Does Retrieval Practice Enhance Learning and Transfer Relative to Restudy for Term-Definition Facts?

Steven C. Pan

Timothy C. Rickard

University of California, San Diego

Word count (main text and references): 10,260

This manuscript was accepted for publication in the *Journal of Experimental Psychology: Applied* on February 3, 2017. This document may not exactly replicate the final version published in the APA journal. It is not the copy of record. The final version is available at: http://dx.doi.org/10.1037/xap0000124

Author Note

Steven C. Pan, Department of Psychology, University of California, San Diego; Timothy C. Rickard, Department of Psychology, University of California, San Diego.

The authors thank Arpita Gopal, Bijan Malaklou, Darian Parsey, Jason Chang, Kayla Hartman, and Maxim Deinitchenko for their assistance with data collection, Anastasia Bogozova, Carissa Gaw, Courtney Lukitsch, Daanish Unwalla, Danielle Emmar, Dominic D'Andrea, Jahan Tajran, Johnny Barry, Ikjot Thind, Michelle Hickman, Sarah Hutter, Thomas

Ting, and Yangyang Liu for assistance with coding data for Experiments 3b and/or 4, and Trinity Carwile for assistance with materials development for Experiment 4.

This research was supported by an American Psychological Association (APA) Early Graduate Student Researcher Award and a National Science Foundation (NSF) Graduate Research Fellowship to S. C. Pan.

Please address correspondence to: Timothy C. Rickard, Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109. Email: trickard@ucsd.edu Phone: 858-822-0122; Fax: 858-534-7190

Abstract

In many pedagogical contexts, *term-definition* facts which link a concept term (e.g., "*vision*") with its corresponding definition (e.g., "*the ability to see*") are learned. Does retrieval practice involving retrieval of the term (given the definition) or the definition (given the term) enhance subsequent recall, relative to restudy of the entire fact? Moreover, does any benefit of retrieval practice for the term transfer to later recall of the definition, or vice versa? We addressed those questions in four experiments. In each, subjects first studied term-definition facts and then trained on two-thirds of the facts using multiple-choice tests with feedback. Half of the test questions involved recalling terms; the other half involved recalling definitions. The remaining facts were either not trained (Experiment 1) or restudied (Experiments 2-4). A 48 hr delayed multiple-choice (Experiments 1-2) or short answer (Experiments 3a-4) final test assessed recall of all terms or all definitions. Replicating and extending prior research, retrieval practice yielded improved recall and positive transfer relative to no training. Relative to restudy, however, retrieval practice consistently enhanced subsequent term retrieval, enhanced subsequent definition retrieval only after repeated practice, and consistently yielded at best minimal positive transfer in either direction. Theoretical and practical implications are discussed.

Public Significance Statement: This research reveals that taking multiple-choice practice tests on term-definition facts results in better learning than does studying. However, that benefit is limited to the tested part of the fact and does not transfer to the reverse case (i.e., retrieving the definition given the term, after having previously retrieved the term given the definition, or vice versa). Accordingly, learners should be aware that testing can yield relatively specific fact learning benefits.

Keywords: retrieval practice, testing effect, transfer, term-definition, fact learning

Does Retrieval Practice Enhance Learning and Transfer Relative to Restudy for Term-Definition Facts?

"What is vision?" "What is the ability to see called?" Those two questions refer to one another: the first asks for the definition of a term (e.g., *vision*), while the second asks for the reverse. The underlying fact that links these two questions ("*Vision is the ability to see*") is a *term-definition* fact: an "A-is-B" formatted fact that links a concept (i.e., term) with its corresponding definition. Term-definition facts are ubiquitous across numerous subject areas, ranging from commonly learned academic concepts (e.g., the definition of *photosynthesis* in biology) to more specialized domains (e.g., the definition of *aileron* in aeronautical engineering). Mastery of such facts provides a prerequisite foundation for thinking, problem solving, and other higher order skills (Willingham, 2009). To be competent in a given domain, one must master important terminology and concepts, and the learning of term-definition facts can be essential toward that end.

One promising method for learning term-definition facts is *retrieval practice*, a technique which involves attempting recall of to-be-learned information (e.g., by taking a practice test). The benefits of retrieval practice (also known as the *testing effect* and *test-enhanced learning*, among other appellations), relative to control tasks such as no training or restudy, have been demonstrated across a wide range of materials, for learners both young and older (Meyer & Logan, 2013), and across different levels of memory ability (Pan, Pashler, Potter, & Rickard, 2015). Accordingly, many learning researchers now classify retrieval practice as one of the most robust learning techniques available (e.g., Brown, Roediger, & McDaniel, 2014; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013) and recommend its implementation in schools (e.g., Pashler et al., 2007).

The two basic methods of using retrieval practice for term-definition facts—namely to

recall the term (*term-response*) given the definition or recall the definition (*definition-response*) given the term—are the focus of this manuscript. More specifically, we address three questions: (a) does retrieval practice in the form of term-response or definition-response confer memory benefits for the same response on a later test?; (b) are any such benefits observed relative to a restudy control condition?; and (c) do any such benefits generalize, or *transfer*, to the reversed case; that is, does learning due to a term-response test transfer to a definition-response test, and vice versa?

The comparison of retrieval practice against a restudy control condition (as opposed to a no training control, which has been used in several prior studies that are relevant to the present topic, as discussed later in this manuscript) addresses the important pedagogical issue of whether there is a benefit of practice testing over alternative learning strategies. Given a finite amount of available learning time, should learners use practice testing or study when learning term-definition facts? Study (and restudy) is used more often by students than is practice testing (e.g., Kornell & Bjork, 2007). Further, unlike a no training condition, the restudy condition allows control for time on task and for reexposure to materials (Carrier & Pashler, 1992).

To date, two studies have addressed the issue of learning and transfer for term-definition facts following retrieval practice (for an overview of the larger literature on retrieval practice and transfer of learning, see Carpenter, 2012; for a meta-analytic review, see Pan & Rickard, 2017). As described in the following section, both studies demonstrated positive transfer. Neither, however, addressed the case of transfer relative to a restudy control for the simplest case implementation of retrieval practice that is most likely to be used in educational contexts. Further, studies involving such basic implementations of retrieval practice on facts that do not explicitly have a term-definition structure have often shown no positive transfer relative to a restudy control (e.g., Agarwal, 2011; Hinze & Wiley, 2011; Pan, Gopal, & Rickard, 2015).

Retrieval Practice and Transfer Between Fact Terms

In a classroom study, McDaniel, Thomas, Agarwal, McDermott, and Roediger (2013; Experiment 1) had 7th grade students take three identical multiple-choice clicker quizzes with feedback on term-definition facts over an approximately 11-day period. The quizzes covered two-thirds of the facts that students were learning across three academic subjects: cells, machines/energy, and animals. Half of the quiz questions required term retrieval, and half required definition retrieval. The remaining one-third of the materials was not practiced. A multiple-choice final test, administered one day after the third practice quiz, had three conditions: (a) questions that were identical to those that were used during practice quizzes (e.g., term-response on both), (b) *term-definition reversals* (e.g., if retrieval practice on a given fact involved term retrieval, the final test question on that fact involved definition retrieval, or vice versa, hence the label "reversal"), and (c) questions on facts that had not been trained. A substantial testing effect was observed relative to no training, as well as substantial transfer of that effect; that is, there was roughly equivalent performance in the first and second conditions, in the context of much poorer performance in the third, no training condition.

The results of McDaniel, Thomas, et al. (2013) suggest that retrieval practice holds substantial promise at enhancing learning and transfer for term-definition facts. However, the authors acknowledged that their experimental design may limit the generalizability of their findings. It was impossible in that experiment to fully control for time on task, outside study activities, and the effects of student motivation on the final exam (which accounted for half of the course grade). For example, it is likely that students attended particularly well to the feedback for incorrectly answered quiz questions, and used that feedback to guide their subsequent outside study activities. Further and perhaps most importantly, transfer was assessed against a no training control. Although the authors speculated that significant testing and

transfer effects would still have been observed had a restudy control condition (which is expected to produce at least some learning) been used, that possibility was not empirically tested.

In relation to that last issue, McDaniel, Bugg, Liu, and Brick's (2015) study of retrieval practice and term-definition facts incorporated a restudy control. In each of two experiments, subjects studied a research methods text and then trained two days later on that text using test-test-test (TTT), test-study-test (TST), or study-study-study (SSS) schedules. Retrieval practice involved answering a multiple-choice term- or definition-response question comparable to those of McDaniel, Thomas, et al. (2013; Experiment 1), as well as a multiple-choice *application* question for each fact. The application question involved identifying the term which best matched a provided example (e.g., *"You take a test and score very high; a week later you take the same test and score moderately; this test lacks _____?"* for which the answer is *reliability*). Feedback was not provided in Experiment 1 but was provided in Experiment 2. On a short answer final test administered four days after training, positive transfer to term-definition reversals was observed for the TTT and TST groups in Experiment 1 and only for the TST group in Experiment 2, relative to restudy (the SSS condition).

McDaniel, Bugg, et al. (2015) reached a similar conclusion as McDaniel, Thomas, et al. (2013), namely that retrieval practice can result in transferrable learning for term-definition facts. However, that evidence of positive transfer was obtained with a multi-session paradigm in which training entailed repeatedly answering two types of questions per fact (a term- or definition-response question as well as an application question), plus (for the more potent TST condition) re-reading of text in between practice tests. Thus, the question of whether a more basic implementation of retrieval practice (i.e., without re-reading of materials between tests, additional application questions, or necessarily involving repeated testing) enhances transfer for term-definition facts, relative to restudy, remains unaddressed.

That simplified form of retrieval practice—henceforth referred to as *unelaborated* retrieval practice (i.e., practice testing with brief correct answer feedback but no other elaborative processing or added post-retrieval activities)—is characteristic of studies that use facts without a term-definition structure. In those studies, unelaborated retrieval practice typically does not enhance transfer from one part of a fact to another, relative to a restudy control. For example, Pan, Gopal, et al. (2015) had subjects study facts from Advanced Placement (AP) History and Biology courses (e.g., "*Jefferson purchased Louisiana from the Spanish"*) and then take practice tests with feedback on individual words from those facts (e.g., "*WHO purchased Louisiana from the Spanish?"*). Across four experiments, they consistently found robust testing effects on final test questions that involved recall of the same words as during training. However, there was no evidence of positive transfer to final test questions that involved recall of different words from previously tested facts (e.g., "*Jefferson purchased WHAT from the Spanish?"*), relative to restudy. That result occurred whether training was in short answer or multiple-choice format, with single or multiple practice trials per fact, and for history and biology facts. Accordingly, the authors concluded that retrieval practice confers potent, but piecewise, learning benefits for AP and similar types of facts.

The findings of Pan, Gopal, et al. (2015) build on prior work which suggests that unelaborated retrieval practice yields highly specific and nontransferable learning for complex facts (e.g., Hinze & Wiley, 2011) and other types of stimuli with multiple elements (e.g., word triplets as in Pan, Wong, Potter, Mejia, & Rickard, 2015). It thus appears that transfer from one response element to another following unelaborated retrieval practice is weak for stimuli with multiple terms or elements, and does not exceed the learning that occurs through restudy. None of those studies, however, used term-definition facts as materials.

Despite the counter-evidence discussed above, one reason to suspect that unelaborated

retrieval practice for term-definition facts may yield transfer relative to restudy is that such transfer is observed for structurally analogous paired associates. Term-definition facts have an apparent two-element (i.e., *term → definition*) associative structure as is the case for paired associates (for prior analogies of fact learning to paired associates, see Mozer, Howe, & Pashler, 2004; Rohrer, Taylor, & Sholar, 2010; for related discussion see McDaniel, Thomas, et al., 2013). Carpenter, Pashler, and Vul (2006) demonstrated potent transfer of learning from one response element of a paired associate to another following cued recall practice tests (e.g., for the word pair *beach → blanket*, a cued recall test on *beach → ?* resulted in better final test performance on *blanket → ?*, relative to restudy). This finding was subsequently replicated by Cheng (2014) using a multiple-choice training format (in which the cue from a word pair was presented on each practice test trial, with subjects having to select one of four possible answers). These results suggest that retrieval practice may yield transfer when the to-be-learned materials involve forming associations between only two elements, as is the case, on the surface at least, with terms and their definitions.

The Current Study

To investigate whether unelaborated retrieval practice enhances learning and transfer for term-definition facts, we conducted four experiments. The design and procedures were adapted from the paradigm used in Pan, Gopal, et al. (2015). In contrast with prior work on term-definition facts, we did not pair each term- or definition-response practice test question with an additional application question, nor did we provide the opportunity for additional exposure to the facts outside the experiment. Our implementation of retrieval practice can be compared to similar implementations of practice test questions in educational contexts, such as those provided at the end of textbook chapters or given by an instructor in advance of a high-stakes final exam.

Experiment 1 was designed as a conceptual replication of McDaniel, Thomas, et al.'s

(2013) results, in which retrieval practice and transfer effects were observed relative to a no training control. To foreshadow, the results fully replicated their findings. In Experiments 2-4 we investigated, for the first time, the learning of term-definition facts under conditions of unelaborated retrieval practice vs. a restudy control (which, as we argued above, is more informative from an educational perspective).

Experiment 1

The first experiment aimed to determine whether the positive transfer relative to a no training control, as observed by McDaniel, Thomas et al. (2013; Experiment 1), would replicate in a laboratory experiment with university students and for the case in which retrieval practice consisted of a single practice test trial with feedback per fact. Besides changes in setting, subjects, and amount of retrieval practice, there were four additional design differences between the prior and current experiments: use of university-level factual materials, no extra study opportunities outside the experiment, a retention interval of 48 hrs, and a between- rather than within-subjects manipulation of question type (term- or definition-response) on the final test.

Method

Subjects. In this and all subsequent experiments, undergraduate students recruited from the University of California, San Diego subject pool participated in exchange for course credit. The target sample size, which is comparable to that of prior retrieval practice and transfer studies (e.g., Hinze & Wiley, 2011; McDaniel & Fisher, 1991) and also applied to Experiment 2, was 50. Fifty-five undergraduate subjects participated. All but one subject completed both sessions of the experiment; data from the remaining 54 subjects (*term-response* group, n =28; *definition-response* group, n =26) was analyzed. The entire study was conducted with the approval of the university's institutional review board, and all subjects provided written informed consent prior to participating.

Materials. Thirty-six term-definition facts, each averaging 17 words in length, were extracted from a widely-used undergraduate introductory neuroscience textbook, *Biological Psychology: An Introduction to Behavioral, Cognitive, and Clinical Neuroscience* (Breedlove & Watson, 2013), and its publisher-provided test bank. Versions of this textbook have been used in prior retrieval practice studies (e.g., McDaniel, Anderson, et al., 2007). Each fact defined a concept in biology or neuroscience (e.g., *consciousness*). One to three facts were extracted from each of the textbook's nineteen chapters, with each fact having a term-definition "A-is-B" structure in which (a) the concept term was at the start of the sentence (either starting with the first word or after the articles "a", "an", or "the"), (b) that term was followed by the verb "is" or a comparable verb (e.g., "refers"), and (c) that verb itself was followed by the definition of the term. An example fact is: "*Consciousness is personal awareness of one's own emotions, thoughts, movements, and experiences*." Terms from all 36 facts are listed in Appendix A.

For each fact, two multiple-choice questions were created: a term-response question and a definition-response question. For both questions, four answer choices were created: the correct answer and three plausible lures. The correct answers and lures did not overlap with those for any of the other facts. Examples of facts, test questions, answers, and lures are included in Appendix B.

For counterbalancing purposes, three sub-lists of 12 facts each were randomly drawn from the full set of 36 facts, with no fact overlap across sub-lists. Six experimental lists of 24 facts each were then created using all possible pairings of two of the three sub-lists. Within each experimental list, one sub-list of facts were assigned to be trained using term-response test questions, and the other sub-list to be trained using definition-response test questions. An additional 12 facts not in that experimental list were assigned to the no training condition. This design enabled each fact to appear in each of the three training conditions (i.e., trained with a

term-response question, trained with a definition-response question, or not trained at all) with equal frequency over subjects.

Design and procedure. The experiment entailed two sessions separated by a 48 hr delay. As shown in Figure 1, Session 1 contained two phases: the study phase and the training phase. During the study phase, subjects viewed all 36 facts, one at a time, for 10 seconds each, and in a random order determined anew for each subject. All facts were studied once.

The training phase, in which each subject trained on one experimental list, followed immediately afterward. Subjects were assigned to one of the six experimental training lists in counterbalanced fashion. There was one training phase test trial per fact (24 trials total), and test questions were presented in random order determined anew for each subject. For each training phase test question, subjects were given 12 s to select one of the four possible answer choices by typing the letters A, B, C, or D. Answer choice order was randomized on each trial for each subject. After 12 s had elapsed, the correct answer was subsequently shown for 3 s. That correct answer constituted feedback. When the feedback appeared, the test question, the four possible answer choices, and the subject's selected response remained on screen, thus providing subjects with all relevant information with which to study the feedback (similar to McDaniel, Thomas, et al., 2013). After the training phase ended, subjects were reminded of their Session 2 appointment, falsely informed that they would return to study new facts (a deception designed to prevent outside study), and dismissed.

In Session 2, which occurred after a 48 hr delay, subjects completed the final test. The question type on the final test (term- or definition-response) was manipulated between-subjects, with random assignment to either the term-response group or the definition-response group. Based on this assignment, subjects were assessed on recall of either all terms or all definitions for each of the 36 facts, with one test trial per fact. On each trial, a multiple-choice test question

and four possible answer choices were presented, and subjects input their answer by typing the letters A, B, C, or D. As occurred during Session 1, answer choice order was randomized on each test trial for each subject. No feedback was provided on the final test, and subjects had unlimited time to respond on each trial.

Data analysis. Final test data (i.e., proportion correct on final test questions) were analyzed using a 3 (Final Test Condition: tested-same, tested-different, and not trained) x 2 (Question Type: term-response and definition-response) mixed factors design, with Final Test Condition as the within-subjects variable and Question Type as the between-subjects variable. *Tested-same* indicates that training and final test questions for a given fact were identical (e.g., term-response in both cases), while *tested-different* indicates that training and final test questions for a given fact were different (i.e., term-definition reversals, such as term-response for training and definition-response on the final test).

Results

Training. Proportion correct was $M = 0.80$, $SE = 0.021$ for term-response questions and $M = 0.70$, $SE = 0.021$ for definition-response questions. The higher accuracy on term-response questions was statistically significant (here and in all subsequent analyses we use $\alpha = 0.05$), $t(53) = 4.35$, $p < .0001$, $d = 0.59$. That pattern suggests that recall of definitions is more difficult than recall of terms, matching an observation previously made by Lipko-Speed, Dunlosky, and Rawson (2014) and patterns observed in McDaniel, Thomas, et al. (2013).

Final test. The mean proportion correct across the two question types and the three final test conditions is shown in Figure 2, Panel a. For comparison, the results of McDaniel, Thomas, et al. (2013; Experiment 1) are depicted in Panel b. A factorial Analysis of Variance (ANOVA) on subject-level proportion correct scores yielded a statistically significant effect of Final Test Condition, $F(2, 104) = 30.82$, $MSE = 0.34$, $p < .0001$, $\eta_p^2 = 0.37$, no significant effect of

Question Type, $F(1, 52) = 1.12$, $MSE = 0.070$, $p = .29$, $\eta_p^2 = 0.021$, and no significant Final Test Condition by Question Type interaction, $F(2, 104) = 2.07$, $MSE = 0.023$, $p = .13$, $\eta_p^2 = 0.038$.

Comparison of Panels a and b shows that the final test results of Experiment 1 are highly analogous to those of McDaniel, Thomas, et al. (2013; Experiment 1). Most critically, in both experiments there were large performance differences between the tested-same and tested-different conditions on one hand, and the no training condition on the other, a result that was observed for both the term- and definition-response groups. Hence, in both experiments there was substantial evidence that retrieval practice yielded transfer relative to a no training control. The only statistical difference between the experiments was marginally better performance in the tested-same than in the tested-different condition for the term-response group in our experiment, $t(27) = 2.17$, $p = .039$, $d = 0.41$, but not in the McDaniel, Thomas, et al. experiment.

Effect of prior course experience. In exit surveys, well over half ($n = 38$) of the subjects reported having previously taken a biological psychology or neuroscience course. While prior experience translated into better overall performance during training ($M = 0.77$, $SE = 0.070$ vs. $M = 0.71$, $SE = 0.084$) and on the final test ($M = 0.73$, $SE = 0.069$ vs. $M = 0.63$, $SE = 0.067$), there was no effect of that expertise difference in the pattern of testing and transfer effects. Moreover, in Experiments 2-4 (for which the percentage of subjects reporting prior course experience was 47%, 45%, 17%, and 47%, respectively), a similar pattern of better overall performance for experienced subjects was also observed, but again that improved performance did not change the resulting testing effect and transfer patterns. Results of exit surveys are not discussed further.

Discussion

The results of Experiment 1 confirm that, relative to a no training control, retrieval

practice (unelaborated practice in this case) enhances final test performance for both the previously tested response (i.e., a testing effect) and the reverse response (i.e., a transfer effect). That replication and extension of the findings of McDaniel, Thomas, et al. (2013) shows that positive testing and transfer effects for term-definition facts relative to a no training control are robust across subject populations, different implementations of retrieval practice, and the various other differences in experimental design.

Experiment 2

In this experiment we explored the question, heretofore unaddressed, of whether unelaborated retrieval practice enhances learning and transfer for term-definition facts relative to a restudy control. This experiment retained all design and procedural features of Experiment 1 with one exception: the control condition was changed from no training to restudy. From the applied perspective, this comparison of retrieval practice relative to a non-retrieval reference task is arguably more important than comparison to a no training control.

Method

Subjects. Sixty-one undergraduate students participated for course credit. All but two subjects completed both sessions of the experiment; data from the remaining 59 subjects (*term-response* group, $n = 31$; *definition-response* group, $n = 28$) was analyzed.

Materials, design, and procedure. All aspects of this experiment's design and procedure were identical to its predecessor (see Figure 1), with the exception that the 12 facts that were not trained during the training phase of Session 1 were instead restudied. Each restudy training trial involved the presentation of an entire fact for 15 s each. Thus, the total duration of each restudy trial during training was identical to each practice test trial. Restudy trials were randomly mixed with term- and definition-response practice test trials, for a combined total of 36 training trials (i.e., one trial per fact). As in the preceding experiment, each fact appeared in each

training condition (i.e., trained with a term-response question, trained with a definition-response question, or restudied) with equal frequency over subjects, and question type on the final test (term-response or definition-response) was manipulated between-subjects.

Results and Discussion

Training. Accuracy on the initial test was as follows: term-response questions, M = 0.71, SE = 0.037; definition-response questions, M = 0.62, SE = 0.028. Performance on term-response questions during training was again significantly better than definition-response questions, $t(58) = 3.99$, $p < .001$, $d = 0.52$.

Final test. An ANOVA identical to that performed for Experiment 1 yielded a statistically significant effect of Final Test Condition, $F(2, 114) = 5.60$, $MSE = 0.071$, $p = .0048$, $\eta_p^2 = 0.089$, no significant effect of Question Type, $F(1, 57) = 0.004$, $MSE = 0.00033$, $p = .95$, $\eta_p^2 < 0.001$, and a significant Final Test Condition by Question Type interaction, $F(2, 114) = 3.80$, $MSE = 0.049$, $p = .025$, $\eta_p^2 = 0.063$. Inspection of Figure 3, Panels a and b provides insight into that interaction. For the term-response group (Panel a) there was a testing effect (tested-same vs. restudied), $t(30) = 3.78$, $p < .001$, $d = 0.68$, but there was no observable transfer (tested-different vs. restudied), $t(30) = 1.18$, $p = .25$, $d = 0.21$. For the definition-response group (Panel b), performance was statistically indistinguishable across conditions; there was no evidence for either a testing effect or a transfer effect.

A cross-experiment analysis of Experiments 1 and 2 confirmed that the transfer results depend critically on control task. In a 2 x 2 x 2 factorial ANOVA with factors of Experiment (1 vs. 2), Final Test Condition (control and tested-different conditions only, a comparison which focuses specifically on the transfer issue), and Question Type (term- or definition-response), there was a significant main effect of Final Test Condition, $F(1, 109) = 24.57$, $MSE = 0.32$, $p < .00001$, $\eta_p^2 = 0.18$, a significant Experiment by Final Test Condition interaction, $F(1, 109) =$

10.47, $MSE = 0.14$, $p = 0.0016$, $\eta_p^2 = 0.088$, and no other significant main effects or interactions ($ps < .41$). The Experiment by Final Test Condition interaction confirms the large differences in transfer results that are evident in a comparison of Panel a of Figure 2 vs. Panels a and b of Figure 3. Specifically, transfer results are markedly different when measured against no training than against restudy.

The results described above suggest that the bulk of the test-enhanced learning and transfer effect in the prior term-definition studies (McDaniel, Bugg, et al., 2015; McDaniel, Thomas, et al., 2013) reflects either use of a no training control, or, in the former study, use of various forms of elaborated retrieval practice. It appears that retrieval practice itself, at least in the form of a multiple-choice test with only brief correct answer feedback, can produce a testing effect, at least for term retrieval, but does not yield a transfer effect.

Experiments 3a and 3b

In the first two experiments we employed multiple-choice initial and final tests, just as did McDaniel, Thomas, et al. (2013). In Experiments 3a and 3b, we switched to a short answer final test format, similar to McDaniel, Bugg, et al. (2015), while retaining multiple-choice practice test questions. The change in final test format was motivated by two considerations. First, whereas multiple-choice tests are commonly used in educational settings, and hence constitute a viable test format for the training test, short answer questions are far more common outside of classroom settings, and therefore constitute a more appropriate format for the final test. Second, if a short answer final test yields larger testing effects than does a multiple-choice test (e.g., Kang et al., 2007; for related discussion see Carpenter & DeLosh, 2006; Halamish & Bjork, 2011), then such a test may also be more sensitive to possible transfer effects.

Method

Experiments 3a and 3b were completed over non-overlapping date ranges using

independent samples from the same subject pool. Experiment 3a involved only term-response final test questions, whereas Experiment 3b involved only definition-response questions. Otherwise those two experiments were identical.

Subjects. Although the small standard errors in Experiment 2 indicate that its results are reliable, we chose to roughly double the sample size in Experiments 3a and 3b for increased statistical power. One-hundred four undergraduate students, all recruited from the same population as in the prior experiments, participated for course credit. All but three subjects in Experiment 3a completed the entire study. Experiments 3a and 3b each involved the same target sample size, 50, as in the preceding experiments. Data from 49 subjects were collected in Experiment 3a (*term-response* only), and data from 52 subjects were collected in Experiment 3b (*definition-response* only).

Materials, design, and procedure. Both experiments were largely identical to the respective term- or definition-response groups of Experiment 2 (see Figure 1), with the primary exceptions that (a) the final test involved short answer test questions instead of multiple-choice, and (b) ten of the facts were shortened with easier to score one word terms (e.g., *absolute refractory period* was changed to *refractory*) and practice and final test questions altered to match. The short answer questions on the final test were minimally modified versions of their multiple-choice counterparts (see Appendix B for examples). Prior to the first final test trial, subjects were told that their typed responses should be spelled as accurately as possible, but that if they were unsure of the correct spelling, to still make their best possible attempt at an answer. The format of the short answer test questions was as follows: the question appeared on the screen, while an empty text box with a cursor appeared directly underneath. Subjects typed their answer and pressed the Enter key to advance to the next trial. As in the preceding experiments, subjects had unlimited time to enter a response on each final test trial.

Data coding and analysis. The nature of typed short answer responses, in which spelling and grammatical errors were possible and indeed often occurred, necessitated a different scoring procedure than in the preceding experiments. Moreover, as term- and definition-response answers differ in length (single word vs. sentence-length answers), dedicated scoring procedures were developed for each.

Term-response scoring. To avoid penalizing subjects for misspelled responses that could be unambiguously identified as referring to correct answers, all final test responses in Experiment 3a were analyzed in Microsoft Excel using the Fuzzy Lookup (Microsoft Research, Redmond, WA) add-in (cf. Metcalfe, Kornell, & Finn, 2009; Pashler, Cepeda, Wixted, & Rohrer, 2005) before statistical analyses were performed. This add-in compares each response to a master list of correct answers, computes a Jaccard similarity score (Microsoft Corporation, 2011), and if a close match is found, identifies the most closely related and accurately spelled answer choice (in this case, the add-in was configured to do so for similarity scores of 0.6 or higher). A comparison with traditional human scoring methods, which involved scoring one-fifth of all responses using the add-in as well as a human rater, indicated that reliability was high (0.98) between the add-in and human scoring. After close misspellings were analyzed and corrected, we used a letter matching algorithm to score for accuracy (where an exact match was scored as correct and all other cases were scored as incorrect).

Definition-response scoring. To score the sentence-length responses in Experiment 3b, and to provide leeway for unambiguously identifiable misspellings, a point-based coding method was developed. Under this method, three to five idea units per fact were identified (cf. Lipko-Speed et al., 2014; Rawson & Dunlosky, 2011; McDaniel, Bugg, et al., 2015), with each idea unit allotted one point. Scoring involved reading each response and assigning yes/no ratings for each idea unit per fact. Eight research assistants trained on this method were assigned up to five

facts each; each assistant scored all of the responses in Experiment 3b for their assigned facts while remaining blind as to training condition assignment. To verify the consistency of this scoring technique, raters also scored an identical subset of the data (comprised of a sample of 5% of the responses for each of the 36 facts). These scores where then compared with scores generated by the authors; there was high reliability (≥ 0.80) between the raters and the authors. For statistical analyses, scores for each response (i.e., number of yes ratings divided by total number of points possible) were dichotomously transformed (i.e., into correct or incorrect, which corresponds with the scoring outcomes on the term-response final test) using the following criterion: > 60% of idea units had to be correctly recalled in order for a response to be scored as correct. Summary data for analyses using continuous scores (i.e., untransformed data) are also reported. To foreshadow, the use of either dichotomous or continuous scores did not change the overall pattern of results.

Results and Discussion

Training. Proportion correct on the multiple-choice test in Experiment 3a was $M = 0.76$, $SE = 0.023$, for term-response questions and $M = 0.72$, $SE = 0.024$ for definition-response questions. While mean performance was numerically better for term- vs. definition-response questions, matching the pattern observed in the prior experiments, in this case the difference was not statistically significant, $t(48) = 1.66$, $p = .10$, $d = 0.24$. In Experiment 3b, accuracy on the initial test was $M = 0.68$, $SE = 0.027$ for term-response questions and $M = 0.62$, $SE = 0.030$ for definition-response questions. Mean performance on term-response questions was significantly better than on definition-response questions, $t(51) = 2.79$, $p = .0074$, $d = 0.39$, again matching the pattern observed in prior experiments.

Final test. Final test results for Experiment 3a are shown in Figure 4, Panel a. A within-subjects one-way ANOVA on proportion correct scores yielded a statistically significant main

effect of Final Test Condition, $F(2, 96) = 6.08$, $MSE = 0.12$, $p = .003$, $\eta_p^2 = 0.10$. Inspection of Figure 4 reveals that this effect is driven primarily by a testing effect for the tested-same condition in the absence of substantial transfer to the tested-different condition. That pattern was confirmed by two tests. The first, limited to the tested-different and restudied conditions, yielded no significant effect of Final Test Condition, $t(48) = 0.62$, $p = .53$, $d = 0.089$. The second, limited to the tested-same and tested-different conditions, yielded a significant effect of Final Test Condition, $t(48) = 2.73$, $p = .009$, $d = 0.39$. Experiment 3a thus fully replicated Experiment 2: for the case of unelaborated retrieval practice vs. a restudy control, testing enhances term retrieval, but that enhancement does not transfer to the reverse case of definition retrieval. The same ANOVA for Experiment 3b (with dichotomous scores, as shown in Figure 4, Panel b) yielded no significant main effect of Final Test Condition, $F(2, 102) = 2.29$, $MSE = 0.026$, $p = .11$, $\eta_p^2 = 0.043$, while the same ANOVA using continuous scores (mean accuracy (*SE*) of 0.30 (0.020), 0.28 (0.022), 0.28 (0.018) in the tested-same, tested-different, and restudied conditions, respectively) also yielded no significant main effect of Final Test Condition ($p = .70$). Thus, just as for Experiment 2, there was no statistical evidence for a testing or transfer effect for definition-responses. It should be noted, however, that there was a numerical trend in favor of a testing effect in Experiment 3b.

A cross-experiment analysis indicated that performance was better overall for the case of term-response (Experiment 3a) than definition-response (Experiment 3b) short answer final tests. In a 2 x 3 factorial ANOVA with factors of Experiment (3a vs. 3b; using dichotomous scores) and Final Test Condition (tested-same, tested-different, and restudied), there was a significant main effect of Experiment, $F(1, 99) = 25.16$, $MSE = 1.90$, $p < .0001$, $\eta_p^2 = 0.20$, and Final Test Condition, $F(2, 198) = 8.25$, $MSE = 0.13$, $p < .0001$, $\eta_p^2 = 0.87$, and no Experiment by Final Test Condition interaction, $F(2, 198) = 1.24$, $MSE = 0.019$, $p = .29$, $\eta_p^2 = 0.012$. The main effects are

apparent upon examination of Panels a and b of Figure 4: performance was better overall on the term-response final test, and performance in the tested-same condition tended to be better than the tested-different and restudied conditions in both experiments. Moreover, the absence of an interaction indicates that the relative performance among the tested-same, tested-different, and restudied conditions did not significantly differ between the two experiments (although, as noted previously, pairwise analyses did find evidence of a testing effect for the case of term-response in Experiment 3a, but not for definition-response in Experiment 3b).

Overall, Experiments 3a and 3b fully replicate Experiment 2 with respect to both testing and transfer effects. There was statistical evidence of a testing effect for term- but not definition-responses on short answer final tests, and there was again no evidence of positive transfer in either direction.

Experiment 4

For the fourth experiment we investigated whether increasing the number of practice test trials from one to three per fact yields different learning and/or transfer results. The use of repeated test trials was motivated in part by the lack of a statistically significant testing effect for definition retrieval in Experiments 2 and 3b. Many demonstrations of the testing effect use repeated training tests (e.g., Butler, 2010; Pan, Rubin, & Rickard, 2015; Roediger, Agarwal, McDaniel, & McDermott, 2011), and it has been suggested the repeated testing strengthens the benefits of retrieval practice (e.g., McDaniel, Thomas, et al., 2013), including when the training test format is multiple-choice (e.g., McDaniel, Wildman, & Anderson, 2012; McDermott, Agarwal, D'Antonio, Roediger, & McDaniel, 2014). Thus, by increasing the training "dosage" in this experiment, we sought to increase experimental sensitivity to detect a testing effect for definition retrieval, if it exists in the population. If we observe a testing effect for definition retrieval, then we will also be in a position to further address the issue of transfer for that case.

Method

Subjects. The minimum sample size to detect a small transfer effect was determined using a priori power analysis. Based on the standard deviation of the tested-different minus restudied condition proportion correct difference scores on the final test of Experiment 3b, a sample size of at least 54 per group is needed to achieve power of 0.8 or greater to detect a proportion correct difference score of 0.05 or greater (based on a one-tailed, one-sample *t* test, $\alpha = 0.05$). Accordingly, 124 undergraduate students, all recruited from the same population as in the prior experiments, participated for course credit. All but eleven subjects completed both sessions of the experiment; data from the remaining 113 subjects (*term-response* group, $n = 54$; *definition-response* group, $n = 59$) was analyzed. Subjects were randomly assigned to the term- or definition-response group.

Materials, design, procedure, and data coding. Nearly all aspects of this experiment's design and procedure were identical to that of Experiments 3a and 3b, with the primary exception being that training involved three practice test or restudy trials per fact. This was accomplished by presenting each of the 36 facts once per training block across three training blocks. Assignment of fact to training condition was kept consistent across all three blocks. The design of each training block was functionally identical to that used in the preceding experiment, with the sole exception being that paraphrased questions, reworded lures, and slightly reworded correct definition responses were used for test questions on each block (examples are included in Appendix C), along with paraphrased facts. The purpose of presenting subjects with modified questions and answers on each training block was to encourage careful reading of each question or fact on each practice trial. The scoring procedure for the term- and definition-response final tests was identical to that used in the preceding two experiments; as before, analyses of definition-response final test data using dichotomous scores are reported in their entirety and

analyses using continuous scores are summarized (the use of either scoring methods did not change the overall pattern of results).

Results and Discussion

Training. As shown in Figure 5, accuracy for both term-response and definition-response questions improved across the three blocks of the initial test. A 2 x 3 factorial ANOVA with factors of Question Type (term- or definition-response) and Block (1 vs. 2 vs. 3) yielded a significant main effect of Question Type, $F(1, 112) = 167.9$, $MSE = 4.45$, $p < .0001$, $\eta_p^2 = 0.60$, and Block, $F(2, 224) = 132.9$, $MSE = 1.86$, $p < .0001$, $\eta_p^2 = 0.55$, as well as a significant Question Type by Block interaction, $F(2, 224) = 3.47$, $MSE = 0.040$, $p = .033$, $\eta_p^2 = 0.030$. Performance on term-response questions was better than definition-response questions, in line with the pattern observed in the preceding experiments. The pattern of improvement across blocks indicates that repeated testing generated additional learning, and the magnitude of block-to-block improvements (from the first to the third block, proportion correct improvement of $M = 0.20$ for term-response and $M = 0.15$ for definition-response) was slightly greater for term-response questions.

Final test. An ANOVA identical to that performed for Experiments 1 and 2 (and using dichotomous scores for definition-responses; results shown in Figure 6) yielded a statistically significant effect of Final Test Condition, $F(2, 222) = 37.19$, $MSE = 0.56$, $p < .0001$, $\eta_p^2 = 0.25$, no significant effect of Question Type, $F(1, 111) = 0.73$, $MSE = 0.082$, $p = .39$, $\eta_p^2 < 0.01$, and a significant Final Test Condition by Question Type interaction, $F(2, 222) = 5.25$, $MSE = 0.079$, $p = .006$, $\eta_p^2 = 0.045$. The same ANOVA using continuous scores for definition-responses (mean accuracy (*SE*) of 0.49 (0.017), 0.43 (0.022), 0.42 (0.022) in the tested-same, tested-different, and restudied conditions, respectively) also yielded the same pattern: a significant effect of Final Test Condition and a significant Final Test Condition by Question Type interaction (*p*s < .0001). As

inspection of Figure 6 shows, there were highly significant performance differences between the tested-same and tested-different conditions (excluding the restudied condition) in both the term-response group, $t(53) = 7.11$, $p < .00001$, $d = 0.97$, and the definition-response group, $t(58) = 2.81$, $p = .007$, $d = 0.37$. However, there was again minimal positive transfer from term retrieval to definition retrieval, or vice versa; as evident in Figure 6, mean proportion correct in the restudy conditions was within about one standard error of that in the tested-different conditions. Thus, there were testing effects in this experiment for both term and definition retrieval, but at best minimal transfer in either case.

The source of the Final Test Condition by Question Type interaction is evident upon comparison of Panels a and b of Figure 6: the testing effect for the term-response group is larger than that for the definition-response group. This was confirmed statistically by comparing the difference scores for the tested-same minus tested-different conditions for both groups; the mean difference score in the term-response group was significantly larger, $t(110) = 2.63$, $p < .01$, $d = 0.49$. That finding of a larger testing effect for term- vs. definition-response is broadly consistent with the fact that, at the lower dosage levels of Experiments 2, 3a, and 3b, testing effects were observed on the term- but not definition-response final tests.

General Discussion

In the present experiments we investigated the utility of unelaborated retrieval practice for learning term-definition facts, including whether it generates testing effects for term- and definition-responses, and whether it yields transfer from terms to corresponding definitions and vice versa. Experiment 1 conceptually replicated McDaniel, Thomas, et al. (2013; Experiment 1), in which positive testing effects and strong transfer was observed with middle school students relative to a no training control condition, for both term- and definition-response final tests. Experiment 1 also extended that result from children to adult students, and over changes in

setting, materials, extent of training, and other differences in experiment procedure. Thus, testing and transfer effects for term-definition facts relative to a no training condition appear to be robust. However, when the control condition was switched to restudy in Experiments 2-4, there was at best weak evidence for transfer from either term to definition retrieval or the reverse. Further, percent transfer (bounded by restudy and tested-same performance) decreased rather than increased in Experiment 4, suggesting that transfer wanes at dosages at or beyond that which yields robust testing effects for both term and definition retrieval.

Term-Definition Facts and Stimulus-Response Rearrangement

Our findings indicate that term-definition facts should be added to the list of materials for which transfer to stimulus-response rearranged items (i.e., where the required response was previously a cue, and vice versa) on the final test is minimal or absent relative to a non-testing reexposure control (e.g., restudy), and particularly when the practice tests do not involve extensive feedback in the form of re-reading (e.g., a text passage) or application questions. Those materials include multi-sentence college biology facts (Hinze & Wiley, 2011), AP History and Biology facts (Pan, Gopal, et al., 2015), and word triplets (Pan, Wong, et al., 2015). Analogous failure of transfer has also been observed for both adult's (Rickard & Bourne, 1996) and children's single-digit arithmetic (Walker, Bajic, Mickes, Kwak, & Rickard, 2014). Overall, it appears that specific and nontransferable learning for rearranged stimulus-response components on a final test is a likely outcome of unelaborated retrieval practice for a wide range of facts and other materials with multiple testable components. As detailed in the following section, the only known exception to this pattern involves an even simpler type of stimuli, namely paired associates.

Term-Definition Facts vs. Paired Associates

Some researchers have noted the existence of structural similarities between paired

associates and facts (e.g., McDaniel, Thomas, et al., 2013; Mozer et al., 2004; Rohrer et al., 2010), and have speculated that such similarities may drive comparable learning and transfer processes. That hypothesis is motivated by the fact that both paired associates and facts contain two components to be associated (for paired associates: *cue* → *target*; for term-definition facts: *term* → *definition*). Of most interest here, practice test questions involving term-definition facts appear to reflect that structure. That is, the learner is given the term and asked for the full corresponding definition, or is given the full definition and asked for the corresponding term.

In contradiction to that hypothesis is the current finding that, whereas strong positive transfer (in some cases up to 100%) is consistently observed following various forms of retrieval practice on paired associates (e.g., in multiple-choice or cued recall format, and with single or repetition practice; Carpenter et al., 2006; Cheng, 2014; Vaughn & Rawson, 2014), little or no transfer is observed for term-definition pairs. It thus appears that the greater complexity of term-definition factual materials and other authentic educational facts leads to associative learning processes that are different in important respects than those involved in paired associate learning (a possibility also considered by Rohrer et al., 2010, and McDaniel, Thomas, et al., 2013, and now empirically supported by the current work).

The different learning and transfer properties for term-definition facts vs. paired associates may be related to the fact that the definition component of a term-definition fact contains multiple words that are unlikely to have been strongly associated with each other prior to training, and are exceedingly unlikely to have constituted a single, or chunked, memory representation as in the case of a familiar word (other stimuli with three or more elements may also exhibit similar associative properties; e.g., Pan, Wong, et al., 2016). Thus, a simple bi-directional association between a familiar stimulus and response (as presumably forms for word pairs) is not sufficient for learning. The possibility that learning term-definition facts involves

forming multiple memory associations between the constituent words (or concepts) of each fact may also account for performance disparities between term and definition retrieval: on term-response questions, multiple words comprising the definition are presented as cues for a single to-be-retrieved target, whereas on definition-response questions, only a single cue (i.e., the concept term) is presented for multiple to-be-retrieved targets. Retrieval is likely to be more difficult in the latter case (given that there are no additional cues to help facilitate recall, and also because multiple targets need to be retrieved, rather than a single word as in the case of term retrieval). Thus, repeated retrieval practice (as in Experiment 4) may be necessary to strengthen associations between the single presented cue and its multiple targets in order to generate a testing effect for definition responses relative to restudy (alternatively, as pointed out by a reviewer, a more overtly effortful retrieval format such as cued recall may be used). Follow-up theoretical work that investigates the differences between paired associates and term-definition facts (and other materials that may have a similar paired structure, such as number digit or sound pairs), as well as the differences between concept term and definition retrieval, is warranted.

Practical Implications for Learning Term-Definition Facts

Retrieval practice has potent memory benefits relative to restudy across a wide range of materials, including those explored here. That fact, along with recent demonstrations of positive transfer of highly elaborated retrieval practice for term-definition facts (McDaniel, Bugg, et al., 2015; McDaniel, Thomas, et al., 2013), might lead educators to assume that retrieval practice will generally also give rise to transfer. More specifically, an instructor who wishes to incorporate evidence-based learning techniques might assign term- or definition-response practice test questions, expecting that transfer effects will result relative to alternative exercises. Similarly, a student could devote time to answering term- or definition-response practice test questions at the end of a textbook chapter, expecting that the resultant learning will generalize to

entire facts. Our findings indicate that while unelaborated retrieval practice will generate learning dividends relative to a restudy strategy for term retrieval practice and (with repeated training) definition retrieval practice, it is unlikely to yield enhanced recall for unretrieved portions of such facts. This is an important consequence of retrieval practice on term-definition facts—a consequence of which both instructors and students should be aware.

Will the same testing and transfer effects occur if training involves a potentially more effortful retrieval method such as cued recall (i.e., short answer) rather than multiple-choice, as suggested by Kang et al. (2007)? Recent work provides insights. In Pan, Gopal, et al. (2015), both single and repeated trials of short answer retrieval practice consistently generated substantial testing but no transfer effects for the recall of terms from biology and history facts. Moreover, similar testing effects were observed in that study when the practice test format was switched to multiple-choice. Given the analogous results when the Pan, Gopal, et al. paradigm was adapted for the present experiments, it is unlikely in our view that the use of short answer practice tests will appreciably change the effects of unelaborated retrieval practice (with respect to either the testing effect or transfer) for term-definition facts.

The practical implications enumerated here apply specifically to the case of unelaborated retrieval practice involving term or definition retrieval. That method constitutes the most basic and least time intensive implementation of retrieval practice, and it is the most likely to be used in current educational practice (particularly since multiple-choice tests are relatively easy to score). However, as detailed in the Introduction, more elaborate forms of retrieval practice such as practicing on term- or definition-response and application questions for each term-definition fact (as occurred in McDaniel, Bugg, et al., 2015) appear to hold promise for facilitating positive transfer. This may be due to additional or different cognitive processes engendered by such methods (for related discussions see Hinze, Wiley, & Pellegrino, 2013; Jensen, McDaniel,

Woodard, & Kummer, 2014; McDaniel et al., 2012). Other elaborate forms of retrieval practice that incorporate techniques such as spacing or criterion level learning (e.g., Rawson, Dunlosky, & Sciartelli, 2013), as well as repeated quizzing in conjunction with extended feedback opportunities and outside study (e.g., McDaniel, Anderson, et al., 2007; McDaniel, Wildman, et al., 2012) may also yield greater amounts of transfer for similar types of (i.e., factual) materials. However, such elaborated retrieval practice is more time intensive, and it is unclear whether the transfer results engendered exceed the gains that would be observed if that extra time were instead allocated to retrieval practice on both term- and definition-responses. Further research along those lines is needed to optimize uses of both unelaborated and elaborated retrieval practice for learning term-definition facts.

Finally, it should be noted that the current study investigated one educationally valid transfer context (i.e., transfer between fact elements) out of many possible contexts. Instructors and students may also be interested in transfer under different circumstances that do not necessarily involve term or definition retrieval (for examples see Carpenter, 2012). For instance, McDermott et al. (2014; Experiment 3) investigated retrieval practice and transfer of factual knowledge to application questions (i.e., generalizing prior knowledge to new examples); in that case, positive transfer was observed. Further, in a recent meta-analytic review of the retrieval practice and transfer literature, Pan and Rickard (2017) observed that the extent of transfer in that literature varies substantially between different transfer contexts; transfer involving rearranged stimulus-response elements is generally negligible (especially when elaborate forms of retrieval practice are not used), but more substantial transfer is evident for multiple other transfer types (e.g., application and inference questions). Thus, while the extent of transfer in the present study also proved to be minimal, that result does not preclude different findings involving similar materials but under different transfer contexts.

Conclusions

In the present work we demonstrated that unelaborated retrieval practice on term-definition facts, involving recall of previously studied responses followed by brief correct answer feedback, is superior to restudy. However, that retrieval practice benefit does not transfer from either term retrieval to definition retrieval or the reverse. That absence of transfer extends prior results for history and biology facts without a term-definition structure. In conjunction with other recent work, it now appears that, across a wide range of educational materials, retrieval practice *itself* (i.e., excluding any influence of elaborative post-retrieval processing) cannot be expected to yield transfer to stimulus-response rearranged final test questions. From the practical standpoint, students and instructors should not expect that taking a practice test on part of a term-definition fact, or any other type of fact with multiple testable terms, will be sufficient to enhance memory (relative to a study strategy) on a later test wherein a different response from the fact is required. Nevertheless, retrieval practice appears to be the best overall learning strategy because it can clearly facilitate later memory for the practiced response.

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. (UMI No. 3468823)

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin,* 128(4), 612-637. doi:http://dx.doi.org/10.1037/0033-2909.128.4.612

Breedlove, S. M., & Watson, N. V. (2013). *Biological Psychology: An Introduction to Behavioral, Cognitive, and Clinical Neuroscience* (7th ed.). Sunderland, MA: Sinauer Associates.

Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2014). *Make it stick: The science of successful learning* (pp. 23-45). Cambridge, MA: Belknap Press.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 36(5), 1118-1133. doi:http://dx.doi.org/10.1037/a0019902

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*(5), 279-283. doi:http://dx.doi.org/10.1177/0963721412452728

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(2), 268-276. doi:http://dx.doi.org/10.3758/BF03193405

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13*(5), 826-830. doi:http://dx.doi.org/10.3758/BF03194004

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition,*

20(6), 633-642.

Cheng, C. K. (2014). Effect of multiple-choice testing on memory retention—cue-target symmetry. (Doctoral dissertation). Retrieved from https://tspace.library.utoronto.ca/handle/1807/65649

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*(1), 4-58. doi:http://dx.doi.org/10.1177/1529100612453266

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(4), 801-812. doi:http://dx.doi.org/10.1037/a0023219

Haskell, R. E. (2001). *Transfer of learning: Cognition, instruction, and reasoning* (pp. 24-37). San Diego, CA: Academic Press.

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory, 19*(3), 290-304. doi:http://dx.doi.org/10.1080/09658211.2011.560121

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language,* 69(2), 151-164. doi:http://dx.doi.org/10.1016/j.jml.2013.03.002

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(2), 307-329. doi:http://dx.doi.org/10.1007/s10648-013-9248-9

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of*

Educational Psychology, 101(3), 621-629. doi:http://dx.doi.org/10.1037/a0015183

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*(4-5), 528-558. doi:http://dx.doi.org/10.1080/09541440601056620

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*(2), 219-224.

LaPorte, R. E., & Voss, J. F. (1975). Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology*, 67(2), 259-266. doi:http://dx.doi.org/10.1037/h0076933

Lipko-Speed, A., Dunlosky, J., & Rawson, K. A. (2014). Does testing with feedback help grade-school children learn key concepts in science? *Journal of Applied Research in Memory and Cognition, 3*(3), 171-176. doi:http://dx.doi.org/10.1016/j.jarmac.2014.04.002

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*(4), 476-490. doi:http://dx.doi.org/10.3758/BF03210951

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*(4-5), 494-513. doi:http://dx.doi.org/10.1080/09541440701326154

McDaniel, M. A., Bugg, J. M., Liu, Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention? *Journal of Experimental Psychology: Applied, 21*(4), 370-382. doi:http://dx.doi.org/10.1037/xap0000063

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, 16(2), 192-201.

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*(2), 200-206. doi: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*(3), 360-372. doi: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 and Cognition,* 1(1), 18-26. doi:http://dx.doi.org/10.1016/j.jarmac.2011.10.001

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & 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(1), 3-21. doi:http://dx.doi.org/10.1037/xap0000004

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition, 37*(8), 1077-1087. doi:http://dx.doi.org/10.3758/MC.37.8.1077

Meyer, A. N. D., & Logan, J. M. (2013). Taking the testing effect beyond the college freshman: Benefits for lifelong learning. *Psychology and Aging*, 28(1), 142-147. doi:http://dx.doi.org/10.1037/a0030890

Microsoft Corporation (2011). *Microsoft Fuzzy Lookup Add-In for Excel.* Retrieved from: https://atidan.files.wordpress.com/2013/08/fuzzy-lookup-add-in-for-excel.pdf

Mozer, M. C., Howe, M., & Pashler, H. (2004). Using testing to enhance learning: A comparison of two hypotheses. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the*

Twenty-Sixth Annual Conference of the Cognitive Science Society (pp. 975–980). Mahwah, NJ: Erlbaum.

Pan, S. C., Gopal, A., and Rickard, T. C. (2015). Testing with feedback yields potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology 107*(4). doi:http://dx.doi.org/ 10.1037/edu0000074

Pan, S. C., Pashler, H., Potter, Z. E., and Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language 83*, 53-61. doi:10.1016/j.jml.2015.04.001

Pan, S. C., Rubin, B. R., and Rickard, T. C. (2015). Does testing with feedback improve adult spelling skills relative to copying and reading? *Journal of Experimental Psychology: Applied 21*(4). doi:http://dx.doi.org/ 10.1037/xap0000062

Pan, S. C., and Rickard, T. C. (2017). Transfer of test-enhanced learning: meta-analytic review and synthesis. Manuscript under review.

Pan, S. C., Rubin, B. R., and Rickard, T. C. (2015). Does testing with feedback improve adult spelling skills relative to copying and reading? *Journal of Experimental Psychology: Applied 21*(4). doi: 10.1037/xap0000062

Pan, S. C., Wong, C., Potter, Z., Mejia, J., & Rickard, T. C. (2016). Does test-enhanced learning transfer for triple associates? *Memory & Cognition*, doi: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 learning* (NCER 2007–2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Available from: http://ncer.ed.gov.

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*(1), 3-8. doi:http://dx.doi.org/10.1037/0278-7393.31.1.3

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General, 140*(3), 283-302.

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review, 25*(4), 523-548. doi:http://dx.doi.org/10.1007/s10648-013-9240-4

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*(5), 1281-1295. doi: http://dx.doi.org/10.1037/0278-7393.22.5.1281

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 233-239. doi: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*(6), 1432-1463. doi:http://dx.doi.org/10.1037/a0037559

Vaughn, K. E., & Rawson, K. A. (2014). Effects of criterion level on associative memory: Evidence for associative asymmetry. *Journal of Memory and Language, 75*, 14-26. doi:http://dx.doi.org/10.1016/j.jml.2014.04.004

Walker, D., Bajic, D., Mickes, L., Kwak, J., & Rickard, T. C. (2014). Specificity of children's arithmetic learning. *Journal of Experimental Child Psychology, 122*, 62-74. doi:http://dx.doi.org/10.1016/j.jecp.2013.11.018

Willingham, D. T. (2009). Why don’t students like school? Because the mind is not designed for

thinking. *American Educator*. Available at: http://www.aft.org/sites/default/files/periodicals/WILLINGHAM%282%29.pdf

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 and Cognition*, *3*(3), 214-221. doi:http://dx.doi.org/10.1016/j.jarmac.2014.07.001

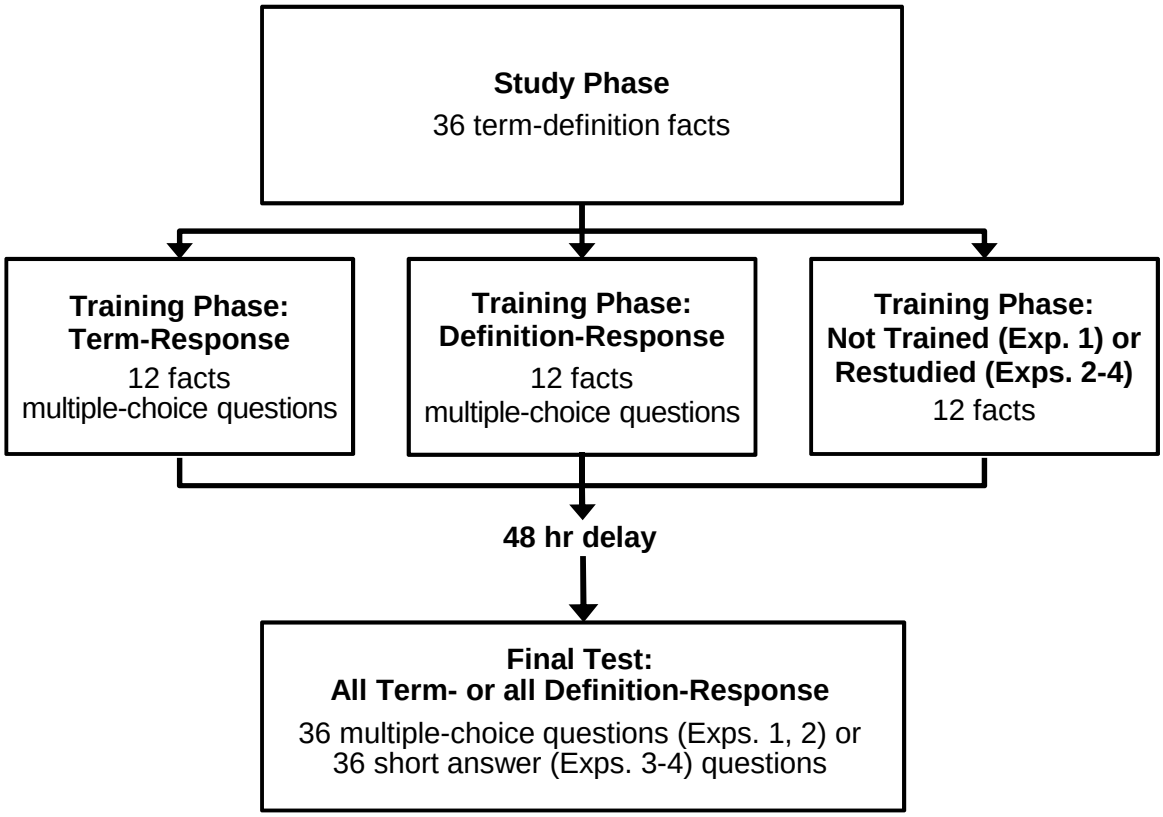

Figure 1. Procedure for Experiments 1-4. Session 1 began with the Study Phase, wherein all 36 facts were studied, followed by the Training Phase, wherein 12 facts each were trained using one (Experiments 1-3b) or three (Experiment 4) *term-response* or a *definition-response* test question(s) with feedback; the remaining 12 facts were either not trained (Experiment 1) or restudied (Experiments 2-4). After 48 hrs, subjects returned for Session 2, the Final Test, and were tested on all term-response or definition-response questions in multiple-choice (Experiments 1, 2) or short answer format (Experiments 3a-4).

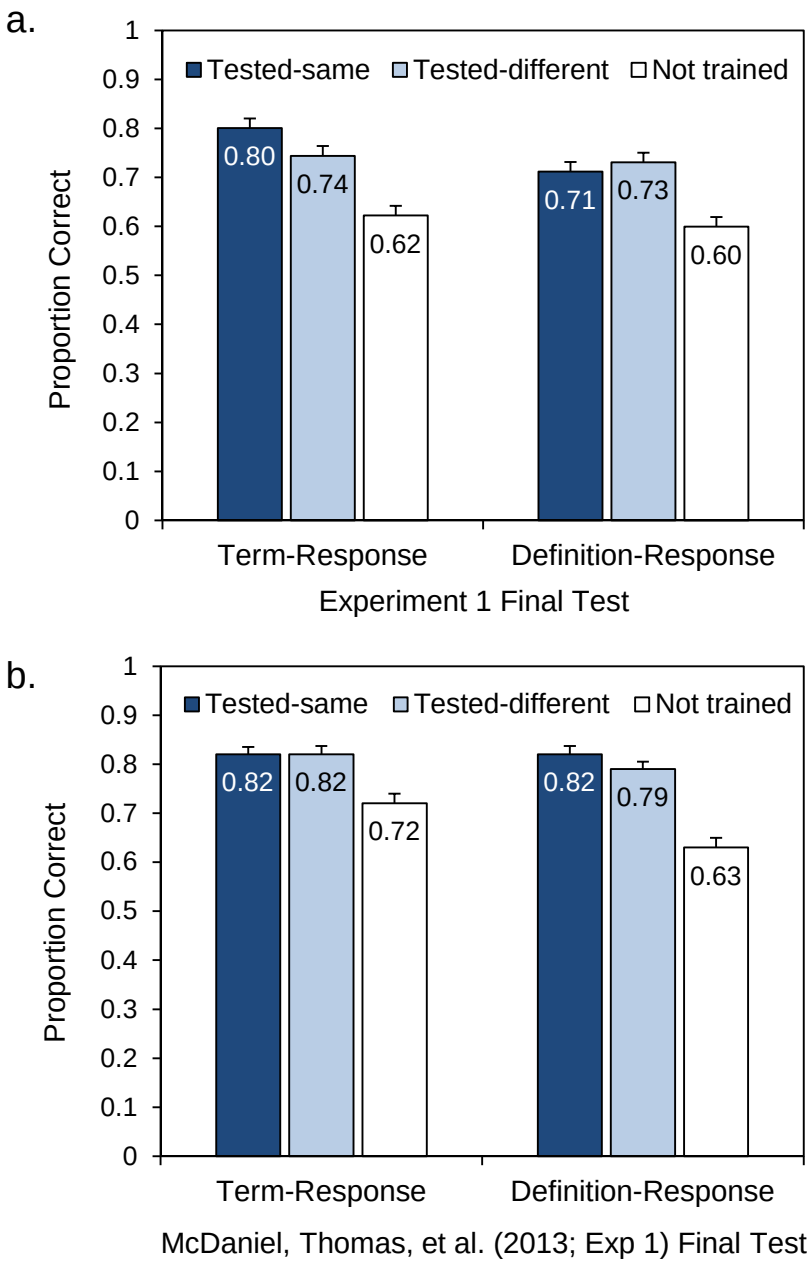

Figure 2. Final test performance on multiple-choice term-response and definition-response questions. Panel a: results from Experiment 1 of the current study. Panel b: results from McDaniel, Thomas, et al. (2013; Experiment 1). *Tested-same* indicates that the training question type matched the final test type; *tested-different* indicates that the training and final test question types did not match. Error bars in Panel a are standard errors based on the interaction error term of a mixed-factors ANOVA on subject mean accuracy scores (based on Loftus & Masson, 1994).

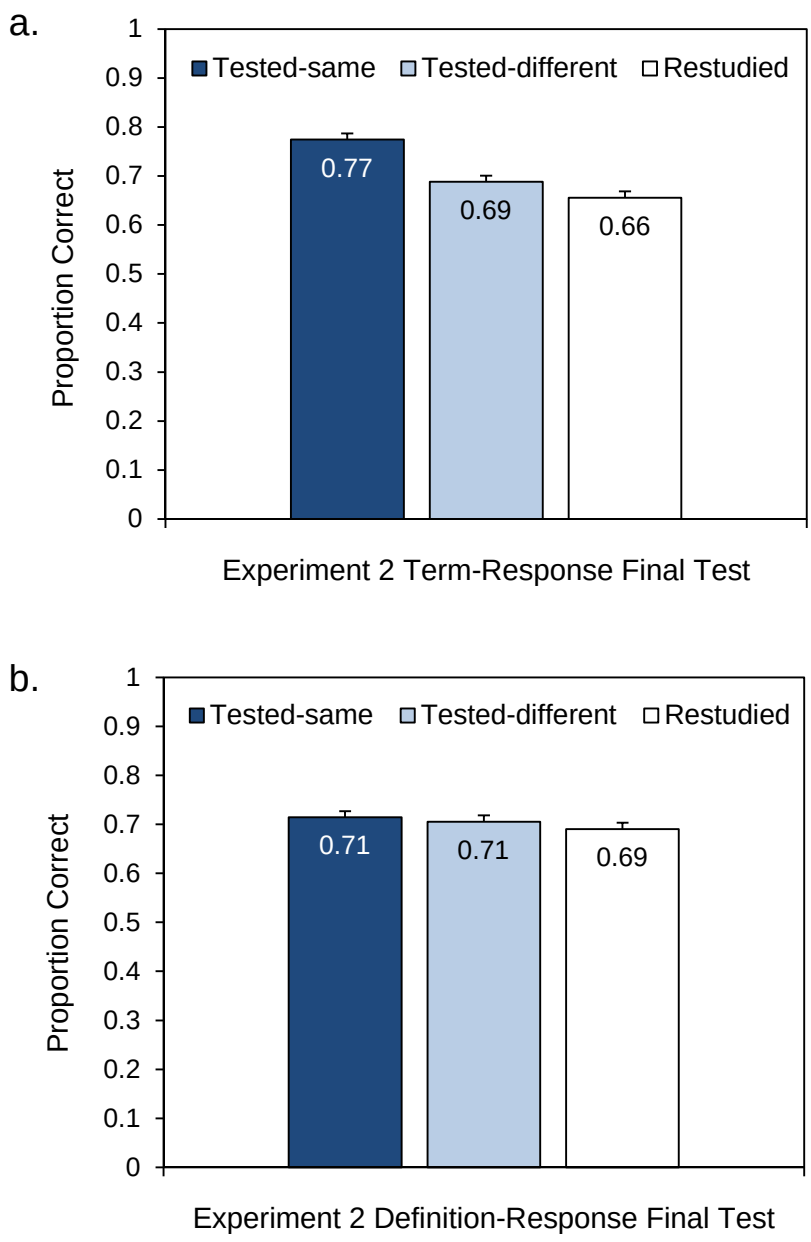

Figure 3. Final test performance on multiple-choice term-response and definition-response questions in Experiment 2. Panel a: term-response. Panel b: definition-response. *Tested-same* indicates that the training question type matched the final test type; *tested-different* indicates that the training question type did not match the final test type. Error bars are standard errors based on the interaction error term of a mixed-factors ANOVA on subject mean accuracy scores (based on Loftus & Masson, 1994).

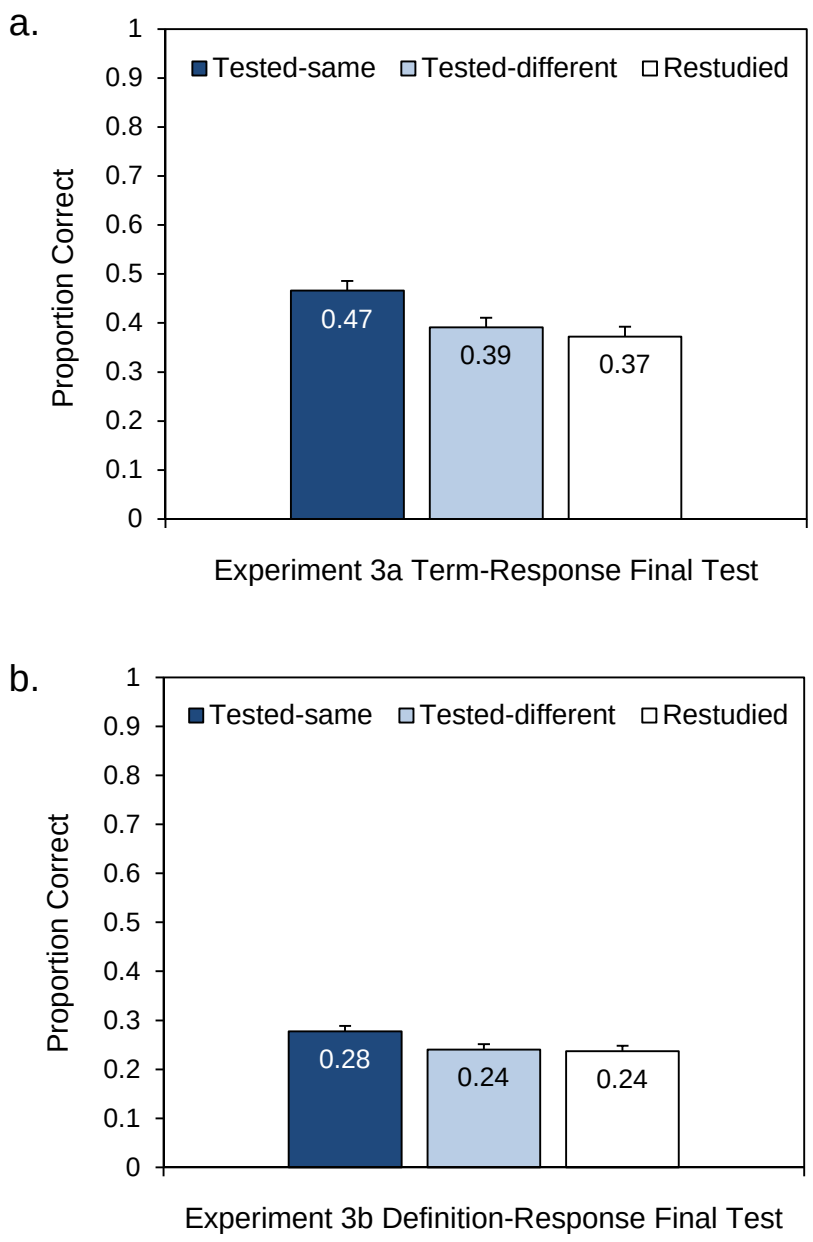

Figure 4. Final test performance on short answer term-response and definition-response questions in Experiments 3a and 3b. Panel a: term-response. Panel b: definition-response. *Tested-same* indicates that the training question type matched the final test type; *tested-different* indicates that the training question type did not match the final test type. Error bars are standard errors based on the error term of a within-subjects ANOVA on subject mean accuracy scores, performed separately for the two experiments (based on Loftus & Masson, 1994).

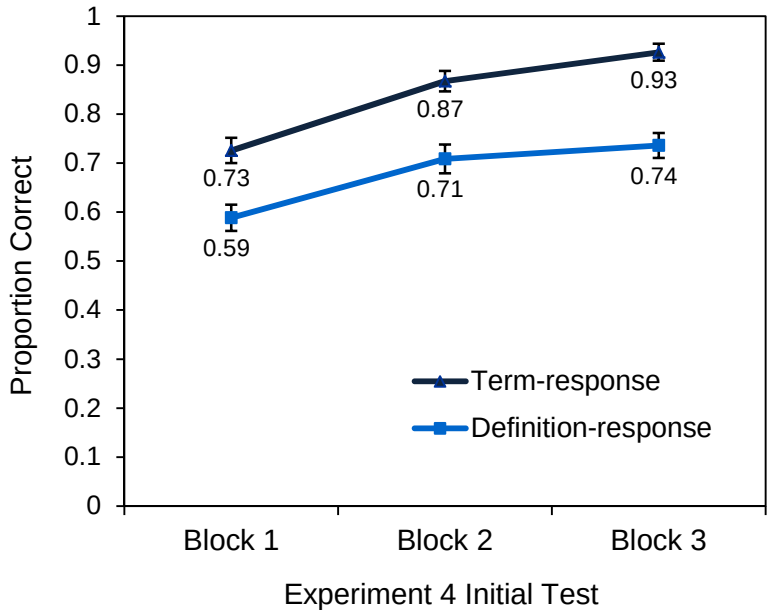

Figure 5. Initial test performance for multiple-choice term-response and definition-response questions across the three training blocks of Experiment 4. Error bars are standard errors of the means.

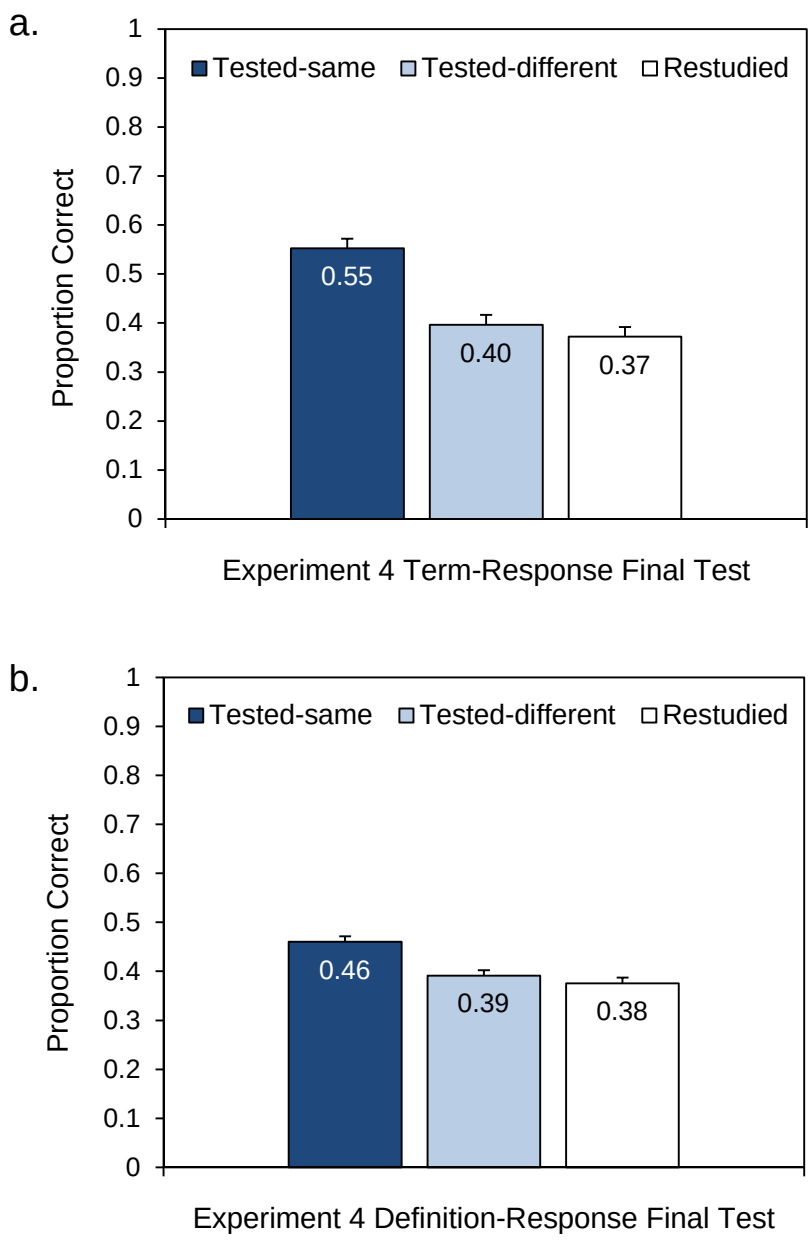

Figure 6. Final test performance on short answer term-response and definition-response questions in Experiment 4, in which each fact was trained three times. Panel a: term-response. Panel b: definition-response. *Tested-same* indicates that the training question type matched the final test type; *tested-different* indicates that the training question type did not match the final test type. Error bars are standard errors based on the interaction error term of a mixed-factors ANOVA on subject mean accuracy scores (based on Loftus & Masson, 1994).

Appendix A

Term-Definition Fact Terms

Fact number	Term
1	Absolute refractory period
2	Acetylcholine
3	Adaptation
4	Cingulate cortex
5	Cocktail party effect
6	Consciousness
7	Dermatome
8	Dermis
9	Dualism
10	ED-50
11	Encoding
12	Focal seizure
13	Glucose
14	Immunocytochemistry
15	Inattentional blindness
16	Infradian cycle
17	Labeled lines
18	Meninges
19	Neuroleptics
20	Neuron doctrine
21	Nissl stain
22	Oculomotor apraxia
23	Oligodendrocytes
24	Osmosensory neurons
25	Oxytocin
26	Paracrine
27	Plasticity
28	Reductionism
29	Sexual selection
30	Somatic intervention
31	Substantia nigra
32	Supplemental motor area
33	Sylvian
34	Temporoparietal junction
35	Tolerance
36	Wernicke's aphasia

Appendix B

Term-Definition Fact and Test Question Examples (used in Experiments 1-4)

Term	Stimulus type	Example
Dermis	Fact	Dermis is the layer of skin that contains a rich web of nerve fibers in a network of connective tissue and blood vessels.
	Term-Response (MC training or final tests, Exps 1-4)	What is the layer of skin that contains a rich web of nerve fibers in a network of connective tissue and blood vessels? a. epidermis b. dermis c. hypodermis d. epithelium
	Definition-Response (MC training or final tests, Exps 1-4)	What is the definition of dermis? a. The layer of skin that contains a rich web of nerve fibers in a network of connective tissue and blood. b. The layer of skin that contains no blood vessels and must receive nourishment from underlying tissue. c. The deepest layer of skin on the body. d. The protective layer of chitin that covers vulnerable body parts.
	Term-Response (SA final test, Exps 3a, 4)	The layer of skin that contains a rich web of nerve fibers in a network of connective tissue and blood vessels is called what?
	Definition-Response (SA final test, Exps 3b, 4)	What is the definition of dermis?
ED-50	Fact	ED-50 is the dose at which a drug has a half-maximal response.
	Term-Response (MC training or final tests, Exps 1-4)	What is the dose at which a drug has a half-maximal response? a. saturated b. therapeutic c. ED-50 d. LD-50
	Definition-Response (MC training or final tests, Exps 1-4)	What is the definition of ED-50? a. The dose at which a drug has a half-maximal response. b. The dose at which a drug is halfway metabolized by the body. c. The dose at which a drug results in a 50 percent mortality rate. d. The maximum dose at which a drug can be safely administered.
	Term-Response (SA final test, Exp 3a, 4)	The dose at which a drug has a half-maximal response is called what?
	Definition-Response (SA final test, Exp 3b, 4)	What is the definition of ED-50?

(appendix continues)

Appendix B *(continued)*

Term	Stimulus type	Example
Focal seizure	Fact	A focal seizure is a seizure that initially affects only one hemisphere of the brain.
	Term-Response (MC training or final tests, Exps 1-4)	What is a seizure that initially affects only one hemisphere of the brain called? a. petit b. grand c. complex d. focal
	Definition-Response (MC training or final tests, Exps 1-4)	What is the definition of a focal seizure? a. A seizure that initially affects only one hemisphere of the brain. b. A seizure which occurs suddenly without warning and results in total unresponsiveness. c. A seizure that does not involve the entire brain, and is characterized by sudden violent involuntary movements. d. A seizure that occurs with slow, gradual onset and results in mild impairment.
	Term-Response (SA final test, Exp 3a, 4)	A seizure that initially affects only one hemisphere of the brain is called what?
	Definition-Response (SA final test, Exps 3b, 4)	What is the definition of a focal seizure?

Note: Exp = Experiment, MC = multiple-choice, SA = short answer.

Appendix C

Paraphrased Term-Definition Practice Question Examples (used in Experiment 4)

Term	Stimulus type	Example
Neuroleptic	Term-Response (version 1)	What type of drug is effective at reducing the symptoms of schizophrenia? c. anxiolytic d. tricyclic c. neuroleptic d. sedative
	Term-Response (version 2)	A drug that is capable of reducing a patient's schizophrenic symptoms is called...? a. analgesic b. neuroleptic c. amphipathic d. neurolytic
	Term-Response (version 3)	A drug that can reduce schizophrenic symptoms effectively is called..? a. neuroleptic b. inhalant c. prophylactic d. inhibitor
Consciousness	Definition-Response (version 1)	What is consciousness? a. personal awareness of one's own emotions, thoughts, movements, and experiences. b. the mental state comprised of the id, ego, and superego. c. the ability to recognize oneself as a separate and distinct entity from other individuals in the environment. d. the ability to observe and empathize with one's actions and those of others.
	Definition-Response (version 2)	What does consciousness mean? a. components of the mind such as the id and the ego. b. the awareness of one's distinct existence apart from other people in an environment. c. self-awareness of personal emotional, physical, and mental experiences. d. the capacity for empathy when observing the actions of oneself and others.
	Definition-Response (version 3)	What does consciousness refer to? a. the perception of self as a an entity which is separate from other people. b. id, ego, and superego which comprise the mental state. c. mindfulness of internal and external phenomena. d. mindfulness of one's own emotions, thoughts, movements, and experiences.

Note: Exp = Experiment.