Journal of Educational Psychology

0022-0663/15/$12.00 http://dx.doi.org/10.1037/edu0000074

Testing With Feedback Yields Potent, but Piecewise, Learning of History and Biology Facts

Steven C. Pan, Arpita Gopal, and Timothy C. Rickard
University of California, San Diego

Does correctly answering a test question about a multiterm fact enhance memory for the entire fact? We explored that issue in 4 experiments. Subjects first studied Advanced Placement History or Biology facts. Half of those facts were then restudied, whereas the remainder were tested using "5 W" (i.e., *who*, *what*, *when*, *where*, or *why*) or analogous questions. Each question assessed a specific critical term of the fact. In the first 3 experiments, 1 test question was posed per tested fact; in the fourth experiment, up to 3 different test questions were posed per tested fact. After a delay of at least 24 hr, a final test involved questions that assessed the same terms that were tested during training, as well as questions that assessed a different term from that previously tested. Results showed that testing produced piecewise fact learning: Tested terms benefited relative to restudy, but untested terms did not. That pattern held when either fill-in-the-blank or multiple-choice questions were used during training, when 1 or 2 test trials were used during training, for both history and biology facts, and when more than 1 term from each fact was tested during training. Thus, across a range of circumstances, taking tests on complex facts results in a selective memory benefit for tested terms. In analogous applied settings, testing on multiple response terms should promote more comprehensive retention.

Keywords: memory, testing effect, transfer, cognitive processes, fact learning

"Winston Churchill was Prime Minister of the United Kingdom during World War II." In domains ranging from astronomy to zoology, fact learning is essential for building foundational knowledge. Moreover, fact learning is often the chief goal of learners. The Advanced Placement (AP) exams, which over 2.2 million students in the United States and Canada take annually (College Board, 2013a), are a case in point. Those exams typically assess knowledge of 60 to 80 facts each—a subset of the many facts that students typically study for a year or more. Given the prevalence of fact learning, the question follows: What is the best way to learn and retain facts? One answer that is strongly backed by learning science: By taking tests.

The use of tests to improve memory, or *test-enhanced learning*, is strongly endorsed by many cognitive and educational psychologists (e.g., Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Pashler et al., 2007). This recommendation is backed by extensive work over decades and across numerous content domains (e.g., Bjork, 1975; Carrier & Pashler, 1992; Gates, 1917; Glover, 1989; for reviews, see Roediger & Karpicke, 2006, and Rowland, 2014). Carpenter, Pashler, and Cepeda (2009), for example, showed that taking tests on facts during an eighth grade U.S. history course boosts long-term retention by as much as 41% by the end of the course, relative to restudy. However, most reports of test-enhanced learning (also called the *testing effect*) are subject to a major caveat: The same questions are used during initial training and final tests (for discussion, see McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; Rohrer, Taylor, & Sholar, 2010). The important question of whether test-enhanced learning of facts *transfers* beyond tested content has received comparatively less attention.

In a recent brief review, Carpenter (2012) concluded that testing effects successfully transfer across changes in test format and to application questions. Kang, McDermott, and Roediger (2007), for example, found positive transfer of fact learning when the test format was changed from short answer during training to multiple choice on the final test. Further, Butler (2010) reported positive transfer for application questions that require inferences from initially tested facts (e.g., from animal wings to aircraft wings). Successful transfer to application questions has been reported elsewhere (e.g., Chan, 2009, 2010; Chan, McDermott, & Roediger, 2006; Foos & Fisher, 1988; Karpicke & Blunt, 2011; McDaniel, Howard, & Einstein, 2009), although no transfer in similar contexts has also been observed by Agarwal (2011) and Tran, Rohrer, and Pashler (2015). Moreover, there remain other categories of transfer of test-enhanced learning that have seen little research to date.

Steven C. Pan, Department of Psychology, University of California, San Diego; Arpita Gopal, Division of Biological Sciences, University of California, San Diego; Timothy C. Rickard, Department of Psychology, University of California, San Diego.

We thank Robert A. Bjork, Courtney Clark, and two anonymous reviewers for their comments on this manuscript. Thanks also to Bijan Malaklou for assistance with data collection, as well as Michelle Hickman and Jonathan Mejia for materials research. Steven C. Pan is supported by a National Science Foundation (NSF) Graduate Research Fellowship. Work on this article was also supported by a Psi Chi APS Convention Society Research Award to Steven C. Pan.

Correspondence concerning this article should be addressed to Timothy C. Rickard, Department of Psychology, #0109, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0109. E-mail: trickard@ucsd.edu

Testing and Transfer Between Fact Terms

The topic of transfer of test-enhanced learning between terms of a fact, which is the focus of this article, is qualitatively distinct from the types of transfer in the prior examples. It is an important issue because test questions usually do not assess entire facts, but rather parts of facts. This is especially true for *multiterm* facts, or facts with multiple critical terms (e.g., the fact that begins this article has four: *Winston Churchill*, *Prime Minister*, *United Kingdom*, and *World War II*). In the classroom and in everyday life, one may need to retrieve any of several different terms from a fact. Yet it remains unclear whether a test question covering one, two, or three terms of a fact will be sufficient to memorize the entire fact for later recall. Prior results relevant to that topic are mixed: Two studies show positive transfer of learning from one question to another from the same fact, whereas two others suggest no transfer.

McDaniel, Anderson, Derbish, and Morrisette (2007) administered online fill-in-the-blank or multiple-choice quizzes on neuroscience facts, and then provided subjects with delayed, detailed feedback on their performance. They observed positive transfer to multiple-choice unit tests in which previously untested parts of each fact were assessed (e.g., from the quiz question "All preganglionic axons, whether sympathetic or parasympathetic, release _____ as a neurotransmitter" to the unit test question "All _____ axons, whether sympathetic or parasympathetic, release acetylcholine as a neurotransmitter"). McDaniel et al. (2013, Experiment 1) reported positive transfer for *term-definition* reversals in which multiple-choice classroom quiz questions requiring key terms from science facts as correct responses were switched to final multiple-choice test questions requiring definitions as correct responses, or vice versa (e.g., from the quiz question "What is the definition of temperature?" and answer "The measure of the average kinetic energy of the particles in a substance," to the final test question "What is the measure of the average kinetic energy of the individual particles in an object?" and answer "Temperature").

In contrast to those studies, neither Hinze and Wiley (2011, Experiments 1 and 2) nor Pan, Wong, Potter, Mejia, and Rickard (in press) observed transfer. Hinze and Wiley administered fill-in-the-blank tests on portions of factual science texts, using two blanks per sentence or paragraph. Uniquely among the studies covered here, no feedback was provided; the facts themselves were also lengthier (e.g., "In mitosis, _____ are created from a parent cell. Each new cell contains a complete set of chromosomes which guarantees that they are _____. When the chromatids align this forms the metaphase plate which later becomes the location where the cell is split in two"). The final test entailed presenting the same portions of text, but with the two previous blanks filled and two new ones added; no positive transfer was found. Recently, Pan et al. (in press) demonstrated that test-enhanced learning does not transfer for triple associates (word triplets such as *lion, hunt, meat*), even when correct feedback was provided after each test trial. Performance gains for previously tested questions (e.g., *lion, meat, ?*) did not transfer, relative to restudy, to new questions on previously tested triplets (e.g., *meat, hunt, ?*). Although triplets lack sentence structure, those results reinforce the possibility that the benefits of testing can be highly specific to tested parts of a fact-like concept.

Overview of the Present Experiments

The present experiments, which directly extend the Pan et al. (in press) triplet design using authentic educational materials, assessed the consequences of answering one test question on a critical term of a fact (Experiments 1 to 3), or up to three questions, each on different critical terms of a fact (Experiment 4), for mastery of that entire fact. In contrast to Hinze and Wiley (2011), correct-answer feedback was provided during training. If the lack of feedback in the Hinze and Wiley study was solely responsible for their finding of no transfer, then positive transfer should be observed here. Our feedback is confined, however, to presenting the correct answer after each test trial. If the extensive delayed feedback in the McDaniel et al. (2007) study was critical to their positive transfer effects, then, based on the findings of Hinze and Wiley and Pan et al., no transfer is expected here. If the term-definition structure of the materials in McDaniel et al. (2013) was critical to their positive transfer results, then again no transfer would be expected here.

A secondary goal of the current study was to explore the role of fact domain expertise on testing and transfer effects. In each experiment, subjects with prior AP United States History, World History, or Biology experience in high school (constituting 25% to 50% of subjects per experiment), as well as subjects with no prior AP course experience, participated. It is possible that the greater expertise for AP students (if confirmed by overall performance) affords a richer knowledge base that will promote more integrative processing through testing, and hence better transfer of learning.

Experiment 1

Method

Subjects. The target sample size in this experiment was 40, which is comparable with that of prior testing-effect studies (e.g., Carpenter et al., 2009; Hinze & Wiley, 2011; McDaniel et al., 2007). Undergraduate students were recruited from the subject pool at the University of California, San Diego, and received course credit for their participation. Students from both lower- and upper-division courses were eligible to participate. Subject ages were $M = 21.03$, $SD = 2.53$, and ranged from 18 to 31 years. The majority (79%) of the sample was female. Data from two subjects were not analyzed: one as the result of a computer error generating unusable data and the other as the result of the subject not returning for the second session. The resulting sample size for analysis was 38.

Materials. Thirty-six history facts were obtained using the AP United States History and World History preparatory texts produced by *Barron's* and *Princeton Review* (Armstrong, Daniel, Kanarek, & Freer, 2014; McCannon, 2014; Meltzer & Bennett, 2014; Resnick, 2014). The history facts, which average 11 words in length, all contain three or more one-word critical terms that address any of the "5 W's" (i.e., *who*, *what*, *when*, *where*, or *why*) that are essential for the comprehension of that given fact. Moreover, most critical terms address different categories (e.g., *who* vs. *what*), and thus exhibit minimal semantic overlap. An example history fact (with critical terms italicized) is: "*Overlord*, an operation led by *Eisenhower*, began with the invasion of *Normandy*." For three critical terms in each fact, one fill-in-the-blank test question (e.g., "Overlord, an operation led by Eisenhower, began

with the invasion of _____") and one corresponding short-answer test question (e.g., "Overlord, an operation led by Eisenhower, began with the invasion of WHERE?") was created; both questions had the same one-word correct answer. Although similar in wording and structure to its fill-in-the-blank counterpart, the wording of the short-answer test question was modified when necessary to maintain grammatical accuracy. Overall, the full set of materials included three fill-in-the-blank and three short-answer questions for each fact (further examples are included in the Appendix).

Design and procedure. As illustrated in Figure 1, Session 1 contained two phases: the study phase and the training phase. During the study phase, subjects viewed all 36 history facts, one at a time, for 8 s each, and in random order determined anew for subjects. All facts were studied once.

In the subsequent training phase, facts were randomly assigned for each subject to one of two lists: the restudy list or the testing list. Facts in the restudy list were presented for 8 s each, using a procedure identical to that of the study phase. Facts in the testing list were presented as fill-in-the-blank questions with feedback (6 s to type the missing term followed by the correct answer being shown for 2 s, during which subjects were no longer allowed to type responses). For tested facts, only one of the three critical terms per fact was tested, and the tested term per fact was counterbalanced across subjects. All 36 facts were presented in random order. After all facts had been trained once, the training phase ended, and subjects were reminded to return in 48 hr for Session 2.

Session 2, the final test, assessed recall for the entire set of 108 answer terms across all 36 history facts. As illustrated in Figure 2, the final test involved three 36-trial blocks. Within each block, (a) each fact was assessed once in random order, (b) the assessment was a short-answer test question, (c) subjects had unlimited time to provide a one-word answer, and (d) no feedback was provided. In each block, the test question for each fact had a different missing term, such that the three critical terms of each fact were separately assessed over the three blocks. Further, within each block, six of the previously tested facts had questions with the same missing term that was to be retrieved as during training (*tested* questions) and 12 had a different missing term (*transfer* questions); the other 18 questions assessed facts with no prior retrieval practice (*restudied* questions). Thus, the six facts in the tested condition in block 1 (i.e., facts having the same missing term to be retrieved as was the case during training) were assessed in the transfer condition of Block 2 (with the missing term being different from that which was tested during training), and analogously, six of the facts in the transfer condition in Block 1 were assessed in the tested condition of Block 2. Block 3 completed that cycling of questions over the full set of 18 questions that were tested during training. Question assignment to blocks on the final test was counterbalanced and block order was randomized for each subject.

The testing of all three critical terms on the final test allowed for a comprehensive assessment of the consequences of testing for each fact. The choice of short-answer test questions for the final test (instead of, e.g., multiple-choice) was intended to approximate fact retrieval conditions under everyday circumstances, in which cued recall from memory is most common. Moreover, using different types of questions on the training and final tests (rather than fill-in-the-blank in both sessions) served to minimize the possibility that any transfer effects, if obtained, could be explained by superficial learning (i.e., learning related to adjacent words or sentence structure that could be used as retrieval cues, as opposed to learning of the fact itself that is generalizable to later contexts in which retrieval cues are not identical to those of prior testing).

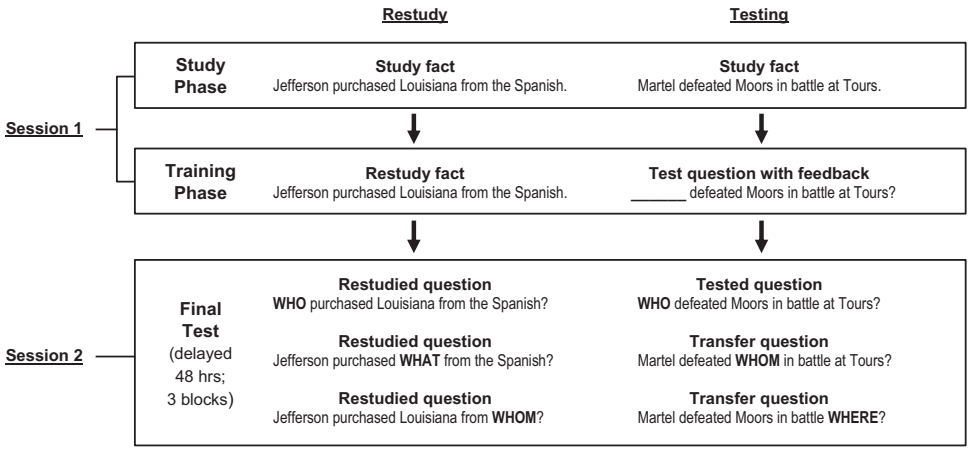

Figure 1. Experiments 1 to 3 procedure and example stimuli (Experiment 1) involving training through testing (right) and training through restudy (left). Training condition was manipulated within-subjects. (1) Study phase: subjects view all 36 facts, one at a time. (2) Training phase: facts trained via testing with feedback or restudy. In Experiments 1 and 2, fill-in-the-blank test questions are used; in Experiment 3, multiple-choice test questions are used. (3) Final test: recall for all 108 critical terms across 36 facts is assessed. There are three 36-trial blocks, within each of which one question per fact is shown. For previously restudied facts, all three final test questions are in the restudied condition; for previously tested facts, two questions are in the transfer condition, and one is in the tested condition.

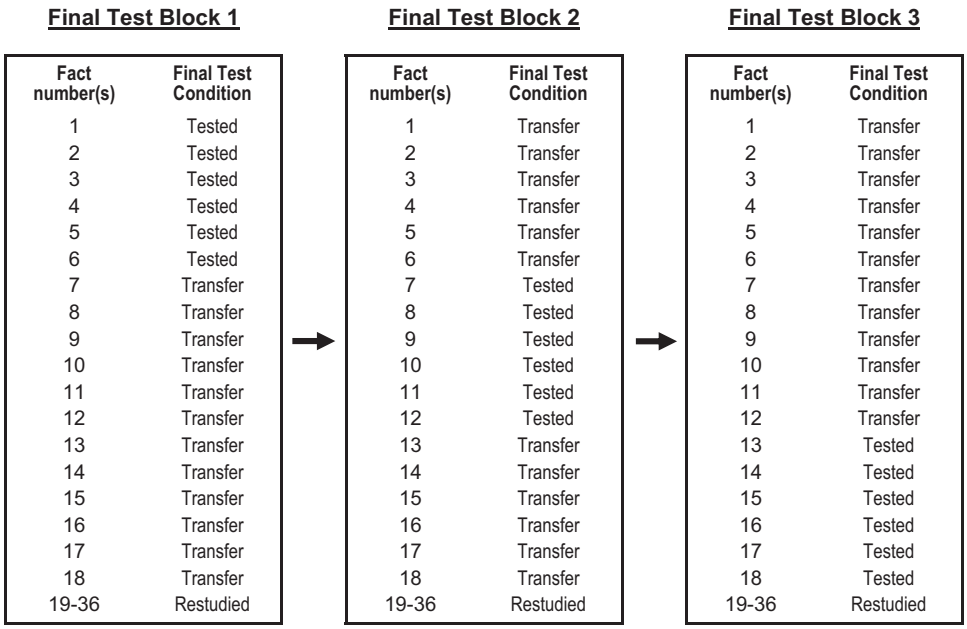

Figure 2. Example final test block design used in Experiments 1 to 3, with hypothetical fact numbers for illustrative purposes. During the final test in Session 2, one critical term from each fact was tested per block, and all 36 facts appeared once per block. All three terms per fact were tested over three blocks. Final test condition indicates whether the missing term was previously retrieved (tested), not previously retrieved, but from a tested fact (transfer), or from a fact that was not previously tested (restudied), during Session 1.

After completion of the experiment, subjects were asked whether they had taken an AP History course (on any history topic) in high school.

Data coding and analysis. The sole dependent variable in both the initial and final test analyses was accuracy. Because many of the missing terms are easy to misspell, two research assistants who were blind to question type checked subject responses for misspelled answers that could be unambiguously matched to correctly spelled answers; these were coded as correct responses. Here and in Experiments 2 and 3, separate analyses performed with misspelled words coded as incorrect responses yielded lower overall accuracy but no difference in relative performance in the tested, transfer, and restudied conditions.

Results and Discussion

Training phase performance. Overall accuracy on the initial test (involving 18 fill-in-the-blank test questions for each subject) was $M = 0.27$, $SE = 0.03$.

Final test performance. We performed a within-subjects factorial analysis of variance (ANOVA) on subject-level mean accuracy scores (see Figure 3) with factors of Final Test Condition (tested vs. transfer vs. restudied) and Block (1 vs. 2 vs. 3). In this and all subsequent analyses, alpha was set at 0.05. There were statistically significant effects of final test condition, $F(2, 74) = 25.90$, mean squared error (MSE) $= 0.027$, $p < .0001$, $\eta_p^2 = 0.41$, and block, $F(2, 74) = 41.88$, $MSE = 0.019$, $p < .0001$, $\eta_p^2 = 0.53$, but no significant final test Condition × Block interaction, $F(4, 148) = 1.21$, $MSE = 0.024$, $p = .31$. The significant improvement over blocks may reflect intermediate priming effects (e.g., Pan et al., in press); because each test question per block displayed the correct answers to test questions in subsequent blocks, answer priming effects may have occurred for the second and third blocks of the final test. Inspection of Figure 3 illustrates the main effect of final test condition: In all three blocks, proportion correct was higher in the tested condition than in either the restudy or transfer condition. A planned follow-up ANOVA limited to the transfer and restudied conditions yielded a nonsignificant effect of final test condition, $F(1, 37) = 2.75$, $MSE = 0.015$, $p = .11$, and no interaction between test condition and block, $F(2, 74) = 2.50$, $MSE = 0.010$, $p = .089$. Overall, the critical finding of Experiment 1 is that testing strongly enhances fact learning, but affords no, or only very limited, benefits to transfer questions relative to restudy.

Effect of prior AP experience. In exit surveys, nearly half ($n = 18$) of the subjects reported having previously taken an AP History course. Test performance during the training session was indistinguishable for AP and non-AP students, $t(30) = 1.49$, $p = .15$. Overall performance on the final test, however, was substantially better for AP students, $M = 0.40$ versus $M = 0.28$, $t(29) = 2.32$, $p = .0028$, $d = 0.77$, suggesting better retention for AP students. The overall pattern of no transfer relative to restudy on the final test held equivalently, however, for AP and non-AP students. The percent transfer (in which 0% means that transfer and restudy performance were equivalent, and 100% means that transfer and tested performance were equivalent) was −3% and −2% for AP and non-AP students, respectively; for both groups, performance on transfer questions was no better, and numerically worse, than that in the restudied condition.

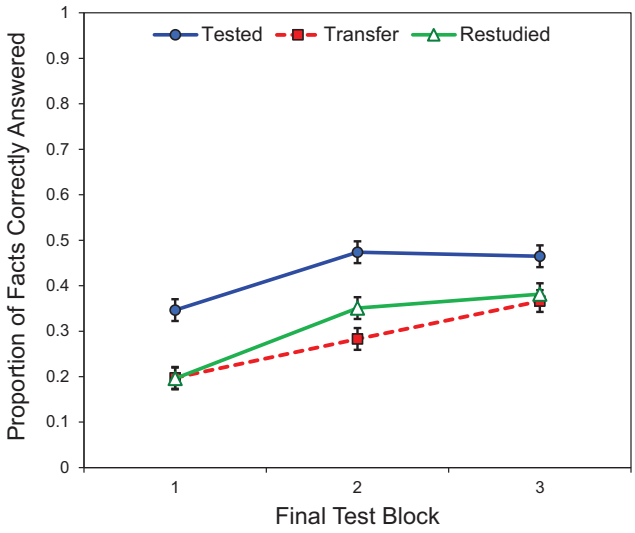

Figure 3. Results from Experiment 1: mean accuracy performance for history fact questions on the final test as a function of whether the missing term was previously retrieved (tested), not previously retrieved, but from a tested fact (transfer), or from a fact that was not previously tested (restudied). Error bars are standard errors based on the interaction error term of a within-subjects analysis of variance on subject mean accuracy scores (based on Loftus & Masson, 1994). See the online article for the color version of this figure.

Experiment 2

In the second experiment, we investigated whether the transfer properties of test-enhanced learning as established for history facts in Experiment 1 would extend to a physical science domain, namely, biology. As evident from a comparison of AP Biology and AP History exams, the structure of biology facts differs substantially from history facts (e.g., reduced frequency of *who* and *when* type terms and more terms referring to objects and processes). It is thus possible that test-enhanced learning for science facts will yield better transfer than it does for history facts.

Method

Subjects. Minimum sample size to detect a modest transfer effect was determined using a priori power analysis. Based on the standard deviation of the final test transfer minus restudied condition proportion correct difference scores in Experiment 1, a sample size of at least 47 is needed to achieve power of 0.95 to detect a mean proportion correct difference score of 0.05 or greater (based on a one-tailed, one-sample *t* test, $\alpha = .05$). When subject sign-up and data collection were completed, 58 undergraduates had participated for course credit. Subject ages were $M = 21.12$, $SD = 3.072$, and ranged from 17 to 38 years. Two thirds (67%) of the sample was female. All subjects finished both sessions of the study.

Materials. Thirty-six biology facts were obtained using the AP Biology preparatory texts produced by *Barron's* (Goldberg, 2014) and *Princeton Review* (Magloire, 2014). The biology facts had the same defining characteristics (e.g., at least three one-word critical terms each, average 11 words in length) as the history facts from Experiment 1. However, unlike in the preceding experiment, the information covered by each fact did not include geopolitical concepts, persons, or unique historical events. An example biology fact (with critical terms in italics) is "The *Krebs* cycle occurs in the *mitochondria* and produces *ATP*." As before, a fill-in-the-blank test question and a short-answer test question were created for each of three critical terms per fact (additional examples are included in the Appendix).

Design and procedure. Two sessions occurred in the same manner as in Experiment 1, with the following exception: During Session 1, after training on all 36 biology facts, subjects trained on all 36 facts for a second time using identical procedures (i.e., the same fill-in-the-blank question was shown for each tested fact, and the complete fact was displayed for each restudied fact). Hence, for tested facts, subjects were asked to retrieve the same missing term twice. This modification was motivated by pilot tests, which suggested that the biology facts may be more difficult than the history facts.

Results and Discussion

Training phase performance. Mean accuracy on the initial test was $M = 0.37$, $SE = 0.03$ for the first training block, and $M = 0.64$, $SE = 0.03$ for the second training block. The improvement between blocks was statistically significant, $t(57) = 13.63$, $p < .0001$, $d = 1.79$.

Final test performance. A within-subjects factorial ANOVA on subject-level mean accuracy scores (see Figure 4) with factors of Final Test Condition (tested vs. transfer vs. restudied) and Block (1 vs. 2 vs. 3) found statistically significant effects of final test condition, $F(2, 114) = 57.46$, $MSE = 0.032$, $p < .0001$, $\eta_p^2 = 0.50$,

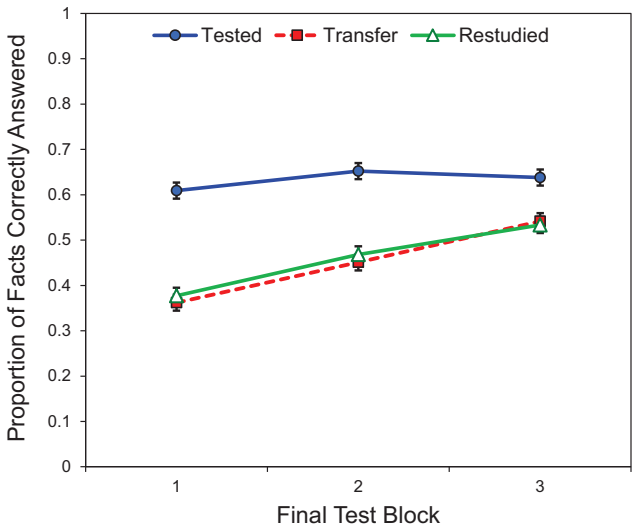

Figure 4. Results from Experiment 2, in which subjects trained twice per fact during training: mean accuracy performance for biology fact questions on the final test as a function of whether the missing term was previously retrieved (tested), not previously retrieved, but from a tested fact (transfer), or from a fact that was not previously tested (restudied). Error bars are standard errors based on the interaction error term of a within-subjects analysis of variance on subject mean accuracy scores (based on Loftus & Masson, 1994). See the online article for the color version of this figure.

and block, $F(2, 114) = 38.27$, $MSE = 0.017$, $p < .0001$, $\eta_p^2 = 0.40$, and a significant Final Test Condition × Block interaction, $F(4, 228) = 5.60$, $MSE = 0.018$, $p = .00026$, $\eta_p^2 = 0.090$. The interaction corresponds to an increased pattern of improvement for transfer and restudied questions across blocks; however, even by the final block, performance on tested questions is still greater than for either transfer or restudied questions by $M = 0.10$, $SE = 0.03$. As in Experiment 1, a follow-up ANOVA limited to the transfer and restudied conditions found no significant effect of final test condition, $F(1, 57) = 0.26$, $MSE = 0.022$, $p = .61$, a significant effect of block, $F(2, 114) = 61.62$, $MSE = 0.013$, $p < .0001$, $\eta_p^2 = 0.52$, and no significant Block × Condition interaction, $F(2, 114) = 0.40$, $MSE = 0.014$, $p = .67$. These results replicate the findings of Experiment 1: Testing strongly benefits fact learning relative to restudy, but in a manner that is specific to the exact tested term.

Effect of prior AP experience. There was a trend toward better training-phase test performance for AP ($n = 15$) than for non-AP subjects, $M = 0.28$ versus $M = 0.16$, $t(34) = 1.85$, $p = .07$, $d = 0.50$. Overall performance on the final test was again significantly higher for AP subjects, $M = 0.71$ versus $M = 0.45$, $t(38) = 6.27$, $p < .0001$, $d = 1.48$, suggesting better retention for AP students. The overall pattern of no transfer relative to restudy held equivalently, however, for AP and non-AP students, with the percent transfer being −6% and 1%, respectively.

Experiment 3

The third experiment investigated whether specificity of test-enhanced learning would manifest when multiple-choice test questions, rather than fill-in-the-blank test questions, are used during training. Multiple-choice questions are conveniently graded and are frequently used in educational contexts. The structure of multiple-choice questions often requires careful consideration of, and adjudication between, several possible responses.

Method

Subjects. Fifty-eight undergraduates participated for course credit. Subject ages were $M = 20.37$, $SD = 2.38$, and ranged from 18 to 28 years. The majority (83%) of the sample was female. Six subjects did not return to complete the second session; data from the 52 remaining subjects was analyzed.

Materials. The 36 history facts from Experiment 1 served as stimuli for Experiment 3. The only change was the format of test questions during training: Instead of fill-in-the-blank test questions, multiple-choice test questions were used (examples are included in the Appendix). Each multiple-choice test question covered one of the three possible missing terms per fact. There were four one-word answer choices per multiple-choice test question (A, B, C, or D; one correct answer and three distractors). Each distractor was designed to be competitive enough to warrant consideration by subjects (Little & Bjork, 2015; Little, Bjork, Bjork, & Angello, 2012).

Design and procedure. Two sessions occurred in the same manner as in the prior experiments, except for the use of multiple-choice test questions during Session 1. For each multiple-choice test question, subjects were given 6 s to select one of the four possible answer choices; the correct answer was subsequently shown for 2 s. There was one training trial per fact.

Results and Discussion

Training phase performance. Mean accuracy on the initial test (18 multiple-choice test questions) was $M = 0.73$, $SE = 0.03$.

Final test performance. A within-subjects factorial ANOVA on subject-level mean accuracy scores (see Figure 5) with factors of Final Test Condition (tested vs. transfer vs. restudied) and Block (1 vs. 2 vs. 3) yielded statistically significant effects off Test Condition, $F(2, 102) = 34.19$, $MSE = 0.94$, $p < .0001$, $\eta_p^2 = 0.40$, Block, $F(2, 102) = 67.22$, $MSE = 1.73$, $p < .0001$, $\eta_p^2 = 0.57$, and no significant Final Test Condition × Block interaction, $F(4, 204) = 1.70$, $MSE = 0.037$, $p = .15$. A follow-up ANOVA limited to the transfer and restudied conditions found no significant effect of final test condition, $F(1, 51) = 2.48$, $MSE = 0.067$, $p = .12$, a significant effect of block, $F(2, 102) = 69.59$, $MSE = 1.43$, $p < .0001$, $\eta_p^2 = 0.58$, and no significant Block × Condition interaction, $F(2, 102) = 1.86$, $MSE = 0.024$, $p = .16$. These results replicate and extend the findings of the prior experiments: Multiple-choice testing strongly benefits fact learning, but only for the tested term.

Little et al. (2012; see also Little & Bjork, 2015), who also used multiple-choice tests, had distractors during training tests reappear as correct answers to final transfer test questions. They observed partial transfer of learning for multiple-choice training tests relative to cued recall training tests (for which no distractors were shown). Little et al. suggested that the act of considering plausible distractors that were answers to later test questions was the driver of improved transfer performance. This contrasts with the present experiment, in which distractors used on multiple-choice training test questions did not reappear as correct answers to transfer questions on the final test.

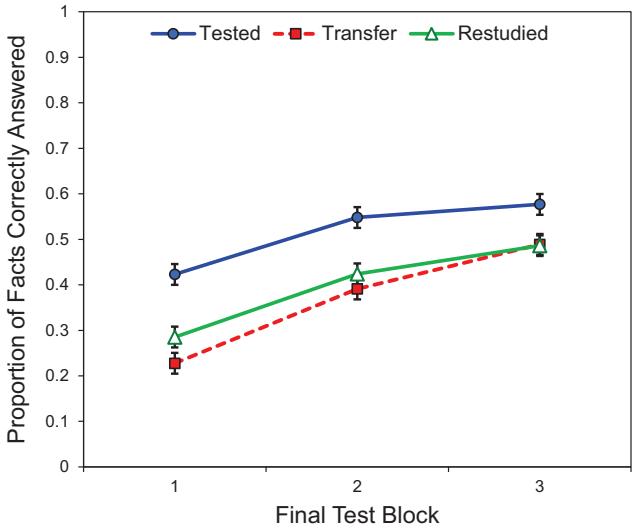

Figure 5. Results from Experiment 3, in which multiple-choice test questions were used during training: mean accuracy performance for history fact questions on the final test as a function of whether the missing term was previously retrieved (tested), not previously retrieved, but from a tested fact (transfer), or from a fact that was not previously tested (restudied). Error bars are standard errors based on the interaction error term of a within-subjects analysis of variance on subject mean accuracy scores (based on Loftus & Masson, 1994). See the online article for the color version of this figure.

Effect of prior AP experience. There was no difference in mean test performance during the training session for AP ($n = 26$) than for non-AP subjects, $t(49) = 1.30$, $p = .20$. Overall performance on the final test was also indistinguishable in this case, $M = 0.41$ versus $M = 0.40$, $t(49) = 0.18$, $p = .86$. Percent transfer was 1% and −5% for AP and non-AP students, respectively.

Experiment 4

The fourth experiment investigated the effects of training on more than one critical term per fact. For example, given the fact (with critical terms italicized), "*Jefferson* purchased *Louisiana* from the *Spanish*," it is possible to test each critical term separately with a different test question (e.g., by asking about *who*, *what*, and *whom* regarding that fact). This design addressed two questions. First, does training on two critical terms per fact promote transfer to a third term on the final test? If testing with feedback on the second term during training reactivates memory for the preceding test on the first term, then the second test may result in a more integrated, or holistic, representation that will promote transfer to a third term on the final test. Second, are there increasing or diminishing returns from expanding the number of answer terms that are tested during training? We approached these questions by manipulating the number of different answer terms that subjects were asked to retrieve per fact during training (one, two, or three), and, correspondingly, the number of times that subjects were asked to restudy facts during training.

Method

Subjects. Fifty-two undergraduates participated for course credit. Subject ages were $M = 19.91$, $SD = 1.68$, and ranged from 17 to 26 years. The majority (73%) of the sample was female. All but seven subjects completed both sessions of the experiment; data from the remaining 45 subjects was analyzed.

Materials. Fifty-four AP United States History and AP World History facts were used for this experiment. This included the 36 history facts from Experiments 1 and 3, along with 18 additional facts that were extracted from the same sources. More facts were necessary because of the different training conditions in Session 1. The extra facts met the same selection criteria as in prior experiments, and three fill-in-the-blank and three short-answer test questions were also created for each fact.

Design and procedure. Two sessions occurred in the same manner as Experiment 1, with the following modifications. In the training phase of Session 1, of the 27 facts assigned to testing, one third each were assigned to have one, two, or three critical terms tested. Of the 27 facts assigned to be restudied, a corresponding number were restudied once, twice, or three times. Assignment of facts to restudy or testing, as well as to one, two, or three test or restudy trials during training, was counterbalanced across subjects. As illustrated in Figure 6, the training phase featured three blocks of 36 trials each; within each block, there were six trials for the nine facts that were tested on one term or restudied once across the three blocks, 12 trials for the nine facts that were tested on two

Training Block 1

Fact number(s)	Training Condition
1	1 term tested
2	1 term tested
3	1 term tested
10	2 terms tested
11	2 terms tested
12	2 terms tested
13	2 terms tested
14	2 terms tested
15	2 terms tested
19-27	3 terms tested
28	1 restudy trial
29	1 restudy trial
30	1 restudy trial
37	2 restudy trials
38	2 restudy trials
39	2 restudy trials
40	2 restudy trials
41	2 restudy trials
42	2 restudy trials
46-54	3 restudy trials

→

Training Block 2

Fact number(s)	Training Condition
4	1 term tested
5	1 term tested
6	1 term tested
16	2 terms tested
17	2 terms tested
18	2 terms tested
13	2 terms tested
14	2 terms tested
15	2 terms tested
19-27	3 terms tested
31	1 restudy trial
32	1 restudy trial
33	1 restudy trial
43	2 restudy trials
44	2 restudy trials
45	2 restudy trials
40	2 restudy trials
41	2 restudy trials
42	2 restudy trials
46-54	3 restudy trials

→

Training Block 3

Fact number(s)	Training Condition
7	1 term tested
8	1 term tested
9	1 term tested
10	2 terms tested
11	2 terms tested
12	2 terms tested
16	2 terms tested
17	2 terms tested
18	2 terms tested
19-27	3 terms tested
34	1 restudy trial
35	1 restudy trial
36	1 restudy trial
37	2 restudy trials
38	2 restudy trials
39	2 restudy trials
43	2 restudy trials
44	2 restudy trials
45	2 restudy trials
46-54	3 restudy trials

Figure 6. Example training block design used in Experiment 4, Session 1, with hypothetical fact numbers for illustrative purposes. For tested facts, a different missing term was assessed on each test trial across all three blocks. For restudied facts, the entire fact was presented on each restudy trial. Of the 54 facts in this experiment, nine each were assigned to the one-, two-, and three-tested term or restudy trial training conditions. In each 36-trial block, equal distribution of facts in the different training conditions was accomplished by presenting three facts in the one-term-tested condition, six facts in the two-terms-tested condition, and nine facts in the 12-terms-tested condition. A corresponding number were presented in each of the restudied conditions.

terms or restudied twice across the three blocks, and 18 trials for the nine facts that were tested on three terms or restudied three times across the three blocks. Fact presentation was randomly ordered within each block, no fact was repeated within each block, and there were no breaks between blocks. By the end of the three training blocks, each third of the 54 facts had been tested or restudied once, twice, or three times. Session 1 ended with a reminder for subjects to return the following day to complete the second part of the experiment.

Session 2, the final test, occurred 24 hr after training and assessed subjects' recall of the entire set of 162 answer terms across all 54 history facts. This involved three 54-trial blocks that were based on the design of the final test in Experiments 1 to 3. As in the prior experiments, each fact appeared once per block, and a different critical term was assessed per fact and per block without repeats. In each block, there were nine facts each per training condition (one, two, or three tested terms or restudy opportunities per block). Of the nine facts that had been tested once during training, there were three tested and six transfer questions per final test block. Of the nine facts that had been tested twice during training, there were six tested and three transfer questions per final test block.

Results and Discussion

Training phase performance. Mean proportion correct across all three training blocks of the initial test was $M = 0.31$ ($SE = 0.040$) for facts that had one term tested, $M = 0.39$ ($SE = 0.031$) for facts that had two terms tested, and $M = 0.40$ ($SE = 0.032$) for facts that had three terms tested. A within-subjects one-way ANOVA yielded a significant effect of terms trained (i.e., facts that were trained on one, two, or three terms across the three training blocks), $F(2, 88) = 6.76$, $MSE = 0.098$, $p = .0019$, $\eta_p^2 = 0.13$. In the context of correct-answer feedback, which allows for restudy of the entire fact on each test trial regardless of accuracy, the improved performance over the number of terms trained was expected.

Final test performance. We first performed a within-subjects factorial ANOVA on subject-level mean accuracy scores (see Figure 7) with factors of Final Test Condition (tested vs. transfer vs. restudied), Block (1 vs. 2 vs. 3), and Number of Trained Terms or Restudy Trials (one vs. two). As there were no transfer questions for facts that had three terms tested during training, this analysis was confined to data from the facts that had two tested terms or two restudy trials. The ANOVA yielded statistically significant effects of final test condition, $F(2, 88) = 35.32$, $MSE = 2.27$, $p < .0001$, $\eta_p^2 = 0.44$, and block, $F(2, 88) = 28.78$, $MSE = 1.53$, $p < .0001$, $\eta_p^2 = 0.40$, no significant effect of the number of trained terms or restudy trials, $F(1, 44) = 1.94$, $MSE = 0.11$, $p = .17$, and no significant interaction effects ($ps > .17$). A follow-up ANOVA limited to the transfer and restudied conditions (and again limited to facts that were tested on two terms or had two restudy trials) yielded no significant effect of final test condition, $F(1, 44) = 3.17$, $MSE = 0.14$, $p = .082$, a significant effect of block, $F(2, 88) = 28.39$, $MSE = 1.35$, $p < .0001$, $\eta_p^2 = 0.39$, no significant effect of trained terms or restudy trials, $F(1, 44) = 3.19$, $MSE = 0.21$, $p = .081$, and no significant interactions ($ps > .35$). These results replicate the findings of no transfer relative to restudy observed in the prior experiments. They also extend those

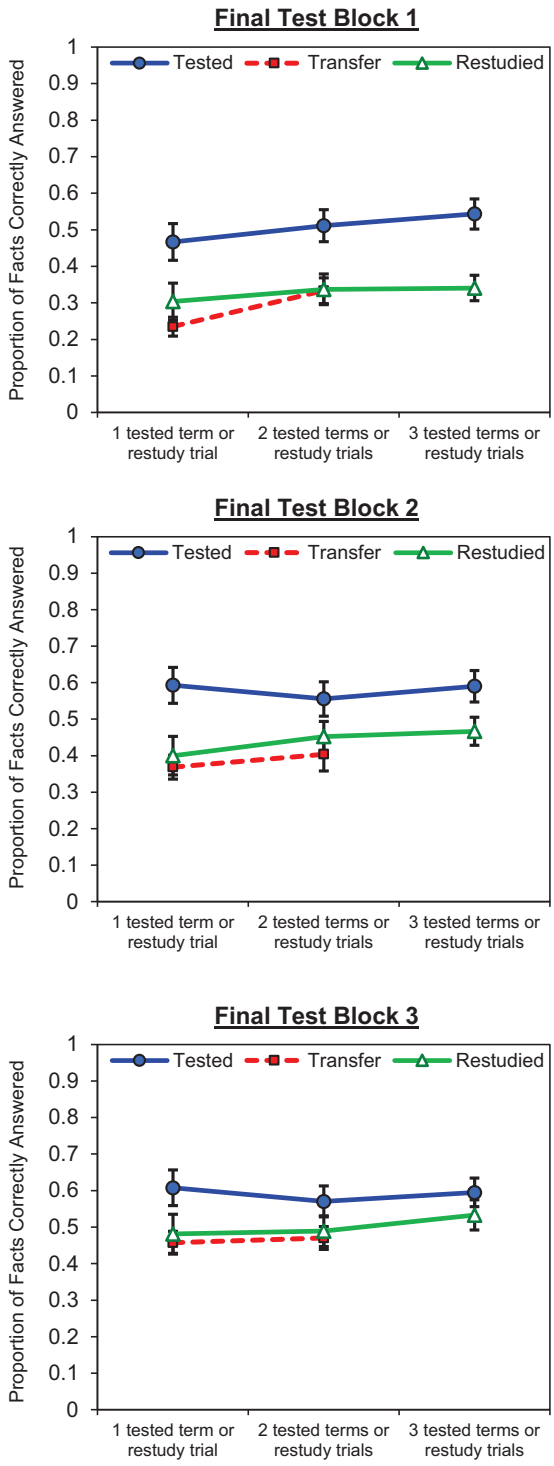

Figure 7. Results from Experiment 4: mean accuracy performance for history fact questions by training condition (one, two, or three tested terms or restudy trials) per final test block as a function of whether the missing term was previously retrieved (tested), not previously retrieved, but from a tested fact (transfer), or from a fact that was not previously tested (restudied). Error bars are standard error of the mean. See the online article for the color version of this figure.

findings in an important respect: Even when two terms are testing during training, there is no transfer of test-enhanced learning to a previously untested term.

To assess the effect of expanding the number of answer terms tested during training on overall acquired knowledge for each fact, we performed a within-subjects factorial ANOVA on mean accuracy scores (see Figure 8) from the first final test block, with factors of Final Test Condition (*combined* tested and transfer vs. restudied) and Number Of Tested Terms or Restudy Trials (1 vs. 2 vs. 3). Blocks 2 and 3 were not included in this analysis because the overall pattern of results in Block 1 was replicated in those blocks in the prior analyses, and because Block 1 provides the purest and most sensitive measure of training effects. By combining the tested and transfer conditions in this analysis, we were able to assess the effects of testing on one, two, or three terms per fact on memory for the entire fact. That approach to data analysis directly addresses the educationally important question of whether testing with feedback yields diminishing, constant, or increasing learning effects for the fact as a whole, as a function of the number of terms tested.

There were statistically significant effects of final test condition, $F(1, 44) = 36.49$, $MSE = 1.38$, $p < .0001$, $\eta_p^2 = 0.45$, and number of trained terms, $F(2, 88) = 20.84$, $MSE = 0.58$, $p < .0001$, $\eta_p^2 = 0.32$. The apparent interaction between those factors in Figure 8, however, did not reach statistical significance, $F(2, 88) = 2.33$, $MSE = 0.12$, $p = .10$. Thus, with respect to fact memory as a whole, test-enhanced learning appears to increase linearly as a function of the number of trained terms, whereas repeated restudy may have diminishing returns. The linear increase in proportion correct for tested items is consistent with the hypothesis that learning in the training phase occurs independently for each question-term combination of a fact (for related discussion, see Pan et al., in press).

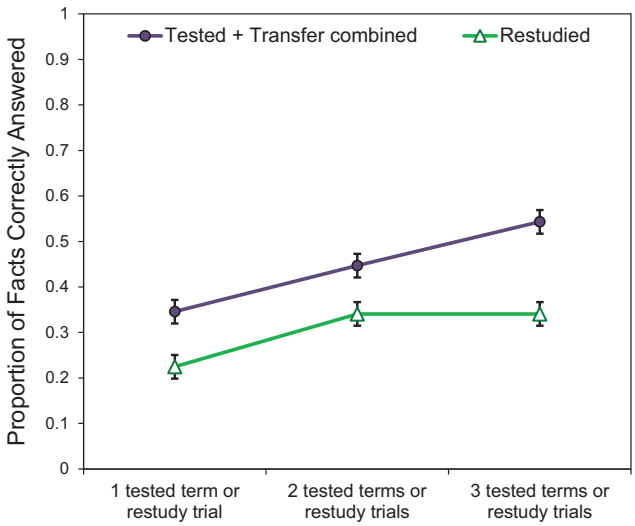

Figure 8. Experiment 4 mean accuracy performance for history fact questions by training condition (one, two, or three tested terms or restudy trials) on the first block of the final test, as a function of whether the fact was previously tested at least once (tested + transfer questions combined) or not previously tested at all (restudied). Error bars are standard errors based on the interaction error term of a within-subjects analysis of variance on subject mean accuracy scores (based on Loftus & Masson, 1994). See the online article for the color version of this figure.

Effect of prior AP experience. There was significantly better training phase test performance for AP ($n = 14$) than for non-AP subjects, $M = 0.46$ versus $M = 0.34$, $t(37) = 2.07$, $p = .046$, $d = 0.61$. Overall performance on the final test was also significantly higher for AP subjects, $M = 0.60$ versus $M = 0.44$, $t(37) = 32.26$, $p = .0090$, $d = 0.86$. Extending the results from the preceding experiments, the percent transfer, averaged over all final test blocks and across the one, two, and three items tested or restudy trials training conditions, was −9% and −5% for AP and non-AP students, respectively.

General Discussion

The foregoing experiments assessed whether testing on one or more critical terms of a fact benefits long-term memory for the entire fact. The results demonstrate that, under educationally important circumstances, testing can produce potent, but piecewise, fact learning. Despite consistently large testing effects (38%, 38%, 29%, and 41% higher proportion correct for testing vs. restudy in Experiments 1, 2, 3, and 4, respectively), those effects were entirely specific to tested terms, and there was no positive transfer to untested terms relative to restudy. Indeed, in each experiment, there was a nonsignificant trend toward negative transfer. That lack of transfer held for both history and biology facts, when either fill-in-the-blank or multiple-choice questions were used during training, when a single term was tested twice during training (Experiment 2), and when more than one term was tested during training (Experiment 4). Accordingly, we expect that these findings will generalize to other subject materials and types of training tests.

Conditions of Transfer and Theoretical Implications

As noted in the introduction, the results from the prior literature on transfer between terms of facts are mixed, with two of four studies demonstrating positive transfer relative to restudy. We suggest that those contrasting results can be integrated by two principles. First, if there is extensive feedback about the entire fact between training and the final test, positive transfer may be observed. In McDaniel et al. (2007), subjects had unlimited opportunities to view online feedback without any time limit over a period of 1 week, in preparation for a later test. That extended feedback is likely to have been processed most thoroughly for incorrectly answered test questions, and that processing, which constitutes a form of restudy, may have yielded positive transfer relative to restudy on the final test. By that account, the event of testing in itself was not the primary basis for transfer in that report; rather, it is additional learning that was the result of further study opportunities. Second, positive transfer may occur when facts have an explicit term-definition structure, as in McDaniel et al. (2013). If subjects have a learning goal of associating a term and a definition (perhaps because they expect a test of that type), then testing's benefits may be analogous to that for paired associates, which also exhibit positive transfer (Carpenter, Pashler, & Vul, 2006). In other words, if a given fact can be logically organized as two components, it may be the case that such a structure is more easily learned, and hence more transferrable, than more complex

materials. We have neither evidence nor intuition about that possibility, but rather note it here as a plausible integrative account. Finally, if neither of those conditions holds, as in the current experiments, the Pan et al. (in press) experiments, and the Hinze and Wiley (2011) experiments, then no transfer will be observed relative to restudy. Further research to test those hypotheses is needed.

From the broader theoretical perspective, the current results support specificity-of-learning accounts of memory representations following retrieval practice for multiterm stimuli, such as the identical elements model (Rickard & Bajic, 2006; Rickard & Bourne, 1996; Rickard, Healy, & Bourne, 1994). That model, originally developed to account for response time gains following training on arithmetic facts, holds that successful retrieval after initial study establishes a new and separate recall representation for that particular stimulus–response configuration. That memory representation is only accessible when the stimulus–response configuration that was previously trained is presented; for instance, a particular set of operands in a multiplication fact (e.g., Rickard et al., 1994) or specific words from a word triplet (e.g., Pan et al., in press). The current experiments add a more complex category of stimuli, namely, multielement facts, to the list of materials that are subject to the specificity principles outlined by the identical elements model. It remains to be determined, however, whether the same results would apply if the to-be-learned materials are different from the facts used in the current study (such as, e.g., conceptual knowledge).

The current results also suggest that transfer of test-enhanced learning is not substantially modulated by expertise level, at least for multiterm facts. In three of the four experiments, AP students performed significantly and substantially better overall on the final test than did non-AP students, confirming the expected higher expertise. However, there were no significant differences in transfer results for those two subject groups. Indeed, performance in the transfer condition for both groups was slightly poorer than performance in the restudy condition in all but one instance (AP students: −3%, −6%, −1%, and −9%; non-AP students: −2%, 1%, −5%, and −5%, for Experiments 1, 2, 3, and 4, respectively). Based on these results, higher expertise in fact domains apparently does not result in more integrated processing during the initial test that could support transfer.

A caveat to that conclusion is that specific expertise for the assessed AP facts was not directly manipulated in the current experiments. In future studies, the role of expertise on transfer could be further examined by providing additional time in one condition to review relevant or related subject matter prior to the training phase (thus facilitating greater expertise and contextual knowledge), or by presenting related facts sequentially in one condition, as, for example, in Chan (2009, Experiment 1).

The trend toward poorer performance in the transfer condition than in the restudy condition, although not significant, raises the possibility of negative transfer. Anderson, Bjork, and Bjork (1994) and others have also observed negative transfer following testing using the retrieval-induced forgetting paradigm. That paradigm differs markedly from that of the current study, however. In the case of retrieval-induced forgetting, testing on specific category exemplars typically results in poorer recall of other category exemplars, relative to exemplars from categories that were not tested or restudied at all, and often only after a relatively short (<24 hr) delay interval (Chan, 2009). In the current study, the terms of each tested fact were from different categories, transfer was compared against restudy rather than no reexposure, and longer delays were used. It is thus difficult to draw inference about possible relations between those paradigms.

Practical Implications for Testing on Facts

The present transfer results are educationally important, given that a large portion, and perhaps the majority, of facts that learners encounter in many educational contexts are multiterm facts rather than paired-associate or term-definition type materials. AP facts exemplify this pattern. In the latest AP World History, United States History, and Biology practice tests (College Board, 2011, 2012, 2013b), assessed facts have, by our analysis, an average of four testable terms, and few have an explicit paired-associate or term-definition structure (although clearly such facts are also encountered in other educational contexts). Multiterm facts can also be found in a wide variety of high school, college, technical, and other courses. Thus, knowledge of the transfer properties of multiterm facts is essential to efforts to optimize learning in applied contexts.

Of immediate impact to learners, our results indicate that instructors should not expect that one test question will typically be sufficient to memorize an entire fact. Unless training tests target multiple (or in the best-case scenario, all) terms of each fact, a selective memory benefit can result. This finding is particularly important given that practice tests—for instance, those at the end of textbook chapters—often assess factual knowledge with a single fill-in-the-blank, short-answer, or multiple-choice question per fact. Fortunately, redesigning practice tests to solve this problem is a relatively simple task: Ask more than one question per fact, and ensure that questions comprehensively target all parts of each fact. The results from Experiment 4 indicate that learning rate does not diminish as the number of tested terms is increased.

Although the present experiments were not focused on a comparison of test formats, the results from Experiments 1 and 3, which were identical except for the format of test questions during training (fill-in-the-blank vs. multiple-choice), indicate that both types of tests yield potent testing benefits, but that fill-in-the-blank questions with feedback may produce a larger benefit for tested terms (9% greater mean difference in proportion correct vs. restudy). That apparent disparity in final test performance is consistent with prior findings of larger testing effects following training with short-answer versus multiple-choice tests (e.g., Butler & Roediger, 2007; Duchastel, 1981; McDaniel et al., 2007).

Maximizing Test-Enhanced Fact Learning

Future work may examine whether there are training conditions that reliably produce transfer of test-enhanced fact learning from one term to another, thus increasing the degree to which tests benefit fact learning. Allowing more extensive feedback, as in McDaniel et al. (2007), is one possible strategy. Instructional strategies that promote more integrative processing (e.g., Chan, 2009; Chan et al., 2006) are another option.

From the perspective of optimizing learning efficiency, however, the extra time required for such approaches, relative to that required for testing on a variety of answer terms, must be factored into any systematic evaluation. Moreover, until such techniques are consistently demonstrated, a strategy of testing on as many parts of each fact as possible stands as the most promising solution for successful fact learning.

References

Agarwal, P. K. (2011). Examining the relationship between fact learning and higher order learning via retrieval practice (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Order No. 3468823)

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1063–1087. http://dx.doi.org/10.1037/0278-7393.20.5.1063

Armstrong, M., Daniel, D., Kanarek, A., & Freer, A. (2014). *The Princeton Review: Cracking the AP World History exam*. New York, NY: Random House.

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

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., 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. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21,* 279–283. http://dx.doi.org/10.1177/0963721412452728

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology, 23,* 760–771. http://dx.doi.org/10.1002/acp.1507

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13,* 826–830. http://dx.doi.org/10.3758/BF03194004

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633–642. http://dx.doi.org/10.3758/BF03202713

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61,* 153–170. http://dx.doi.org/10.1016/j.jml.2009.04.004

Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory (Hove, England), 18,* 49–57. http://dx.doi.org/10.1080/09658210903405737

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135,* 553–571.

College Board. (2011). *AP World History: Practice exam and notes*. New York, NY: Author.

College Board. (2012). *AP Biology: Practice exam and notes*. New York, NY: Author.

College Board. (2013a). Advanced placement program student score distributions. Retrieved from http://media.collegeboard.com/digitalServices/pdf/research/2013/STUDENT-SCORE-DISTRIBUTIONS-2013.pdf

College Board. (2013b). *AP United States History: Practice exam*. New York, NY: Author.

Duchastel, P. C. (1981). Retention of prose following testing with different types of tests. *Contemporary Educational Psychology, 6,* 217–226. http://dx.doi.org/10.1016/0361-476X(81)90002-3

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4–58. http://dx.doi.org/10.1177/1529100612453266

Foos, P. W., & Fisher, R. P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology, 80,* 179–183. http://dx.doi.org/10.1037/0022-0663.80.2.179

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives de Psychologie, 6* (40).

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81,* 392–399. http://dx.doi.org/10.1037/0022-0663.81.3.392

Goldberg, D. (2014). *Barron's AP Biology* (4th ed.). Hauppauge, NY: Barron's.

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory (Hove, England), 19,* 290–304. http://dx.doi.org/10.1080/09658211.2011.560121

Kang, S. H. K., 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., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331,* 772–775. http://dx.doi.org/10.1126/science.1199327

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition, 43,* 14–26. http://dx.doi.org/10.3758/s13421-014-0452-8

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science, 23,* 1337–1344. http://dx.doi.org/10.1177/0956797612443370

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1,* 476–490. http://dx.doi.org/10.3758/BF03210951

Magloire, K. (2014). *The Princeton Review: Cracking the AP Biology Exam*. New York, NY: Random House.

McCannon, J. (2014). *Barron's AP World History* (6th ed.). Hauppauge, NY: Barron's.

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. http://dx.doi.org/10.1080/09541440701326154

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science, 20,* 516–522. http://dx.doi.org/10.1111/j.1467-9280.2009.02325.x

McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14,* 200–206. http://dx.doi.org/10.3758/BF03194052

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

Meltzer, T., & Bennett, J. H. (2014). *The Princeton Review: Cracking the AP U.S. History Exam*. New York, NY: Random House.

Pan, S. C., Wong, C., Potter, Z., Mejia, J., & Rickard, T. C. (in press). Does test-enhanced learning transfer for triple associates? *Memory & Cognition*. http://dx.doi.org/10.3758/s13421-015-0547-x

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learn-*

ing (NCER 2007–2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Resnick, E. (2014). *Barron's AP United States History* (2nd ed.). Hauppauge, NY: Barron's.

Rickard, T. C., & Bajic, D. (2006). Cued recall from image and sentence memory: A shift from episodic to identical elements representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 734–748.

Rickard, T. C., & Bourne, L. E., Jr. (1996). Some tests of an identical elements model of basic arithmetic skills. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1281–1295. http://dx.doi.org/10.1037/0278-7393.22.5.1281

Rickard, T. C., Healy, A. F., & Bourne, L. E. (1994). On the cognitive structure of basic arithmetic skills: Operation, order, and symbol transfer effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1139–1153. http://dx.doi.org/10.1037/0278-7393.20.5.1139

Roediger, H. L., III, & Karpicke, J. D. (2006). 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

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 233–239. http://dx.doi.org/10.1037/a0017678

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. http://dx.doi.org/10.1037/a0037559

Tran, R., Rohrer, D., & Pashler, H. (2015). Retrieval practice: The lack of transfer to deductive inferences. *Psychonomic Bulletin & Review, 22,* 135–140. http://dx.doi.org/10.3758/s13423-014-0646-x

Appendix

Training and Final Test Question Examples

Experiment	Subject	Type	Example
1,4	AP History	Training fact	An edict signed by Henry in Nantes gave rights to Calvinists.
		Fill-in-the-blank	An edict signed by _____ in Nantes gave rights to Calvinists.
			An edict signed by Henry in _____ gave rights to Calvinists.
			An edict signed by Henry in Nantes gave rights to _____.
		Short answer	An edict signed by WHO in Nantes gave rights to Calvinists?
			WHERE did Henry sign an edict giving rights to Calvinists?
			An edict signed by Henry in Nantes gave rights to WHOM?
2	AP Biology	Training fact	During glycolysis, sugar is broken down into pyruvate.
		Fill-in-the-blank	During glycolysis, sugar is broken down into _____.
			During _____, sugar is broken down into pyruvate.
			During glycolysis, _____ is broken down into pyruvate.
		Short answer	Sugar is broken down into pyruvate during WHAT?
			During glycolysis, WHAT is broken down into pyruvate?
			During glycolysis, sugar is broken down into WHAT?
3	AP History	Fact	Darrow defended Scopes over his teaching of evolution.
		Multiple-choice	_____ defended Scopes over his teaching of evolution. a. Bryan c. Hunter b. Darrow d. Butler

(*Appendix continues*)

Appendix (*continued*)

Experiment	Subject	Type	Example
			Darrow defended _____ over his teaching of evolution. a. Scopes c. Hicks b. Stewart d. Raulston
			Darrow defended Scopes over his teaching of _____. a. religion c. evolution b. creationism d. segregation
		Short answer	WHO defended Scopes over his teaching of evolution?
			Darrow defended WHOM over his teaching of evolution?
			Darrow defended Scopes over his teaching of WHAT?

Note. AP = Advanced Placement.

Received March 27, 2015
Revision received June 29, 2015
Accepted June 30, 2015 ■