kahana, 2014), there was also much less retrieval success on the at-the-end questions, due to the longer retention interval between study and initial test. It is also possible that the build-up of proactive interference during study, which would have been eliminated by the interpolated tests in the

interspersed condition (Szpunar et al., 2008), contributed to lower performance on the at-the-end initial questions. In the present experiments, the greater levels of retrieval success in the interspersed condition most likely balanced out the benefits of effortful retrieval in the at-the-end condition.

The pattern of results obtained by Wissman and Rawson (2015) in a set of experiments with free recall of prose passages was the same as what we observed: Subjects performed better on initial free recall tests placed throughout the passage than they did on a free recall test at the end, but the two conditions did not differ on a final assessment. The present experiments examined the effects of interspersed and at-the-end retrieval practice conditions relative to a study-only control condition. This is important because it demonstrates that the lack of difference between interspersed and at-the-end retrieval on the final test does not simply reflect the absence of a retrieval practice effect. Our experiments also generalized the effects to multimedia (PowerPoint) materials, to short-answer question formats commonly used in educational settings, and to authentic classroom conditions (in Experiment 3). Taken together, the present results and Wissman and Rawson's results provide complementary evidence that interspersed retrieval practice produces gains during initial learning, relative to at-the-end retrieval activities, but does not confer benefits on delayed tests.

It is likely that factors unexplored in our set of experiments may moderate the efficacy of practice quiz questions and their best placement. For instance, for very difficult materials using interspersed practice questions might be better than using at-the-end practice questions because the latter will likely reduce retrieval success to a minimum that might override any context reinstatement and/or retrieval effort benefits. However, for easier materials it is likely that at-the-end practice questions would be more efficient than interspersed practice questions. This

possible interaction between materials difficulty and question placement was not investigated in the current set of experiments, and it is a future area that needs clarification. Of course, the role of individual differences between learners and other intrinsic factors such as inclusion and placement of feedback might also have an impact on the ideal placement of practice questions.

So, knowing what we know thus far, what can we recommend to instructors? Is it better to include retrieval practice questions interspersed during study, or at the end of a study session? As we observed, there is not a simple answer to this question. Interspersing produced better initial performance, but at-the-end questions produced less forgetting over time. The effectiveness of interspersed versus at-the-end practice questions may be determined in terms of the amount of information that is forgotten between the initial and final test, in which case at-the-end practice questions may be considered more effective. On the other hand, initial test performance may be an important factor in the instructor's decision of quiz question placement because good performance on interspersed tests can boost class morale and keep students engaged and motivated to keep learning, and in this case, interspersed practice questions may be preferable. However, this boost in class morale could come with metacognitive illusions (i.e., overestimation of learning) that might have a negative effect on future study decisions. Considering the results of our three experiments, ideally, one would maximize retrieval effort with at-the-end practice questions in a learning phase that also maximizes retrieval success (i.e., increasing it to the level of interspersed condition, or to ceiling). We believe this would be the best method for increasing the efficacy of practice questions, although more research is needed to identify a viable method for maximizing both retrieval success and retrieval effort.

References

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs, 11,* 159-177.

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

Bjork, R. A. (2013). Desirable difficulties perspective on learning. In H. Pashler (Ed.), *Encyclopedia of the Mind.* Thousand Oaks: Sage Reference.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society* (pp. 56-64). New York: Worth Publishers.

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, *19*(5), 619-636.

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.

Darley, C. F., & Murdock, B. B. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology*, *91*, 66-73.

Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation*, *53*, 63-147.

Dempster, F. N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology*, *79*, 162-170.

Dunlosky, J., Bottiroli, S., & Hartwig, M. (2009). Sins committed in the name of ecological validity: A call for representative design in education research. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.). (pp. 430-440). *Handbook of Metacognition in Education*. NY: Psychology Press.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *Psychology of Learning and Motivation, Vol. 61* (pp. 237-284). San Diego, CA: Elsevier Academic Press.

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *5*7, 151-162.

Kay, R. H., & LeSage, A. (2009). A strategic assessment of audience response systems used in higher education. *Australasian Journal of Educational Technology*, *25*, 235-249.

Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., ... & Zhang, H. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology*, *34*, 51-57.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger III, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*, 399-414.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*, 494-513.

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.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*, 437-447.

Roediger III, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*, 382-395.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning taking memory tests improves long-term retention. *Psychological Science*, *17*, 249-255.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*, 1432-1463.

Siegel, L. L., & Kahana, M. J. (2014). A retrieved context account of spacing and repetition effects in free recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 40,* 755-764.

Slamecka, N. J., & McElree, B. (1983). Normal forgetting of verbal lists as a function of their degree of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 384-397.

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.

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.

Wissman, K. T., & Rawson, K. A. (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.

Retention interval (# of days)	Section 1	Section 2	Topic
63	interspersed	at-the-end	History of Cognitive Psychology
56	at-the-end	interspersed	Perception
54	interspersed	at-the-end	Mental Imagery
47	at-the-end	interspersed	Divided Attention
42	at-the-end	interspersed	Categorization
26	at-the-end	interspersed	Long-Term Memory
21	interspersed	at-the-end	Encoding and Retrieval
19	interspersed	at-the-end	Autobiographical Memory

Table 1. *Quasi-random counterbalancing for Experiment 3: the second and third columns indicate whether questions for a given topic were interspersed or presented at the end in each section of the class.*

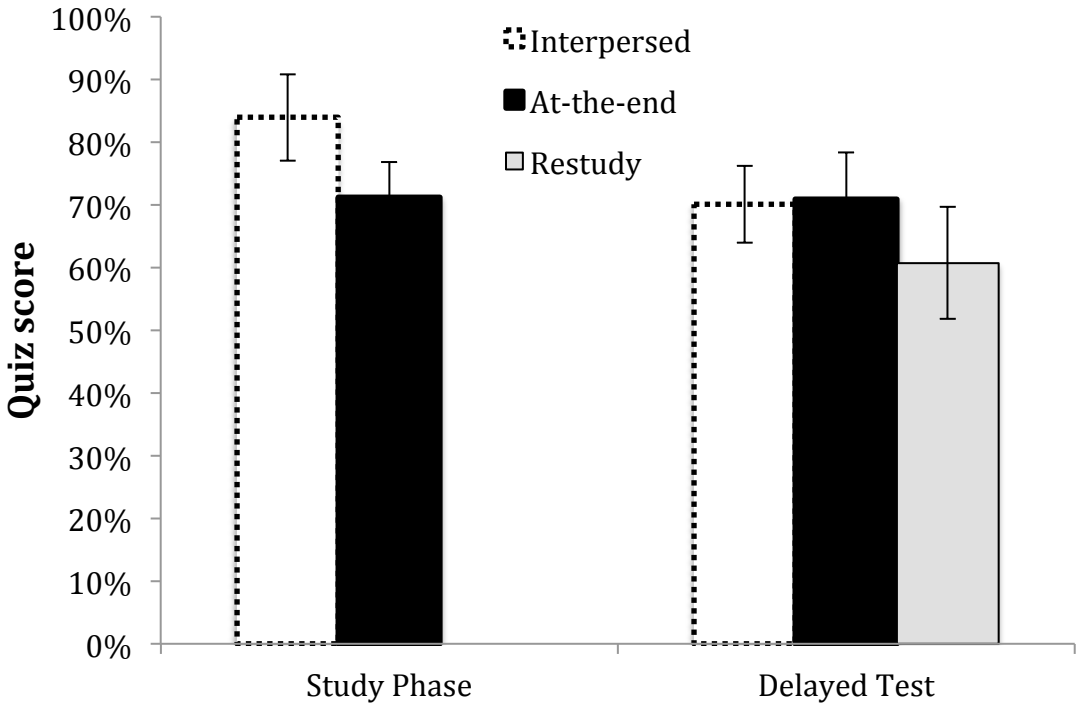


Figure 1. *Test performance in the interspersed and at-the-end conditions in the learning phase, and in the interspersed, at-the-end, and restudy conditions on the delayed test of Experiment 1. Error bars represent 95% confidence intervals.*

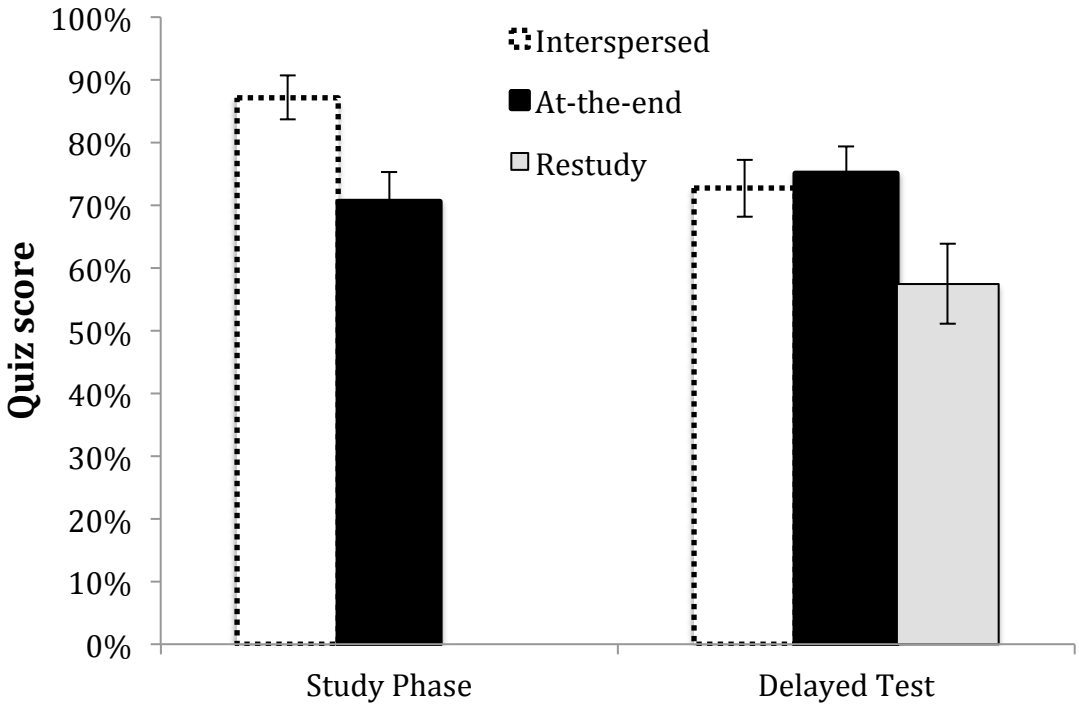


Figure 2. *Test performance in the interspersed and at-the-end conditions in the learning phase, and in the interspersed, at-the-end, and restudy conditions on the delayed test of Experiment 2. Error bars represent 95% confidence intervals.*

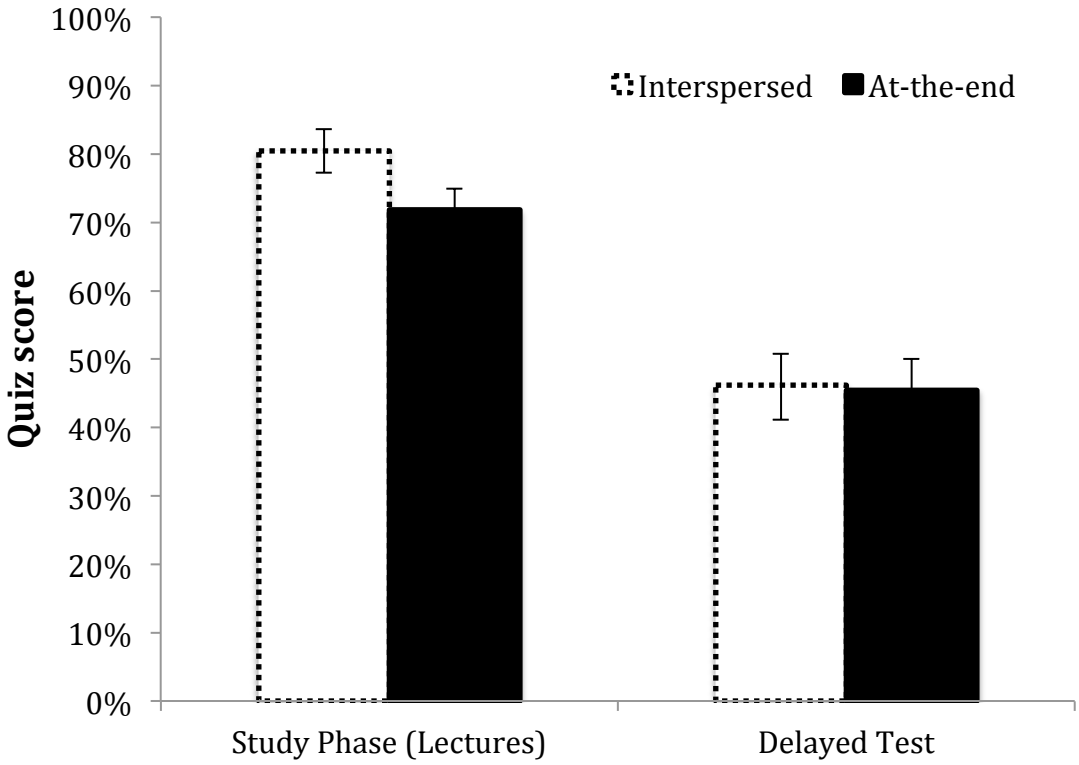
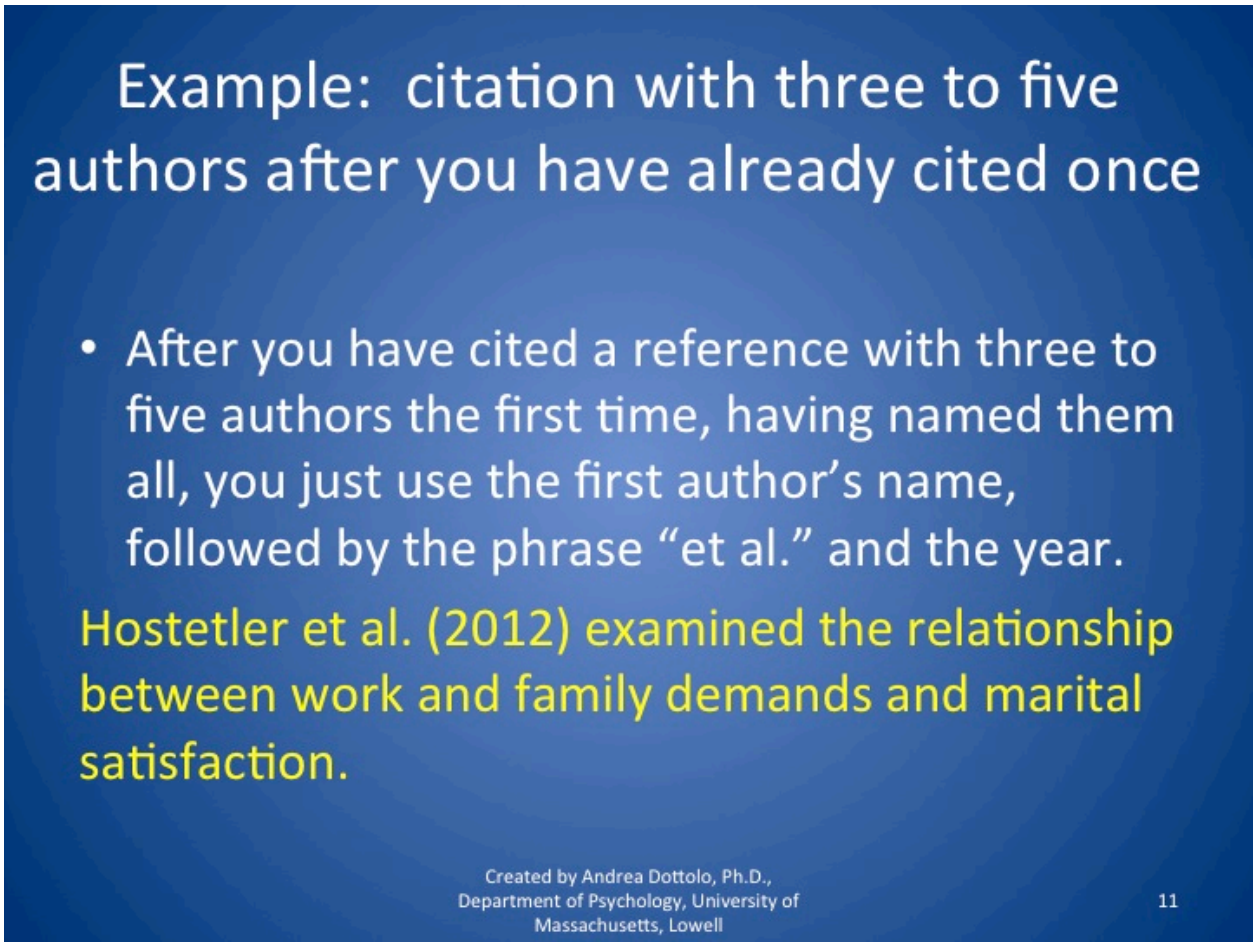


Figure 3. *Test performance in the interspersed and at-the-end conditions, in the study phase lectures and delayed test of Experiment 3. Error bars represent 95% confidence intervals.*

Appendix 1

Sample Slide, Question, and Rubric for Experiments 1 and 2



Question: After citing a reference with three to five authors, how do you subsequently refer to that article?

Correct Answer: First author's name and et al.

Rubric: 1 point for "first author's name", 1 point for "et al." (max 2 points total)

Appendix 2

Questions for In-Text Citation presentation in Experiments 1 and 2

Slide	Question	Answer	Points
1	*no question*		
2	What two pieces of information do you generally include when you are citing an authored publication in the text?	Author(s) and year of publication	2
3	How often do we quote directly from other published work in psychology research papers?	Almost never	1
4	Imagine you are trying to cite a paper that was written by Alyson Traficante in 2012. Fill in the blank below to cite the reference. "Children are often the victims of bullying (_____)."	Traficante, 2012	3
5	Reword the following sentence without the use of parentheses. "A review of the literature suggests that children's parents are often unaware of bullying (Traficante, 2013)."	"Traficante, in her 2013 review of the literature, found that children's parents are often unaware of bullying." Author outside of parentheses, year outside of parentheses, and restatement of sentence	3
6	*no question*		
7	The two authors are linked by what word or character in the following situations a) in parentheses b) not in parentheses	a) & b) and	2
8	When you are citing a two-author paper, when do you have to cite both authors' names?	always	1
9	*no question*		

10	*no question*		
11	After citing a reference with three to five authors, how do you subsequently refer to that article?	First author's name and et al.	2
12	*no question*		
13	*no question*		
14	When there are 6 or more authors in one reference, how do you cite the reference in the text for the first time?	First author's name and et al.	2
15	*no question*		
16	If you are citing multiple works in one sentence, in what order do you list them in parentheses?	alphabetical	1
17	When citing multiple sources in one sentence, the individual references are separated by what word or symbol?	;	1
18	*no question*		

Appendix 3

Questions for References presentation in Experiments 1 and 2

Slide	Question	Answer	Points
1	*no question*		
2	*no question*		
3	Which sources that you cite in a paper must appear in your reference list?	all of them	1
4	What is the name of the section of your paper where you list all the sources that you used?	References	1
5	What is the order of references in a reference list?	Alphabetized, by last name and/or by first author	2
6	*no question*		
7	If an author's name is Alyson Traficante, how would their name appear in the reference list?	last name, comma, first initial (1 point for last name appearing first, 1 point for abbreviating the first name to just an initial (doesn't matter about the period), 1 point for comma)	3
8	*no question*		
9	When referencing a journal article, where and how does the year appear?	After the authors' names, in parentheses	2
10	Which word or words are capitalized in the title of the article in a reference list?	First word, First word after a colon, Proper nouns (any 2 of the 3)	2
11	When citing an article in a reference list, what emphasis is placed on the font of the journal title and what form of punctuation is it followed by?	italics, comma	2
12	*no question*		
13	*no question*		
14	What page numbers appear after an issue number when citing an article in a reference list, and what form of punctuation are they followed by?	first and last page number, period	2
15	What does DOI stand for?	Digital Object Identifier (1 point per word, up to 2)	2

16	Where in the article can the DOI be found?	first page	1
17	*no question*		
18	*no question*		

Appendix 4

Sample Question, and Rubric for Experiment 3

Question: How does Chomsky's viewpoint on language go against behaviorism?

Rubric: To get 1 point, the answer needs to have two aspects. First, it needs to specify one of Chomsky's arguments:

- Language is more than just conditioning
- Children say things they haven't heard
- Children say things that are not rewarded
- Language must be innate

Second, it needs to explain how that goes against the behaviorist viewpoint (i.e., that behaviorists thought everything could be explained by stimulus-response associations).

Give 0.5 points if the student mentions one of Chomsky's arguments but does not link it with an explanation of how that goes against behaviorism, OR gives the explanation without specifying one of Chomsky's arguments.