Journal of Experimental Psychology: Applied
2015, Vol. 21, No. 4, 370–382

© 2015 American Psychological Association
1076-898X/15/$12.00 http://dx.doi.org/10.1037/xap0000063

When Does the Test-Study-Test Sequence Optimize Learning and Retention?

Mark A. McDaniel, Julie M. Bugg, Yiyi Liu, and Jessye Brick
Washington University in St. Louis

In educational learning contexts, unlike typical contemporary laboratory paradigms, students have repeated opportunities to study and learn target material, thereby potentially allowing different sequences of testing and studying. We investigated learning and retention after several plausible sequences that were patterned on a classic memory paradigm. After initially reading a research methods text, 2 days later in 1 condition participants repeatedly restudied the material 3 times (SSS), in another condition they engaged in a test-restudy-test sequence (TST), and in a third condition participants repeatedly tested on the studied material (3 times: TTT). Participants received a final test 5 days later. In Experiment 1, both TST and TTT produced better final performance than did SSS; however, TST was not better than TTT. In Experiment 2 the TST condition was altered so that after the first test, correct/incorrect feedback was provided and the test and feedback were available during the study phase. With this protocol, TST produced better learning and retention than did TTT or SSS. These findings suggest possible critical aspects regarding test feedback and the availability of previous tests for helping students to optimize their restudy efforts after low- or no-stakes quizzes.

Keywords: testing effect, indirect effects of testing, restudy after testing, test-potentiating effects

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Extensive laboratory work and a growing body of classroom research have established that testing (quizzing) enhances learning and retention more than does additional study (see Roediger & Karpicke, 2006a, for a review; see McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013, and McDermott, Agarwal, D'Antonio, Roediger, & McDaniel, 2014, for more recent summaries of classroom experiments). An overarching interpretation of these findings is that the retrieval required by testing enhances learning by directly modifying a person's knowledge (e.g., increases strength of the items or elaboration of the stored content, Carpenter & DeLosh, 2006; Halamish & Bjork, 2011; McDaniel & Masson, 1985). This interpretation is most compelling in laboratory experiments in which participants are not given the opportunity to restudy material in the testing conditions.

In contrast, in authentic educational contexts the situation is more complex. In classroom studies, unlike the laboratory, there is not a single acquisition session in which testing versus restudying of target material is manipulated, followed by a final criterial test. Rather, the testing conditions are embedded in a learning context in which students are presumably reviewing the target material (and are encouraged to do so by instructors) after the initial testing. Accordingly, testing effects in the classroom (e.g., Glass, 2009; Lyle & Crawford, 2011; McDaniel, Wildman, & Anderson, 2012; McDaniel et al., 2013; McDermott et al., 2014; Roediger, Agarwal, McDaniel, & McDermott, 2011) could also be a consequence of several indirect effects of testing. For instance, tests presumably provide students a fairly accurate gauge of what they know and what they do not know, thereby potentially allowing more effective study allocation (what to study) in preparation for a final (summative) assessment, and tests also may potentiate learning on subsequent study (Arnold & McDermott, 2013; Izawa, 1970, 1971; Little & McDaniel, 2015). To better appreciate testing effects that have emerged in authentic contexts in which testing and restudy could have been intermingled (or not), it is important to tease apart the benefits of testing alone relative to testing with restudy opportunities.

Our interest in the present article concerns this more complex and educationally authentic learning context in which there are repeated opportunities to study and learn target material. With repeated learning opportunities, potentially different sequences of testing and studying can be marshaled, and in this article we investigate the relative outcomes of several plausible sequences in terms of learning and retention. Specifically, after an initial study session in which participants read a didactic text on experimental design methods in psychology, participants were required to engage in repeated processing of the material. One condition repeatedly restudied the material (three times: SSS), another condition engaged in a test-restudy-test sequence (TST), and a third condition was repeatedly tested on the studied material (three times: TTT). These conditions reflect plausible candidates for how students in authentic contexts could use repeated learning opportunities. Surveys of student study behaviors report that repeated re-

This article was published Online First October 26, 2015.

Mark A. McDaniel, Julie M. Bugg, Yiyi Liu, and Jessye Brick, Department of Psychology, Washington University in St. Louis.

This research was supported in part by Grant 22020166, Applying Cognitive Psychology to Enhance Educational Practice, from the James S. McDonnell Foundation. Mark McDaniel's effort in preparing the article was supported by Grant R305A110550 from the Institute of Education Sciences, and Yiyi Liu's effort was supported in part by Grant R305A130535 from the Institute of Education Sciences.

Correspondence concerning this article should be addressed to Mark A. McDaniel, Department of Psychology, Washington University, Campus Box 1125, Saint Louis, MO 63130. E-mail: markmcdaniel@wustl.edu

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

studying of material is a favored strategy (Hartwig & Dunlosky, 2012; Karpicke, Butler, & Roediger, 2009). Some have also claimed that interleaving studying and testing is a technique that students adopt to learn new material (Karpicke & Roediger, 2007). Perhaps less frequently used by students, but of theoretical interest, is simply practicing repeated retrieval (TTT). This condition reveals the direct effects of testing, without contamination of potential indirect effects that occur on restudy. Though these sequences have been examined in basic laboratory work using simplistic materials, there is no published work of which we are aware that has directly contrasted the effects of these sequences on learning complex concepts and constructs in an authentic content domain. It is important to do so for both theoretical and practical reasons, as we develop in Experiment 1 when considering the possible patterns that might emerge.

Consider first the results from a multitrial learning experiment using word lists. Karpicke and Roediger (2007, Experiment 1) had participants study a word list and then followed that study with either three more study trials (SSS; our labels focus on the activities following the initial study), three test trials (TTT; in this case free recall), or alternating test and study trials (TST). On a final test given 1 week later, TST and TTT produced better free recall performance than did SSS (these differences reflected medium to slightly under large effect sizes), with TST also showing a slight advantage (small effect size) relative to TTT. The implication is that alternating testing and restudy (TST) may be a preferred method of sequencing relearning opportunities. It appears to provide the direct benefits of testing coupled with the indirect (metacognitive) benefits of increased awareness during restudy of which items were not learned (i.e., during restudy participants might recognize which items they could not previously recall; cf. Karpicke and Roediger, 2007), thereby allowing effective focus on those items during that restudy trial.[1]

A major issue is whether an advantage for intermingling testing and restudy emerges with more educationally authentic materials, test tasks, and relearning intervals following initial study. Briefly, our to-be-learned material was content about research methodology in psychology, our test tasks were multiple choice and short answer (as opposed to free recall, which would be uncommon in classroom tests), and we delayed the relearning session that followed the initial study session by several days to approximate the realistic situation in which students would not restudy immediately after their initial reading (unlike the typical multitrial learning experiment in which all study, testing, and restudying occurs within a single session; Karpicke & Roediger, 2007; Tulving, 1967). The final test was administered 5 days after the relearning session.

Experiment 1

One straightforward prediction is that the basic laboratory findings (with list learning and free recall), will generalize to more complex materials, authentic tests, and spaced relearning opportunities, such that intermingling testing and restudy is better than either restudy alone or testing alone. For our first experiment, we followed the Karpicke and Roediger (2007) procedure of not providing feedback for the tests. Karpicke and Roediger suggested that during the restudy opportunity in TST, learners would recognize what they missed on the preceding test and be able to deploy restudy effort accordingly. For the advantage of TST over TTT to emerge with the present authentic materials, we assumed that a similar dynamic would need to be present.

There are reasons, however, to suspect that TST will not necessarily be superior when applied to more complex materials and authentic tests. First, the potency of repeated retrieval for long-term learning has been shown in laboratory studies that have used relatively complex materials (texts). In particular, when repeated retrieval attempts, with no feedback and no additional restudy, followed initial study (of an educational text) substantial gains in week-long retention have been observed relative to conditions with fewer retrieval attempts and more restudy (see Karpicke, 2012). For complex educational materials, testing (retrieval practice) may be especially advantageous because it provides the learner with more focused exposure (in the present case, three retrieval-practice exposures) to the critical target information (that will appear on the final test). Even with no feedback, successful initial retrieval is reinforced by subsequent retrievals, thereby attenuating forgetting (see Roediger & Karpicke, 2006b). By contrast, with content-rich, complex texts, even when restudy follows an initial test (in the TST condition), learners may not be able to successfully identify and extract all of the critical target information in the text. That is, for complex texts, during restudy but not retrieval practice, learners must be able to effectively identify target material that needs to be restudied, because all content in the text will not be presented on the final test (unlike word list materials tested with free recall, as in Karpicke & Roediger, 2007, Experiment 1).

A second challenge during restudy in the TST condition is not only to identify the critical material within the text (based on the previous test experience) but to recognize which of that material was answered incorrectly on the previous test. The idea here is that TST is an optimal relearning sequence because students' restudy can be effectively directed at material that has not been well learned (cf. Karpicke & Roediger, 2007; Little & McDaniel, 2015). As mentioned above, however, in this experiment no feedback was given on the initial tests. Therefore, to focus restudy on material not well learned, during restudy learners must be able to remember what was tested, how they answered, and then determine which answers were incorrect. It seems possible that learners would not be able to completely remember their test items and answers and would not necessarily be able to determine the correctness of their answers from restudy alone (see Dunlosky, Hartwig, Rawson, & Lipko, 2011). In light of these considerations, it would be expected that TST would fail to produce better final test outcomes than TTT.

One additional feature of the experiment merits mention. We wanted to evaluate the relative effectiveness of the relearning sequences (SSS, TST, TTT) across the range of final test items that reflect those used in educational contexts. In classroom experiments that have evaluated the testing effect (McDaniel et al., 2013; McDaniel et al., 2012; McDermott et al., 2014) and in instructors' spontaneous use of quizzing to augment learning (Wooldridge,

[1] Note that the paradigm repeated these cycles over five blocks so that TTT participants did receive a further study opportunity at the beginning of each block. Thus, even though there were further study opportunities for TTT, the idea is that intermingling testing and study on alternating presentations (TST) provides more advantage than a preponderance of repeated testing.

Bugg, McDaniel, & Liu, 2014), final tests (those on which students' grades are based) may incorporate questions that use question stems that are identical to those presented on the prior quizzes or questions that use different stems from those presented on the prior quizzes (but target identical concepts). Further, both classroom experiments and instructors can test definitional information or application of concepts on exams (indeed test banks accompanying textbooks typically include both questions types). Accordingly, for purposes of generality we designed the final (short-answer) test to include the four question types resulting from the factorial combination of question stem (*identical* or *different* from the stem used on the initial tests) and question type (*definition* or *application*). We thought it possible that TTT might most benefit final test performance (relative to TST) for identical question stems, given the increased direct practice on these questions and answers (see, e.g., McDaniel et al., 2012; Wooldridge et al., 2014). By contrast, the restudy opportunity with TST might allow a richer encoding of the target information than TTT and thereby better support performance on different stems, perhaps especially for application questions.

Method

Participants and design. Eighty-five students from Washington University participated in the experiment; 10 participants were excluded because they did not complete all three sessions of the study. The remaining 75 were randomly assigned to three study strategy conditions: study only (SSS; $n = 25$), quiz only (TTT; $n = 26$), and mixed quiz and study (TST; $n = 24$). Our sample sizes per condition were based on Karpicke and Roediger (2007, Experiment 1; $n = 20$ per condition); to be conservative, we slightly exceeded their sample sizes. The current sample size provided power of .11, .42, and .79 to detect small, medium, and large effects, respectively. Participants were compensated with either $20 or course credit.

The experiment used a 3 (Study Strategy: SSS, TST, TTT) × 2 (Question Type: Application, Definition) × 2 (Question Stem: Same, Different) mixed factorial design with study strategy manipulated between-subjects and question type and question stem manipulated within-subjects.

Materials. Twenty key concepts from research methods were identified as target concepts for the experiment (see Appendix A). Passages that covered these concepts were excerpted from *Research Methods in Psychology, 3rd edition* (Heiman, 2002), and the passages were combined into a 38-page packet. Tables and bullet points that summarized information were whited out but otherwise the materials appeared exactly as they did in the textbook.

The quizzes consisted of 40 multiple-choice questions: a *definition-based* question and an *application-based* question for each of 20 concepts.

For example, a definition question on reliability read:

> What is reliability?
>
> (a) The extent to which a procedure measures what it is intended to measure.
>
> (b) The extent to which our results generalize to other participants and other situations.
>
> (c) The extent to which a measurement reflects the hypothetical construct of interest.
>
> (d) The extent to which a measurement is consistent, can be reproduced, and avoids error.

The corresponding application question read:

> A student complains to her professor that her essay makes the same points as her friend's but she got a lower grade than her friend. She is complaining that the grading lacks _____.
>
> (a) Internal validity
>
> (b) External validity
>
> (c) Reliability
>
> (d) Concurrent validity

Quiz items across the different quizzes were identical, and the quizzes were also structured identically. The first half (i.e., 20 items) of the quiz included one question per concept, with half of the concepts (10 items) tested with application questions (one question per concept) and the remaining half tested with definition questions (one question per concept). The second half of the quiz also included one question per concept, but the type of question switched for each concept from the first half of the quiz. That is, concepts tested with application questions in the first half, were tested with definition questions in the second half, and concepts tested with definition questions initially were tested with application questions in the second half. The presentation order of the questions was randomized within each half of each quiz. Further, across quizzes the selection of whether a particular concept would be tested in the first half with an application or a definition question (and thereby also the second half) was also randomly determined. Participants did not receive feedback on any of the quizzes.

The final test was composed of 40 short answer questions, half of which were definition and half application questions (i.e., all 20 target concepts were tested on both definition and application questions). Eight definition questions and eight application questions used the same question stems as in the quizzes. Twelve definition questions and 12 application questions used different question stems from those in the quizzes (illustrated below). Question items from the quizzes were randomly determined a priori to be constructed as a *same-stem* or a *different-stem* question for the final test. That is 16 particular items from the quiz (eight definition and eight application) were always presented as same-stem questions and the remaining 24 items (12 definition and 12 application) were always presented as different stem questions. Same-stem questions appeared on the final test exactly as they did on the quiz but without the multiple-choice options. Different-stem questions covered the same concept but changed the focus or the context of the question. For definition questions, if the quiz question gave the concept-term and required a definition in the response, then the final test question would provide the definition and required the concept-term as the response (or vice versa). For instance, for the definition quiz question on *reliability* (provided above), the different-stem definition test question was:

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

> _____ is the extent to which a measurement is consistent, can be reproduced, and avoids error.

For application questions, the different-stem question would provide a different context from the one seen on the quiz question. For example, for the application quiz question on *reliability* (see Experiment 1 Materials), the different-stem application test question was:

> You take a test that measures how stubborn you are and score very high. A week later you take the same test and the results show you are only moderately stubborn. The test appears to lack _____.

Note that for the SSS condition, the question stem variable was arbitrary, as it was yoked to the quiz conditions.

Procedure. This was a three-session experiment lasting up to 135 minutes in total, with the time approximately evenly divided over the three sessions. After the initial session, participants returned for the second session 2 days later and the last session was held 1 week after the first session. For example, if a participant attended Session 1 on a Wednesday, he or she would return Friday for Session 2 and the following Wednesday for Session 3. Participants were tested in groups of one to four and each group was randomly assigned to a condition (SSS, TST, or TTT).

In the first session, participants were asked to read an information sheet containing elements of consent and give verbal informed consent. The participants were told that the purpose of the study was to examine how study and retrieval processes impact test performance. They were told that they would read a chapter on the first day, engage in study activities on the second day, and take a final test on the last day. Because of the materials being used, the experimenter asked if they had taken a research methods or experimental psychology class. All of the participants indicated that they never took this class. Following the intake procedure, participants received the 38-page research methods packet. They were instructed to read straight through the packet without marking on the packet for 45 min. At the conclusion of this time, the experimenter thanked the participants and confirmed the time of the next session.

During the second session, participants engaged in various study activities depending on the condition to which they were assigned. In the SSS condition, participants had 10 min to restudy the research methods packet. Then they received a highlighter and had 8 more min to study and highlight the passages they focused on. Finally, they returned the highlighter to the experimenter and had another 10 min to reread and restudy. In the TST condition, participants first took a multiple-choice quiz on the computer. Once they had finished, the experimenter handed them a copy of the packet and a highlighter. They had 8 min to restudy and were asked to highlight the passages they focused on. Finally, they took a second quiz on the computer. In the TTT condition, participants took three successive multiple-choice quizzes on the computer. They were given no feedback and did not have any opportunities to restudy. Participants in all conditions took approximately 30–45 min to complete the second session. At the conclusion of the session, the experimenter thanked the participants and confirmed the time of the last session.

When participants returned for the third session, they took the final test on the computer. They were instructed to type their answers into the black box that appeared below each question. Although they were asked to be concise, participants were allowed as much time as they needed to finish the test. Most participants completed the session within 30–45 min. After completing the task, participants were thanked and were assigned either credit or a payment voucher.

Results

Quiz performance. Across both experiments, all statistical analyses were conducted at a .05 alpha level. Performance on the first quiz was analyzed with a 2 (Study Strategy: TTT, TST) × 2 (Question Type: Definition, Application) × 2 (Question Stem: Same, Different)[2] mixed analysis of variance (ANOVA), with study strategy as the between subjects factor. There was a main effect of question type, $F(1, 48) = 12.13$, $MSE = .02$, $p = .001$, $\eta_p^2 = .20$, such that performance was higher for definition questions than for application questions (Table 1 provides means). No other main effects ($F < 1$ for study strategy effect) or interactions reached significance ($ps > .1$).

A parallel ANOVA conducted for the last quiz (Quiz 3 for TTT and Quiz 2 for TST) found an identical pattern. Performance was higher for definition questions than for application questions (see Table 1 for means), $F(1, 48) = 24.08$, $MSE = .02$, $p < .001$, $\eta_p^2 = .34$, and no other main effects ($F(1, 48) = 1.10$ for study strategy effect) or interactions were significant ($ps > .10$).

Final test performance. The authors created a rubric for scoring the final short answer test. They wrote ideal responses and also identified responses that would qualify for partial credit. Two scores were extracted for the final test. One was based on a strict-scoring procedure, in which participants received either full or no credit for their answer (a response could receive 1 or 0 points depending on if the answer was correct). The other score included partial-credit. For instance, a question on *content validity* read:

> You test the effectiveness of a motivational training program by providing it to half of your college's football team. The remaining members receive no training. The dependent variable is the coach's evaluation of each player. Does your study have content validity? Why?

The answer had two parts (i.e., "if the study has content validity" and "why") to it. For partial-credit scoring, a response could receive 0, .5, or 1 point, depending on how many parts were successfully answered.[3] Because the partial-credit scores are presumably most reflective of the scoring practices in authentic classrooms, we focus on those results. For purposes of completeness, however, we also briefly describe the results from the strict scoring procedures following presentation of the partial-credit score results.

[2] The same or different stem variable refers to whether a particular question was later presented on the short-answer test with the same or a different stem than used on the multiple choice quiz (test).

[3] In only one question, the answer had four parts in it, and thus a response could receive 0, .25, .5, .75, or 1 point, depending on how many parts were successfully answered. The question read: "In your study, you create different conditions by playing different types of music to participants. After a while, you suddenly pull out and shoot a (blank) pistol. You then measure participants' anxiety level to determine whether different types of music cause people to remain more or less calm in the face of startling stimuli. Name three possible conditions and identify one as a control condition."

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 1
Proportion of Correct Responses on the First and Last Quizzes in Experiments 1 and 2 as a Function of the Question Type and the Study Strategy

Experiment		TTT		TST	
		M	*SE*	*M*	*SE*
1	First quiz				
	Application	.71	.02	.67	.02
	Definition	.77	.02	.75	.03
	Last quiz				
	Application	.72	.03	.73	.03
	Definition	.78	.03	.84	.03
2	First quiz				
	Application	.71 (.73)	.03 (.03)	.80 (.80)	.03 (.03)
	Definition	.70 (.71)	.03 (.03)	.80 (.79)	.03 (.03)
	Last quiz				
	Application	.92 (.91)	.02 (.02)	.87 (.87)	.02 (.02)
	Definition	.92 (.90)	.02 (.02)	.90 (.89)	.02 (.02)

Note. For the first quiz in Experiment 2, data were available for 26 participants in the TTT condition and 24 in the TST condition (see Footnote 4). For Experiment 2, the values in parentheses reflect performance for participants with first and last quiz scores (see Footnote 7).

Figure 1 shows the proportion correct on the final test as a function of the question type, the question stem and the study strategy. These data were analyzed via a 3 (Study Strategy: SSS, TTT, TST) × 2 (Question Type: Definition, Application) × 2 (Question Stem: Same, Different) mixed analysis of variance (ANOVA), with study strategy as the only between subjects factor. The most important outcome for present purposes was a significant main effect of study strategy, $F(2, 72) = 7.23$, $p < .001$, $MSE = .072$, $\eta_p^2 = .17$. A Post Hoc Fisher's least significant difference (LSD) test confirmed that TTT ($M = .71$, $SE = .03$) and TST ($M = .69$, $SE = .03$) outperformed SSS ($M = .58$, $SE = .03$), $ps < .01$, whereas the difference between TTT and TST did not reach significance, $p = .46$. To provide further statistical evidence that final-test performance did not differ between TTT and TST conditions, we used the Bayes information criterion (BIC) value to generate the posterior probability of the null hypothesis (see Masson, 2011). The probability of the null, P_{BIC} (H_0|D), was .84, indicating positive support for the null (using Raftery's, 1995 guidelines).

There was a main effect of question stem, $F(1, 72) = 173.99$, $p < .001$, $MSE = .011$, $\eta_p^2 = .71$, such that performance on same-stem questions ($M = .74$, $SE = .02$) was higher than for different-stem questions ($M = .58$, $SE = .02$). Also, performance was slightly but significantly higher on the application questions ($M = .67$, $SE = .02$) than on the definition questions ($M = .65$, $SE = .02$), $F(1, 72) = 5.68$, $p < .05$, $MSE = .01$, $\eta_p^2 = .07$. Question type and study condition significantly interacted, $F(2, 72) = 3.19$, $p < .05$, $MSE = .01$, $\eta_p^2 = .08$. One a priori interpretation of this interaction suggested in the introduction is that TTT and TST groups might have differed on application but not definition conditions. However, inspection of means (see Figure 1) indicated that this pattern did not emerge and moreover, LSD tests revealed no significant differences across TTT and TST groups for either question type. Instead, pairwise tests showed that the interaction reflected that higher performance on application questions relative to definition questions was observed only for the TST study group, $t(23) = 2.13$, $p < .05$; performance on the two question types did not differ for the SSS or TTT study groups, $t(24) = 1.16$ and $t(25) = 1.64$, respectively.

In addition, the interaction between the question type and the question stem also reached significance, $F(1, 72) = 29.00$, $p < .001$, $MSE = .02$, $\eta_p^2 = .29$. The advantage of the application questions over the definition questions appeared within the same stem questions (.79 vs. .69) but not the different stem questions (.55 vs. .60). The three-way interaction did not reach significance, $p > .10$.

Scoring that did not allow partial credit (the strict-scoring procedure) produced patterns that paralleled those just reported. The only difference was minor: With strict scoring the question type and study strategy interaction was just marginally significant ($p = .10$).

Discussion

These results reinforce the potency of testing for enhancing learning and retention of authentic educational content. Both conditions that included testing during the relearning phase produced better final test performance than the restudy only condition. This pattern held both for definitional questions and for application questions, and was present even when the question stems on the final test differed from those on the initial tests. This latter finding adds to the growing body of research indicating that retrieval practice (testing) produces transfer of initially tested content to new questions and applications (e.g., Butler, 2010, in laboratory experiments; McDaniel et al., 2013, in classroom experiments). Importantly for present purposes, repeated retrieval practice (testing) alone produced nominally higher final test performance than the relearning condition that interleaved retrieval practice and restudy (TST).

At first blush, this finding may seem to represent a departure from previous findings with simple laboratory materials (e.g., word lists) that have reported small advantages in final test performance after TST relearning sequences relative to TTT sequences (Karpicke & Roediger, 2007, Experiment 1). However, an

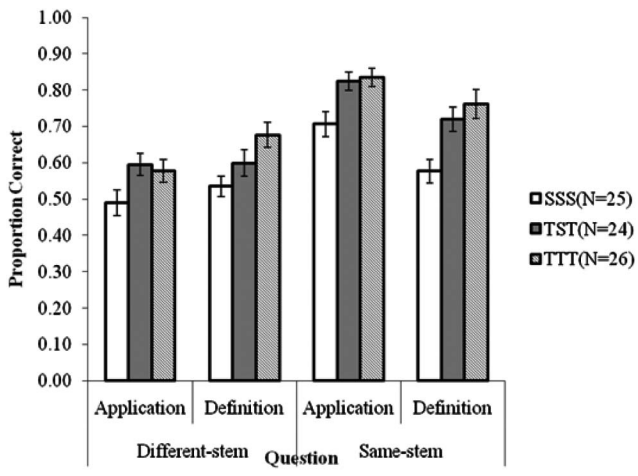

Figure 1. Proportion correct (using the partial-scoring procedure) on the final test in Experiment 1 as a function of the study strategy, the question type, and the question stem. Error bar represents the standard error.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

ambiguous aspect of the previous report is that the advantage of TST relative to TTT was established using only a probability of replication analysis (Killeen, 2005), with the $p_{rep} = .70$. Subsequent to the Karpicke and Roediger (2007) report, p_{rep} was shown to be a poor estimator of the probability of replication, substantially overestimating this probability for small effect sizes and small sample sizes (Iverson, Lee, & Wagenmakers, 2009), such as those present in Karpicke and Roediger (20 per condition). Had a conventional statistic been reported for the comparison of TST with TTT, it seems unlikely that a significant TST advantage would have emerged. Accordingly, a plausible interpretation is that the present results with learning of complex concepts converges with the word list findings (Karpicke & Roediger, 2007) to demonstrate that TST and TTT relearning sequences produce fairly comparable learning outcomes, at least as evident on a delayed final test. Note that these generally similar patterns emerged despite several additional differences (other than materials) between the paradigms. With the word lists, free recall tests (both prior and final tests) were used and the relearning sequences were repeated across five blocks prior to the final test (Karpicke & Roediger, 2007, Experiment 1); by contrast, with the present authentic materials, multiple choice quizzes and a short-answer final test were used to better mimic an educational context, and the relearning sequences were not repeated.

As developed in the introduction, we thought it possible that the TST sequence when implemented without feedback (as in Karpicke & Roediger, 2007) for educational material, might not promote effective restudy after the initial test. Our reasoning was that the learner would need to remember what was tested, remember the answers they gave, and figure out which answers were incorrect while restudying the text. Inspection of Table 1 indicates, however, that the TST participants did significantly improve after restudy on the final quiz relative to the first quiz, $F(1, 48) = 4.82$, $MSE = .01$, suggesting that, at least in the present experiment with a 38-page text on research methods, participants met these challenges. Yet, this improvement was not sufficient to yield a significant advantage for TST (relative to TTT) on the final test. In Experiment 2, we altered the testing and restudy procedure to more closely reflect educational practice, and by so doing anticipated that the effectiveness of TST might be enhanced.

Experiment 2

In terms of applied implications, the paradigm used in Experiment 1 had several limiting features: Feedback was not provided after the tests and learners did not have access to their tests (quizzes) while restudying. Though overlapping with basic laboratory work, this procedure does not necessarily overlap well with authentic educational contexts in which feedback is typically provided. In Experiment 2, after the first test, we provided feedback that indicated which items were answered correctly and which were answered incorrectly. The correct answer for incorrect items was not provided so that learners would have to consult the text during restudy to figure out the correct answer. Further, the test and feedback were available to learners during their restudy so that they would be certain which items they missed and the (incorrect) answer they gave on the initial test. We reasoned that this procedure would provide learners with unambiguous information regarding their comprehension or memory failures (more accurate metacognition) and would stimulate restudy focusing on the targeted information, particularly information in the text that would inform the learners' incorrect answers. To address this latter possibility, in this experiment we analyzed participants' highlighting responses in the restudy phase. With these feedback conditions, we thought it possible that TST might produce more robust learning and retention than TTT. Also, because the differences between TST and TTT have been slight in both a previous experiment with word lists (the TST advantage in Karpicke & Roediger, 2007, Experiment 1) and the current experiment with text, we increased the sample sizes to improve the power to detect an effect, if present.

Note that for the TTT condition, we implemented a similar initial test procedure. Learners received the same kind of feedback as for the TST condition, and they were allowed to look over this feedback before proceeding to the subsequent tests. To further increase the potency of the TTT condition, correct answer feedback was provided after the second and third tests (e.g., see Kang et al., 2007, for increased testing effects when feedback provided).

A second objective was to examine an issue regarding the benefits of testing. For educational materials, summative tests typically do not exhaustively assess the presented content (unlike the laboratory, where free recall tests of the word list or the entire passage are administered). Accordingly, using testing (retrieval practice) as a learning technique may be potent, at least in part, because it focuses the learner directly on the critical target information (that will appear on the final test, as in the present study). Restudy alone (SSS) does not confer that advantage. In the repeated restudy condition, the learners' processing activities are likely inefficient (relative to retrieval practice) because the learners are not able to focus all of their time exclusively on the to-be-tested information in the text. To address this possibility, in the SSS condition, we created a packet of study points that mimicked the content targeted in the initial tests. These study points were presented only in the second study phase, essentially to parallel the anticipated behavior of the participants in the study phase of the TST condition. We reasoned that if the testing effects found in Experiment 1 were largely a consequence of restricting focus to the key information on the final test, then the modified SSS condition could show performance levels approaching that for the TTT condition. However, to the extent that the restudy phase in the TST condition profits from test-potentiated learning processes (as outlined in the introduction; see Arnold & McDermott, 2013; Izawa, 1970, 1971; Little & McDaniel, 2015), TST should continue to demonstrate superior performance to SSS.

One final change from Experiment 1 is that for the final test, all questions used a different stem (or wording) than was present on the initial tests. This change, along with giving feedback on the initial tests, was implemented to create a more authentic context for examining the relative benefits of the SSS, TST, and TTT learning conditions. A recent survey of college psychology instructors' use of quizzing to promote learning (applying the testing effect) indicated that only a small minority (about 25%) retain the identical questions across quizzes and final tests; most instructors target similar content across quizzes and exams but change the wording of quiz questions for the exams (Wooldridge et al., 2014).

Method

Design and participants. The experiment was a 3 (Study Condition: SSS, TTT, TST) × 2 (Final-Test Question Type: Definition, Application) mixed factorial design, with study condition as the between-subjects factor, and final-test question type as the within-subjects factor. Our intent was to double the number of participants sampled in the Karpicke and Roediger (2007) experiment (60 total for the three study conditions). Accordingly, we tested 124 Washington University students. Seventeen participants were excluded, however, because they did not complete all three sessions of the study. The remaining 107 participants were distributed across the three study strategies as follows: study only (SSS; $n = 35$), quiz only (TTT; $n = 36$), and mixed quiz and study (TST; $n = 36$). This sample size provided a power of .13, .55, and .91 to detect small, medium, and large effect sizes, respectively. Participants were compensated with either $20 or partial fulfillment of a course requirement.

Materials. The materials were identical to those used in Experiment 1 with the following exceptions. Unlike Experiment 1, all of the questions on the final test used a different stem from the quiz questions. Thus, each question had a Stem A version and a Stem B version. Accordingly, additional questions had to be created for the questions that had been in the same-stem condition in Experiment 1. The two versions of each question were counterbalanced across the quizzes and final tests, such that half of the participants saw the Stem A questions on the quizzes and the Stem B questions on the final test, whereas the other half saw the Stem B questions on the quizzes and the Stem A questions on the final test. (Note that for the study-only condition the question stem variable was arbitrary, as it was yoked to the quiz conditions.) In addition, two versions of paper-based quizzes (the first test administered in both the TST and TTT conditions) were created (one for each counterbalancing condition: Stem A questions, Stem B questions); paper-based quizzes were used so that they could be returned to TST participants for their restudy phase.

For the SSS condition, three-page packets of study points were created for the second study phase (described below). The study statements reflected the information found in the quiz questions and also had two versions (one for Stem A and one for Stem B). For example, for the application quiz question on *reliability* (see Experiment 1 Materials), the statement version read:

> You take a test that measures how stubborn you are and score very high. A week later you take the same test and the results show you are only moderately stubborn. The test appears to lack reliability.

The statements were presented in the same order in which the corresponding questions appeared on the paper quizzes. One packet of statements corresponded to the Stem A questions and another packet corresponded to the Stem B questions. (Note that sometimes these statements were the same for both the A and B conditions; e.g., for the definition quiz questions on *reliability* (see Experiment 1 Materials), the statements for both stems read: "Reliability is the extent to which a measurement is consistent, can be reproduced, and avoids error.").

Procedure. The procedure paralleled that of Experiment 1. Again there were three sessions (initial reading, relearning phases, and final test) and the length and spacing of the sessions were identical to Experiment 1. The procedure of the first and third session remained the same, but the procedure for the second (relearning) session was altered. During the second session, participants engaged in various study activities depending on the condition to which they were assigned. In the SSS condition, for the first study, participants were instructed to read and study a packet of statements about the chapter that they read during the previous session. The statements reflected the content in the quiz items in the TTT and TST groups. They were allowed as much time as necessary to read through the packet. Most participants needed approximately 10 min. For the second study they received a highlighter and the original 38-page packet in addition to the packet of statements. They had 15 min to restudy the chapter and were told they could use the statements they just read to help them restudy. They were also asked to highlight the passages they focused on while restudying. After 15 min, participants returned the packet of statements, the highlighter, and the original 38-page packet to the experimenter. For the third study, they were represented with the statements, but this time on the computer. The statements were presented in a randomized order. Participants could take as much time to read a statement as they needed and were asked to press a button when they were ready to view the next statement.

In the TST condition, participants first took a paper-based quiz. Participants were allowed to take as much time as they needed and most took approximately 10 min. Once they had finished, the experimenter took away the quiz and marked answers as either correct or incorrect. If incorrect, the experimenter did not supply the correct answer. For the study period, the experimenter then returned the graded quiz to the participants along with the original reading packet and a highlighter. They had 15 min to restudy the chapter and were told they could use the feedback on the quiz to help them restudy. They were also asked to highlight the passages they focused on while restudying (identical to the second study period in the SSS condition). Finally, participants returned the quiz, highlighter, and reading packet and took a second quiz on the computer. Once the participant entered his or her response in self-paced fashion, the computer would display the correct answer.

The first step of the TTT procedure was identical to the first step in the TST condition. After the experimenter returned the graded paper quiz to the participant, however, the TTT group had 5 min to look over the feedback on the quiz without the help of the chapter or highlighter. Participants then returned their graded quiz to the experimenter and took their second quiz. Though this was self-paced, we estimated that participants would spend about 10 min; thus, total time spent on processing Quiz 1 feedback and taking Quiz 2 approached 15 min. Participants then took the third quiz. Both Quiz 2 and 3 were computer presented, and the correct answer to each question was shown immediately after responding.

Both the computer-based read statements and the computer-based quizzes presented items one at a time, and participants could not go back to previous items after proceeding to later items. Participants in all conditions took approximately 30–45 min to complete the second session. At the conclusion of the session, the experimenter thanked the participants and confirmed the time of the last session, which used the same procedure as for the last session in Experiment 1.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Results

Quiz performance.[4] Performance on the first quiz was analyzed with a 2 (Study Strategy: TTT, TST) × 2 (Question Type: Definition, Application) mixed ANOVA, with study strategy as the between subjects factor. There was a main effect of study strategy, $F(1, 48) = 6.17$, $MSE = .23$, $p < .05$, $\eta_p^2 = .11$, such that performance was higher for TST than for TTT (see Table 1). No other main effects or interactions reached significance. We also compared TTT with TST on the last quiz (Quiz 3 for TTT and Quiz 2 for TST) in a similar mixed ANOVA. Definition questions were answered slightly more accurately than application questions, $F(1, 68) = 3.12$, $p = .08$, and TTT performed slightly but not significantly better than TST, $F(1, 68) = 2.76$, $p = .10$. The interaction between the study strategy and the question type was marginally significant, $F(1, 68) = 3.69$, $MSE = .00$, $p = .06$, such that the advantage of TTT relative to TST was significant for application questions, $F(1, 68) = 5.50$, $p < .05$, but not definition questions ($F < 1$).

Final test performance. The two scoring procedures for the final test (partial scoring and strict scoring) were identical to that used in Experiment 1. As in Experiment 1, our main focus is on performance gauged by the partial-credit scoring scheme; however, for completeness analyses of the outcomes with the strict-scoring scheme are also briefly reported. Figure 2 shows the proportion of correct responses (based on the partial-scoring scheme) on the final test as a function of study strategy and question type. These data were analyzed with a 3 (Study Strategy: SSS, TTT, TST) × 2 (Question Type: Definition, Application) mixed ANOVA. Importantly, the ANOVA revealed a significant main effect of study strategy, $F(2, 104) = 4.48$, $p < .05$, $MSE = .04$, $\eta_p^2 = .08$. A post hoc LSD test indicated that participants in the TST condition ($M = .73$, $SE = .02$) outperformed participants in SSS ($M = .63$, $SE = .02$), $p < .01$, and in TTT ($M = .66$ $SE = .02$), $p = .05$. The TTT sequence did not produce significantly better final-test performance than did SSS ($p = .32$). The probability of the null hypothesis (for the TTT vs. SSS comparison), $P_{BIC}(H_0|D)$, was .84, indicating positive support for the null (using Raftery's, 1995, guidelines). There was also a main effect of question type, $F(1, 104) = 35.74$, $p < .001$, $MSE = .01$, $\eta_p^2 = .26$, with higher performance on application questions ($M = .71$, $SE = .01$) than definition questions ($M = .64$, $SE = .02$). There was no interaction between the study strategy and the question type ($F < 1$).

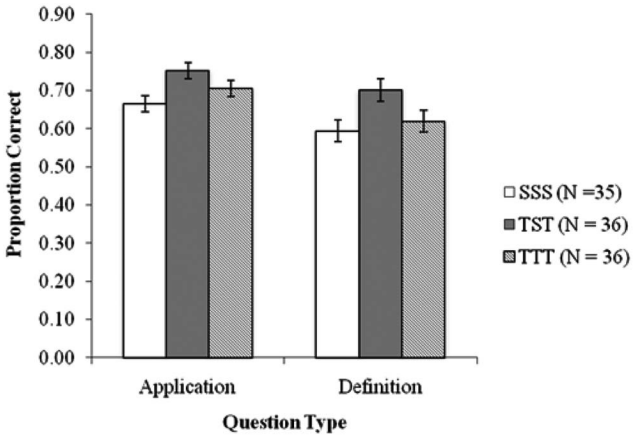

Figure 2. Proportion correct (using the partial scoring procedure) on the final test in Experiment 2 as a function of the study strategy and the question type. Error bar represents the standard error.

Results from the strict scoring procedure were identical to those just reported for the partial-credit scoring procedure. There were significant main effects of study strategy and question type, $F(2, 104) = 4.81$ and $F(1, 104) = 22.81$, respectively, ps $< .05$. As before, the LSD test showed that TST ($M = .67$, $SE = .03$) significantly outperformed TTT ($M = .59$, $SE = .03$, $p < .05$) and SSS ($M = .56$, $SE = .03$, $p < .01$), whereas TTT and SSS did not significantly differ, $p = .32$.

Cross-experimental analyses. To further explore the extent to which the TST versus TTT and TTT versus SSS study-condition patterns reported here reflected differences from Experiment 1, we conducted two additional ANOVAs on final test performance (partial-credit scoring) with experiment and study condition as between-subjects factors and question type as a within-subjects factor (for different stem items only, as these were used in both experiments).[5] In the first ANOVA, the study condition variable contrasted TST versus TTT (for efficiency, we only report study-condition effects). Overall there was no study condition difference ($F < 1$), but study condition marginally interacted with experiment, $F(1, 118) = 3.16$, $p = .08$, $MSE = .04$, $\eta_p^2 = .03$, reflecting the TST advantage in Experiment 2 but not 1. Further, a significant three-way interaction that included question type, $F(1, 118) = 8.54$, $p < .01$, $MSE = .01$, $\eta_p^2 = .07$, indicated that the change in the pattern across experiments of the TST–TTT difference was most robust for definition questions (for which TTT had higher performance in Experiment 1 but TST had higher performance in Experiment 2).

The second ANOVA contrasted SSS with TTT. Overall, TTT produced significantly higher performance than SSS, $F(1, 118) = 7.66$, $p < .01$, $MSE = .04$, $\eta_p^2 = .06$. This effect did not significantly interact with experiment, $F(1, 118) = 2.33$, $p = .13$, nor was the three-way interaction (including question type) significant, $F(1, 118) = 1.71$, $p = .19$.

Highlighting.[6] Participants in the SSS and TST conditions were scored on their highlighting on the original 38-page packet from the 15-min study phase (the second phase in the restudy sequences). The concepts that each participant highlighted were identified, and these were tabulated according to two categories: target concepts from the quiz questions (tested concepts), and untested concepts (see Appendix B for how the untested concepts were identified). The proportion of tested concepts highlighted (out of 20 possible; Appendix A lists the 20 concepts) and the proportion of untested concepts highlighted (out of 29 possible; see Appendix B) were derived for each participant.

[4] Final quiz scores for two participants were not available due to mechanical malfunction during the experimental session. For Quiz 1, 26 quizzes in the TTT condition and 24 quizzes in the TST condition were included in the analysis; 22 of these paper-and-pencil quizzes were misplaced during a relocation of the laboratory.

[5] We thank an anonymous reviewer for suggesting these analyses.

[6] Twenty-five protocols in the SSS condition and 24 in the TST condition were included in this analysis; 22 were misplaced during a relocation of the laboratory.

These proportions were submitted to a 2 (Study Strategy: SSS, TST) × 2 (Concept Type: Tested, Untested) mixed ANOVA. The analysis revealed a significant main effect of the study strategy, $F(1, 47) = 5.73$, $p < .05$, $MSE = .03$, $\eta_p^2 = .11$, indicating that SSS ($M = .40$, $SE = .03$) highlighted significantly more concepts than TST ($M = .31$, $SE = .03$). The main effect of the concept type was also significant, $F(1, 47) = 296.24$, $p < .01$, $MSE = .02$, $\eta_p^2 = .86$, showing that a higher proportion of tested concepts ($M = .57$, $SE = .03$) was highlighted than untested concepts ($M = .14$, $SE = .01$). The interaction between the two factors did not reach significance, $F(1, 47) = 1.37$, $p = .25$.

Next, for each participant we computed the proportion of highlighted concepts that were tested concepts (i.e., the number of tested concepts highlighted relative to the total number of concepts highlighted). A one-way ANOVA comparing the two conditions showed that the proportion of highlighted content that focused on the tested concepts was significantly lower for the SSS group ($M = .71$, $SE = .02$) than for the TST group ($M = .79$, $SE = .02$), $F(1, 47) = 5.86$, $p < .05$, $MSE = .01$, $\eta_p^2 = .11$. That is, for the content highlighted, the TST group was somewhat more focused on the tested concepts than was the SSS group.

We were further interested in whether the TST group used the feedback of the first quiz, which proceeded the study session, to guide their restudying and highlighting. A paired-sample *t* test was conducted to compare the likelihood that a concept was highlighted depending on whether the corresponding quiz question was answered correctly or incorrectly. The analysis revealed that subjects were much more likely to highlight the concepts of the questions answered incorrectly ($M = .78$, $SE = .06$) than those answered correctly ($M = .44$, $SE = .04$), $t(23) = 6.50$, $p < .05$.

Discussion

The key finding was that, unlike in Experiment 1, the TST sequence promoted better learning and retention as evidenced on the final test than did either repeated study (SSS) or repeated testing (TTT). This finding appeared to hinge on critical modifications to the TST procedure relative to Experiment 1: Feedback was provided for the first test (in the TST sequence), and participants had the test and the feedback available during their subsequent study. We reasoned that making the test available during restudy would better allow learners to focus their restudy on tested content. The analysis of highlighting behaviors supported this assumption. Participants in the TST condition were more focused in their highlighting (relative to the second study session in SSS): They highlighted less of the text and a higher proportion of highlighted information was related to the test items. Given that participants in the SSS condition were given "study facts" to guide their second study session, the increased focus on tested material by TST participants underscores the potentiating effects of tests per se (rather than just alerting learners to key information) for focusing students on critical to-be-learned content in assigned texts.

We also posited that giving feedback regarding whether responses were correct or incorrect (but not providing the correct answer) would further guide and motivate learners to especially focus on sections of the text that pertained to incorrectly answered questions. Again, the highlighting responses supported this hypothesis. On average, nearly 80% of the questions that were answered incorrectly were referenced in highlighted content (in TST). By contrast, less than half of the correctly answered questions were referenced in highlighted content. Overall then, the initial test with feedback promoted the expected restudy behaviors, plausibly leading to the superior learning and retention performance of the TST condition evidenced on the final test.

One possible caveat to the above interpretation is that the TST condition displayed higher performance than did the TTT condition on Quiz 1. Thus, after initially reading the text, TST participants showed more learning than TTT participants. However, this advantage was completely eliminated (and even reversed, significantly so for application questions) by the final quiz. That is, TTT produced substantial gains in performance from Quiz 1 to the final quiz, $F(1, 46) = 61.12$, $MSE = .01$; TST produced gains as well, $F(1, 46) = 11.86$, $MSE = .01$[7] (see Table 1), so that by the conclusion of the restudy session (i.e., after the final quiz) TST and TTT were performing at high levels on the information targeted by the final test. As we briefly suggested in the introduction, the subsequent advantage for TST on the final test (relative to TTT) might be because during the restudy opportunity, learners (at least in the present feedback paradigm) achieved a richer encoding and/or better understanding of the target information than learners gained from the test opportunity (Quiz 2 for TTT). This enhanced encoding could support better retention for the final delayed test (administered 5 days after the last quiz), or the richer encoding could better support performance on the different stem (from quiz items) questions than does repeated testing on identical question stems (TTT condition), or both. Further research is needed to directly investigate these possibilities.

In terms of educational applications, the positive finding with incorporating correct/incorrect feedback into the TST procedure may suggest a more nuanced view of what kind of feedback after testing is most effective for promoting learning. Some experiments with educationally authentic materials have found that feedback to initial tests that provides the correct answer is effective for increasing learning, whereas correct/incorrect feedback does not significantly improve later test performance relative to no feedback (e.g., Pashler, Cepeda, Wixted, & Rohrer, 2005). However, in these paradigms learners are not allowed a restudy opportunity after the initial test(s), and thus one would not expect that correct/incorrect feedback would support additional learning. The current findings indicate that under typical educational conditions, in which learners have the opportunity to review material that is quizzed, providing correct/incorrect feedback for quizzes can be effective for stimulating focused restudy. In further work it would be of interest to investigate whether correct-answer feedback on initial tests produces the same positive effects on restudy.

Another relatively novel finding from this experiment was that the SSS condition produced performance that was statistically equivalent to the TTT condition. The absence of a robust testing effect in this experiment is not a consequence of a short (immediate) delay between initial and final testing (cf. Roediger & Karpicke, 2006b), nor likely a consequence of the particular test

[7] These comparisons were conducted with the 25 TTT participants and 23 TST participants for whom both first and final quiz scores were available. The *MSE* for each comparison was derived from an overall analysis contrasting first and final quizzes for TTT and TST conditions.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

formats used (multiple choice formats for the initial tests and a short-answer format for the final test). Supporting this conclusion, the initial-test formats in Experiment 1 were also multiple choice and the final test was short answer, and a robust testing effect was produced.

From a theoretical perspective, these results suggest, at least for authentic educational materials and tests (that do not require recall of the entire text), that one benefit of testing may be that the test items identify specific content that will be subsequently tested. Restudy by contrast, in typical classrooms with authentic materials, often does not provide learners with specific guidance for items that will appear on the final test. Indeed, in laboratory experiments with authentic educational materials in which the restudy condition is given the benefit of a study list (or summary) that provides facts that subsequently appeared on the final test, initial tests may not produce better final performance than restudy (Butler & Roediger, 2007, when the initial test is multiple choice and the final test is short answer; Kang et al., 2007, Experiment 1; also see McDermott et al., 2014, and Roediger et al., 2011, for contrasting findings in middle school classroom experiments). Nevertheless, in light of the absence of a significant interaction between Experiments 1 and 2 and study (TTT–SSS) condition, further research is needed to determine the extent to which the provision of a study sheet in authentic contexts (educational texts and tests) reliably augments the effectiveness of restudy such that final test performance levels produced by repeated testing are approached.

General Discussion

Our objective in this study was to extend the recent research highlighting the effectiveness of using testing to promote learning with authentic educational materials and in authentic educational contexts (e.g., Glass, 2009; Lyle & Crawford, 2011; McDaniel et al., 2011, 2012, 2013; McDermott et al., 2014; Roediger et al., 2011). In authentic contexts, in addition to benefitting from practice tests or quizzes, students may intersperse restudy of to-be-learned material. However, existing testing-effect studies in authentic contexts have not explicitly examined or been sensitive to the potential benefits or consequences of such interspersed restudy when tests are administered to promote learning. Inspired by previous laboratory research with word lists, we were particularly interested in a plausible sequence in which, following initial reading of the text (or some other initial presentation of the material), students would experience an initial test, then restudy the target content, and then take the initial test again. With word lists, this TST sequence has been found to produce slightly better learning and retention than learning sequences that only incorporate testing (TTT; Karpicke & Roediger, 2007). With authentic materials and tests, the present results revealed that the benefits of TST relative to TTT may hinge on important aspects of how it is implemented, as we recapitulate below.

In the classic TST procedure (Karpicke & Roediger, 2007; Tulving, 1967) there is no direct feedback provided to test responses. This procedure is a sensible consequence of the original purpose of interspersing testing with study trials—to gauge learning during repeated study trials in order to document learning rates. However, in light of findings that retrieval is a potent modifier of memory (Carpenter & DeLosh, 2006; Halamish & Bjork, 2011; McDaniel & Masson, 1985), the classic TST procedure has been adopted as a technique for promoting learning and retention. The current Experiment 1 indicated that with authentic, complex materials and educationally relevant test tasks, this classic implementation that withholds feedback on test trials does not produce gains relative to TTT. Providing feedback on test trials appears to be one key component for supporting an advantage of TST over TTT with complex, authentic materials (Experiment 2).

An intriguing component of these different patterns, however, is that in both experiments, final quiz performance significantly improved relative to initial quiz performance after restudy in the TST condition and that final quiz performance in the TST condition was not substantially different from final quiz performance in the TTT condition. Thus, the restudy even in Experiment 1 appeared to provide effective feedback for the initial quiz. Yet, only in Experiment 2 did TST produce an advantage relative to TTT on the final test. It may be that this advantage hinged on further modifications over the standard procedure for providing feedback to initial tests. In studies examining the testing effect with authentic materials in classrooms, correct-answer feedback has been the norm; Lyle & Crawford, 2011; McDaniel et al., 2012, 2013; McDermott et al., 2014; Roediger et al., 2011. By contrast, the current TST condition provided correct/incorrect feedback. In another departure from the standard procedure in testing-effect experiments, we provided the initial test and the feedback to participants during their restudy phase (Experiment 2). One, admittedly speculative, interpretation is that these modifications allowed TST learners to more accurately gauge the correctness of their answers than in Experiment 1 (because Experiment 2 provided direct feedback); thus, learners were better able to devote more concentrated study resources to incorrect items. In addition, because correct answer feedback was not provided, students may have been stimulated to more fully engage during restudy, with such study activities perhaps including figuring out why initial answers were incorrect (more so than in Experiment 1, where students may have not been certain what items they missed). Figuring out "why" can lead to more complete learning and more flexible use of that information (e.g., on the different stem items in Experiment 2; McDaniel & Donnelly, 1996; see also Kendeou, Walsh, Smith, & O'Brien, 2014).

Clearly, Experiment 2 was not designed to determine whether these novel features (relative to existing experimental literature) of providing only correct/incorrect feedback and allowing access to the graded test are necessary for supporting the advantage of TST. Yet, the present results are suggestive that these components might be carefully considered as important techniques to augment test-enhanced learning effects in the classroom. From an applied perspective, these techniques could reflect effective adjustments to the standard practice in many classrooms, which is to give correct answer feedback, to not return quizzes—especially in large college classes where test security can be a concern (e.g., Jensen et al., 2014), or both.

From a theoretical perspective, to explain the advantage of TST over TTT in Experiment 2 we appeal to the just mentioned ideas, as well as mechanisms similar to those proffered by other theorists (e.g., Karpicke & Roediger, 2007; Little & McDaniel, 2015). In general, restudy after initial testing (i.e., the TST sequence) allows advantages of the indirect effects of testing to combine with the direct effects. Specifically, after an initial test (with feedback in the present case), learners likely have a more accurate appraisal of what they know and do not know than after study (SSS; see Little & McDaniel, 2015, for evidence with text materials); learners can then implement discrepancy-based study policies that focus on unlearned (or not well learned) material; and study of the target material may be more effective after the initial test than without the test (a test-potentiated

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

study effect; Arnold & McDermott, 2013). This interpretation finds some support in the study behaviors exhibited by TST participants in Experiment 2: nearly 80% of the initially incorrect items on the initial test were referenced in the highlighting. Moreover, even though SSS participants in Experiment 2, who were provided a study guide for the second restudy session, largely oriented their study efforts toward concepts listed in the study guide, they still did not display final test performances as high as the TST participants. This pattern suggests that an initial test helps potentiate more effective study of complex target material (see also Little & McDaniel, 2015). In sum, TST, at least as implemented in Experiment 2, appears to optimize both the direct and indirect effects of testing to promote learning and retention of authentic and complex educational material.

References

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

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1118–1133. http://dx.doi.org/10.1037/a0019902

Butler, A. C., & Roediger, H. L. M., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19,* 514–527. http://dx.doi.org/10.1080/09541440701326097

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34,* 268–276. http://dx.doi.org/10.3758/BF03193405

Dunlosky, J., Hartwig, M. K., Rawson, K. A., & Lipko, A. R. (2011). Improving college students' evaluation of text learning using idea-unit standards. *The Quarterly Journal of Experimental Psychology, 64,* 467–484. http://dx.doi.org/10.1080/17470218.2010.502239

Glass, A. L. (2009). The effect of distributed questioning with varied examples on exam performance on inference questions. *Educational Psychology, 29,* 831–848. http://dx.doi.org/10.1080/01443410903310674

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 801–812. http://dx.doi.org/10.1037/a0023219

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

Heiman, G. W. (2002). *Research methods in psychology* (3rd ed.). Boston, MA: Houghton Mifflin.

Iverson, G. J., Lee, M. D., & Wagenmakers, E.-J. (2009). p_{rep} misestimates the probability of replication. *Psychonomic Bulletin & Review, 16,* 424–429. http://dx.doi.org/10.3758/PBR.16.2.424

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

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

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test . . . or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review, 26,* 307–329. http://dx.doi.org/10.1007/s10648-013-9248-9

Kang, S. H., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19,* 528–558. http://dx.doi.org/10.1080/09541440601056620

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science, 21,* 157–163. http://dx.doi.org/10.1177/0963721412443552

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory, 17,* 471–479. http://dx.doi.org/10.1080/09658210802647009

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

Kendeou, P., Walsh, E. K., Smith, E. R., & O'Brien, E. J. (2014). Knowledge revision processes in refutation texts. *Discourse Processes, 51,* 374–397. http://dx.doi.org/10.1080/0163853X.2014.913961

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science, 16,* 345–353. http://dx.doi.org/10.1111/j.0956-7976.2005.01538.x

Little, J. L., & McDaniel, M. A. (2015). Metamemory monitoring and control following retrieval practice for text. *Memory & Cognition, 43,* 85–98. http://dx.doi.org/10.3758/s13421-014-0453-7

Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of the lecture improves performance on statistics exams. *Teaching of Psychology, 38,* 94–97. http://dx.doi.org/10.1177/0098628311401587

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43,* 679–690. http://dx.doi.org/10.3758/s13428-010-0049-5

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, 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. http://dx.doi.org/10.1037/a0021782

McDaniel, M. A., & Donnelly, C. M. (1996). Learning with analogy and elaborative interrogation. *Journal of Educational Psychology, 88,* 508–519. http://dx.doi.org/10.1037/0022-0663.88.3.508

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 371–385. http://dx.doi.org/10.1037/0278-7393.11.2.371

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27,* 360–372. http://dx.doi.org/10.1002/acp.2914

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory & Cognition, 1,* 18–26. http://dx.doi.org/10.1016/j.jarmac.2011.10.001

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., III, & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied, 20,* 3–21. http://dx.doi.org/10.1037/xap0000004

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 3–8.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology 1995* (pp. 111–196). Cambridge, MA: Blackwell.

Roediger, 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. http://dx.doi.org/10.1037/a0026252

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

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

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

Tulving, E. (1967). The effects of presentation and recall of material in free recall learning. *Journal of Verbal Learning and Verbal Behavior, 6,* 175–184.

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory & Cognition, 3,* 214–221. http://dx.doi.org/10.1016/j.jarmac.2014.07.001

Appendix A

Tested Concepts

	Concepts
1	Reliability
2	Validity
3	Content validity
4	Construct validity
5	Internal validity
6	External validity
7	Confounding variable
8	Simple random sampling
9	Volunteer bias
10	Between/within subjects design
11	Balancing participants
12	Subject mortality
13	Matching participants/matched-group design
14	Repeated measure design
15	Independent variable
16	Dependent variable
17	Condition
18	State/trait characteristics
19	Control condition
20	Confederate

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix B

Untested Concepts

	Concepts
1	Experimental groups
2	Temporal validity
3	Ecological validity
4	Concurrent validity
5	Error variance
6	Pretest
7	Conceptual replication
8	Literal replication
9	Subject sophistication
10	Subject history
11	Subject maturation
12	Diffusion of treatment
13	Demand characteristics
14	Operational definition
15	Practice effects
16	Fatigue effects
17	Carryover effects
18	Response sets
19	Limiting the population
20	Selection criteria
21	Collapsing
22	Experiment/experimental methods
23	Intervening variable
24	Controlling extraneous variable
25	Intervening variable
26	Pretest-posttest design
27	Random assignment
28	Deciding on controls to use
29	Sample size and representativeness

Note. The untested concepts were subheadings, bolded and italicized terms that appeared in the original packet but were not tested in the quiz and final exam.

Received February 2, 2015
Revision received August 12, 2015
Accepted August 14, 2015 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at http://notify.apa.org/ and you will be notified by e-mail when issues of interest to you become available!