Mem Cogn
DOI 10.3758/s13421-015-0547-x

Does test-enhanced learning transfer for triple associates?

Steven C. Pan[1] · Carol M. Wong[1] · Zachary E. Potter[1] · Jonathan Mejia[1] · Timothy C. Rickard[1]

Abstract Test-enhanced learning and transfer for triple-associate word stimuli was assessed in three experiments. In each experiment, training and final-test trials involved the presentation of two words per triple associate (triplet), with the third word having to be retrieved. In agreement with the prior literature on different stimuli, training through testing with feedback yielded markedly better final-test performance than did restudy. However, in contrast to the positive transfer reported for paired associate stimuli, minimal or no positive transfer was observed, relative to a restudy control, from a trained cue combination (e.g., *A*, *B*, *?*) to other cue combinations from the same triplet that required a different response (e.g., *B*, *C*, *?*). That result also held when two unique cue combinations per triplet were tested during training, and for triplets with low and high average associative strengths. Supplementary analyses provided insight into the overall transfer effect: An incorrect response during training appears to yield positive transfer relative to restudy, whereas a correct response appears to yield no, or even negative, transfer. Cross-experiment analyses indicated that test-enhanced learning is not diminished when two or three cue combinations are presented during training. Thus, even though learning through testing is highly specific, testing on all possible stimulus–response combinations remains the most efficient strategy for the learning of triple associates.

Keywords Cued recall · Retrieval practice · Test-enhanced learning · Testing effect · Transfer

The act of recalling information during a test, otherwise known as *retrieval practice*, enhances memory for that information, above and beyond an equivalent period of time spent restudying the same materials. This *testing effect*, or *retrieval practice effect*, has been replicated numerous times with stimuli ranging from paired associates to prose passages (Roediger & Karpicke, 2006a) and is widely regarded as one of the most robust phenomena in learning science (Butler, 2010; Carpenter, 2012). However, a limitation of the testing-effect literature to date is that the vast majority of published studies have used identical materials during the initial and final tests (for discussion, see McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013). Comparatively less is known about testing's effectiveness for transfer to novel contexts.

Alongside retention, transfer has been described as the ultimate objective of learning (e.g., Carpenter, 2012; Rohrer, Taylor, & Sholar, 2010). Instructors hope that the information that they impart to students will be flexibly accessible and will generalize to different contexts. Therefore, it is crucial to determine the extent to which test-enhanced learning transfers and whether it has limitations (Anderson & Biddle, 1975; Hinze & Wiley, 2011). In a recent review of transfer and the testing-effect literature, Carpenter (2012) concluded that test-enhanced learning transfers well in three categories: (1) across temporal contexts (e.g., different retention intervals), (2) across test formats (e.g., question format alterations such as short-answer to multiple-choice), and (3) across knowledge domains (e.g., from biology to aeronautics). This classification scheme adheres to recommendations that research on transfer requires definitional clarity (Barnett & Ceci, 2002).

Of the categories delineated by Carpenter (2012), the second and third include studies in which the stimulus materials used on a final test are different from those used on prior tests; among these, transfer has been demonstrated for word lists (Carpenter & DeLosh, 2006), prose passages (Butler, 2010;

✉ Timothy C. Rickard
trickard@ucsd.edu

[1] Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109, USA

Published online: 01 September 2015

Chan, 2009, 2010; Chan, McDermott, & Roediger, 2006; Karpicke & Blunt, 2011), and science facts (Hinze, Wiley, & Pellegrino, 2013; McDaniel, Anderson, Derbish, & Morrisette, 2007; McDaniel et al., 2013; Rawson, Dunlosky, & Sciartelli, 2013). Test-enhanced learning also exhibits transfer for less commonly examined materials, such as math functions (Kang, McDaniel, & Pashler, 2011), spatial learning (Carpenter & Kelly, 2012), map learning (Rohrer et al., 2010), and even medical diagnoses (Larsen, Butler, Lawson, & Roediger, 2013). Moreover, transfer has been found when the difference between the originally tested and transfer test materials is relatively small (*near* transfer, involving questions on the same subject; e.g., detail questions, as in Hinze et al., 2013), as well as when it is quite large (*far* transfer, involving application or inference questions on separate topics; e.g., Chan et al., 2006; Johnson & Mayer, 2009; McDaniel, Howard, & Einstein, 2009).

Conversely, there have been some failures to find transfer of test-enhanced learning (e.g., Agarwal, 2011; Hinze & Wiley, 2011; Tran, Rohrer, & Pashler, 2015). Such studies have employed materials similar to those in the studies described above (e.g., textbook passages, science facts, logical premises), and the transfer measured has ranged from near to far. As of this writing, however, failures to find transfer of test-enhanced learning remain the exceptions in the literature.

Positive transfer for paired associates

Carpenter's (2012) second category of transfer also includes studies in which the same materials are presented on the initial and final tests, but in which the stimulus and response elements have been rearranged on the final test. Strong positive transfer of that type had been observed for paired-associate words. Carpenter, Pashler, and Vul (2006) administered cued-recall tests for paired-associate English word lists in one cue-to-target direction (e.g., *beach* → *blanket*) and observed substantial transfer to the reversed, previously untested direction (e.g., *blanket* → *beach*) on a delayed final test. Similarly, Vaughn and Rawson (2014) adapted the Carpenter et al. paradigm using English–English word pairs to assess criterion-level testing (cf. Vaughn & Rawson, 2011). They replicated the finding of positive transfer for the case of subjects trained to a criterion level of one correct trial for each tested item, although the extent of transfer decreased at higher levels of criterion learning.

In Carpenter et al. (2006) and Vaughn and Rawson (2014), transfer was assessed on the final test by the reversal of two elements (e.g., *cue* → *target* reversed to *target* → *cue*). This raises an important yet unaddressed question in the testing literature: Does test-enhanced learning transfer for stimuli with more than two elements, any of which could constitute the answer on the transfer test? This question has not been thoroughly addressed within the testing-effect paradigm using accuracy as the dependent variable, in particular for the type of transfer in which previously studied stimulus and response elements are rearranged on the final test. However, there is some related research, as summarized below.

Transfer research on triple associates

Kahana and Caplan (2002, Exp. 1) explored transfer for word triplets using serial presentation in which each of three words was presented, one at a time, per study trial (all three words were never shown at one time). That study phase was immediately followed by a final test in which either a single word or two words were presented, with subjects being asked to vocalize the missing word or words. A recall advantage, in terms of both accuracy and response time (RT), was observed for both single-word and two-word stimuli in the trained forward direction over the untrained reverse direction (i.e., performance on *A–B–?* was faster and more accurate than performance on *?–B–C*). Kahana and Caplan regarded this finding as evidence of associative asymmetry for the case of triplets, and noted its contrast to the associative symmetry that is characteristic of paired associates (Kahana, 2002). That study did not utilize a testing-effect paradigm, however (i.e., training did not involve a test), and the positive transfer was observed relative to no training. Thus, that finding does not directly address the possibility of transfer for tested materials relative to restudied materials.

Another line of work has investigated the transfer of test-enhanced learning for multi-element stimuli following extended retrieval practice. In these studies, transfer was evaluated relative to a control case of no training and using RTs as the dependent measure. Rickard, Healy, and Bourne (1994; see also Bajic, Kwak, & Rickard, 2011; Rickard, 2005; Rickard & Bourne, 1996) showed that, for both children and adults, RT gains following retrieval practice on arithmetic facts do not transfer to complement facts (e.g., the RT improvement for *5 × 7 =* __ does not transfer positively to *35 ÷ 7 =* __, or vice versa). Of more direct relevance to the present study, the same lack of transfer following extended retrieval practice has also been demonstrated for multi-element verbal stimuli.

Rickard and Bajic (2006) had subjects study a set of 24 word triplets (e.g., *cup, hand, tea*). Each triplet could form three test items, wherein a *test item* refers to a two-element test stimulus in which two of the three words are presented as cues and the missing third word is the to-be-recalled target; ignoring the spatial arrangement of the presented words, three test items are possible for each triplet (e.g., for the triplet *cup, hand, tea*, the three test items are Test Item 1, *cup, hand, ?*; Test Item 2, *cup, tea, ?*; Test Item 3, *hand, tea, ?*). In the subsequent training phase, there were 20 repetitions of testing with feedback for 18 test items from the 36 triplets and the remaining triplets received no training. On the final test, RT performance was assessed for all three possible test items per triplet, which yielded three conditions: (1) tested triplets with the trained response (*tested-identical* items), (2) tested triplets with an untrained response (*tested-inverted* items), and (3) triplets that were not trained at all (*untrained* items). Tested-identical items exhibited more than 650 ms of speedup. That learning, however, did not transfer positively to tested-inverted items; there was no difference in accuracy across the final-test conditions in those experiments (by design, accuracy was high for all test items on the final test).

Specificity of learning and testing on triple associates To explain the specificity of learning in the above RT studies, Rickard et al. (1994; Rickard & Bajic 2006) proposed an identical-elements (IE) model of retrieval practice effects for multi-element stimuli. According to that model, successful retrieval through a study representation establishes a new and separate recall representation for the particular stimulus–response configuration that was tested. That representation, which is assumed to form over the first few successful retrieval attempts, is accessible only for trained stimulus–response combinations. Over the course of retrieval practice, the recall representation becomes the dominant pathway to retrieval, and faster retrieval through that representation as a function of repetition accounts for the RT speedup. Thus, retrieval practice for the test item *cup, tea, ?* will result in substantial RT gains. However, when the tested-inverted item *tea, hand, ?* is later presented, the recall representation for *cup, tea, ?* cannot, according to the IE model, be accessed. Rather, answer retrieval for that test item must rely solely on the initially encoded study representation. The predicted and observed result is no positive transfer of practice effects to tested-inverted items. The IE model also correctly predicts the finding that RT improvement during the transition from reliance on the study representation to reliance on the hypothesized recall (test-based) representation does not follow the nearly ubiquitous power law (Newell & Rosenbloom, 1981) for data averaged over subjects, but rather adheres to a mixture function that also characterizes RT learning in other strategy transition contexts (Rickard, 1997; Rickard & Bajic, 2006).

The studies on multi-element stimuli summarized above and the IE model raise the possibility that transfer of learning will not occur for triplets under the testing-effect paradigm (i.e., assessing transfer in the context of performance on tested vs. restudied materials), despite evidence for positive transfer in that paradigm for paired associates. It is important to emphasize, however, that the aforementioned experiments on word triplets and arithmetic facts differ from those in the testing-effect literature in at least four critical respects that may yield divergent transfer results. First, the focus in most of those studies was transfer of learning as measured by RTs, whereas testing-effect studies primarily use accuracy as the dependent measure. It is plausible that RT gains are highly specific to practiced test items, whereas accuracy gains are not. Second, those studies either included no comparison of testing versus restudy (e.g., Rickard & Bajic, 2006), or testing was entirely absent during training (Kahana & Caplan, 2002). It is possible that transfer may be observed in a testing-effect paradigm that involves testing and restudy during training. Third, training in several of the aforementioned studies (e.g., Rickard et al., 1994; Rickard & Bajic, 2006) involved numerous repetitions (e.g., 20) of each trained test item, on which the observed RT-based specificity of learning effects might have depended. In the present experiments, transfer was assessed after only one test trial per test item. Finally, all experiments described above involved a test immediately after training. In the testing-effect literature, longer delays, such as the one-week delay in the present experiments, have been associated with increased testing-effect magnitudes (Roediger & Butler, 2011; Roediger & Karpicke, 2006a). It is possible that longer delays also yield greater transfer.

The present study

In the present work, we examined test-enhanced learning and transfer for triple associates, relative to a restudy control. In Session 1, subjects studied a complete set of 36 triplets. During the subsequent training phase, they restudied half of the triplets and were tested with feedback on the others. The final test occurred one week after training, and the dependent variable was accuracy (proportion correct). The critical question was whether the expected test-enhanced learning would transfer from tested-identical items to tested-inverted items (where a *test item* referred to a two-element test stimulus that takes the form of two presented words with one missing word to be recalled, as in Rickard & Bajic, 2006). To assess final-test performance on all three possible test items that were derivable from each triplet (i.e., *A, B, ?*; *A, C, ?*; and *B, C, ?*; the spatial order of elements was randomized), we used a three-block design. In each block, one test item (e.g., *A, B, ?*) per triplet was presented, and across all three blocks, each of the three test items (e.g., Test Item 1: *A, B, ?*; Test Item 2: *A, C, ?*; Test Item 3: *B, C, ?*) corresponding to each triplet was presented once.

To explore possible moderating factors on both the magnitude and transferability of the testing effect, in the first two experiments we examined the effects of testing on one versus two test items from each triplet during training. In Experiment 3, we assessed the effects of training on all three test items per triplet during training (thus complementing the first two experiments but not directly assessing transfer), as well as an instructional manipulation. The modifications across experiments were motivated in part by prior suggestions that testing on more than one variant of a stimulus (e.g., Goode, Geraci, &

Roediger, 2008) and employing instructional manipulations (e.g., Chan et al., 2006) can enhance the transferability of the testing effect.

Experiment 1

Subjects first studied 36 word triplets. Each triplet was then presented once again during training, half for restudy and half for testing with feedback. In this and all other experiments, all test trials during both training and the final test always involved presentation of two of the three words as the stimulus, with the third word having to be recalled. As we mentioned in the introduction, a given pair of word stimuli on a test trial is referred to as a *test item* (i.e., two of three words being presented, with the missing third word to be retrieved, constitutes one test item). After a one-week delay, a final test assessed recall for all 108 possible test items across the 36 studied triplets (three blocks, with one test item per triplet being presented in each block). The overall design was modeled closely on prior work (Carpenter et al., 2006; Rickard & Bajic, 2006).

Method

Subjects Forty-six University of California, San Diego, undergraduate students participated for course credit. All but four of the subjects completed both sessions of the experiment. The sample size selected for this study was comparable to those of prior laboratory studies focused on transfer (e.g., Hinze & Wiley, 2011; McDaniel et al., 2007).

Materials The stimuli consisted of 36 triplets, each containing three words of three to seven letters and one to two syllables in length (see the Appendix). All triplets were designed to facilitate the formation of an interactive mental image (e.g., *gift, rose, wine*).

Design and procedure In each session, subjects were individually seated at personal computers running Windows 7 or Windows XP (Microsoft, Redmond, WA) at a viewing distance of 30–40 cm from the computer screen. All experiments were programmed using Adobe Flash Professional CS6 (Adobe Systems, San Jose, CA) and presented using the Mozilla Firefox Web browser (Mozilla Foundation, Mountain View, CA) equipped with the Adobe Flash Player 12 plugin.

The first session consisted of two phases:

In the study phase, subjects read instructions stating that they were to "learn and memorize three concepts (words) at a time," and for each presented triplet to "use your imagination to visualize how the three concepts might interact with one another." After pressing the space bar to proceed, they were shown all 36 triplets one at a time for 8 s each. The order of presentation was random, and there was no delay between trials. Each triplet appeared in columnar fashion (one word above the other), in a large serif font (Garamond, size 40) at the center of the screen, with the instructions "study this triplet" in smaller font underneath. The ordering of each word per triplet (top, middle, or bottom) was randomized on each trial for each subject.

In the training phase, subjects were tested with feedback on one test item from each of 18 randomly selected triplets and restudied each of the remaining 18 triplets, for a total of 36 trials within one uninterrupted block. On each test trial, one test item from a triplet was shown (e.g., *wine, gift, ?*); on each restudy trial, all three words from a triplet were shown. Test and restudy trials were randomly interleaved. At the beginning of the phase, on-screen instructions appeared stating that each of the previously studied triplets would be presented with directions for either restudy ("you will be asked to study, which gives you more time to memorize") or testing with feedback ("you will be shown two words from a triplet and asked to type the third word"; "the correct word will then appear"). Subjects pressed the space bar to proceed to the first trial. For all trials, the stimulus was presented for 8 s, with no breaks between trials.

On trials involving restudy, triplets were presented in a columnar format identical to that of the study phase. On trials involving testing with feedback, test items also appeared in the columnar format, but with the following changes: (1) one word (the response to be retrieved) was removed, and *???* appeared in its place (with the word and *???* ordering in top, middle, or bottom positions randomized on each trial for each subject); (2) an empty text box with a blinking cursor appeared directly underneath the column; and (3) the instructions stated "type the third word of this triplet." Subjects had 6 s to type their answer into the text box, after which no new input was accepted and *???* was replaced by the correct word for 2 s, which constituted feedback. During that 2-s period, the two words of the test item and any typed characters continued to be displayed. Hence, both restudy and testing with feedback trials lasted for 8 s. At the conclusion of training, the experimenter reminded subjects to return at the same time the following week. To minimize the possibility of practice between sessions, subjects were also told that they would be learning a new set of test items.

The final test was administered in the second session. First, subjects were informed that their memory for the triplets that they had learned in the prior week was about to be tested. After pressing the space bar to proceed, subjects were tested on three 36-trial blocks, with one test item from each of the 36 triplets appearing per block, in random order. Thus, no test item was repeated over blocks. Consider the triplet *ant, spray, trash*. On the first final-test block for a given subject, the test item might be *spray, ant, ?*; on the second block, *ant, trash, ?*; and on the third block, *trash, spray, ?*.

Each final-test trial involved the following: (1) Two words were displayed, while the third was absent and replaced by a *???*; (2) the arrangement of those three stimulus components was in an upside-down triangle format, with two elements at the same (upper) level and the third element below. This change in test format from training to final test was made in order to minimize the possible influence of word spatial-location effects from training (as in Rickard & Bajic, 2006); (3) the locations of the three elements (two words and *???*) were randomized on each trial to any of the three positions; and (4) an empty text box with a blinking cursor appeared immediately below. Subjects had 15 s to type an answer into the text box before the program automatically advanced to the next trial. No feedback was provided, and there were no breaks between blocks. The experiment ended after subjects had completed all three blocks.

Across the three final-test blocks, each of the three possible test items for each of the 36 triplets was tested once, yielding 108 test items over 108 consecutive trials (see Table 1). In every block, 18 test items were drawn from triplets that had previously been trained using restudy (*restudied* items), and 18 test items were drawn from triplets that had previously been trained using testing. Six test items from the previously trained triplets were identical to the items that had been presented during training. For example, if the test item *tea, hand, ?* was presented during training, the test item *tea, hand, ?* was presented in the final test. Following Rickard and Bajic (2006), those items on the final test will be referred to as *tested-identical* items. Twelve final-test items from the previously trained triplets were inverted (i.e., one or both of the two presented words were not the same as on a previously trained test item). For example, if the test item *tea, hand, ?* was presented during training, the inverted test item *hand, cup, ?* was presented in the final test. Those items on the final test will be referred to as *tested-inverted* items. Thus, there were three test item type conditions on each block of the final test: tested-identical, tested-inverted, and restudied. Positive transfer of test-enhanced learning would correspond to higher accuracy in the tested-inverted condition than in the restudied condition.

Results and discussion

Training The mean accuracy over subjects for the 18 tested triplets was .60, *SE* = .032.

Final test A within-subjects factorial analysis of variance (ANOVA) was performed on the subject-level proportions correct (Fig. 2), with the factors Final-Test Condition (tested-identical vs. tested-inverted vs. restudied) and Block (1 vs. 2. vs. 3). We found highly significant effects of final-test condition, $F(2, 82) = 18.38$, $p < .001$, $MSE = 0.025$, $\eta_p^2 = 0.31$, and block, $F(2, 82) = 206.01$, $p < .001$, $MSE = 0.024$, $\eta_p^2 = .83$, as well as a significant interaction, $F(4, 164) = 3.83$, $p < .005$, $MSE = 0.021$, $\eta_p^2 = .085$. The main effect of block corresponds to the overall pattern of improvement during the final test, a result that will be discussed further below. The final-test condition by block interaction reflects the decreasing performance difference among final-test conditions over successive blocks.

Of greatest interest is the main effect of final-test condition. Inspection of Fig. 1 suggests that the training effect is primarily driven by the difference between the tested-identical condition and the other two conditions. A follow-up ANOVA limited to the tested-inverted and restudied conditions confirmed no significant main effect of final-test condition, $F(1, 41) = 0.41$, $p = .53$, $MSE = 0.021$, $\eta_p^2 = .0099$, and no significant interaction with block, $F(2, 82) = 2.12$, $p = .13$, $MSE = 0.011$, $\eta_p^2 = .049$. Thus, in marked contrast to the procedurally analogous paired-associate task of Carpenter et al. (2006), test-enhanced learning for triplets appeared to transfer minimally, if at all, to tested-inverted items.

The substantial performance improvement over the three final-test blocks (see Fig. 1) occurred despite that fact that no feedback was provided. That pattern, which we replicated in Experiments 2 and 3, implies transfer from one test item from a triplet to another across blocks, and thus appears to be inconsistent with the finding of little or no transfer from tested-identical to tested-inverted items within each block. Those two transfer patterns, however, are likely in our view to reflect independent memory phenomena. The lack of transfer to tested-inverted items after the one-week delay clearly reflects the specificity of learning in long-term memory, and that specificity is also likely to drive the more compressed condition effects in Blocks 2 and 3. The transfer across the three tested items from a triplet over blocks is more likely to reflect priming, or increased response strength (e.g., Bjork & Bjork, 1992), for individual elements, or perhaps “learning-to-learn” effects (Postman & Stark, 1967). On the first block, two of the three elements from each triplet were presented as the stimuli on each trial, and those elements always constituted the answers to the other two test items from the same triplet that would be presented on Blocks 2 and 3. It is likely in our view that priming of those elements on Block 1 made them more available for retrieval on Blocks 2 and 3. If priming for both of those stimulus elements was largely saturated on Block 1, then the performance improvement due to priming would be expected to be relatively large from Block 1 to Block 2, and smaller from Block 2 to Block 3, as was observed.

Experiment 2

Having found no evidence for the transfer of test-enhanced learning after training on one test item per triplet, we next examined the possibility that training on two different test items per triplet could enhance transfer performance on the

Table 1 Example final-test block design used in Experiment 1, with hypothetical triplet numbers for illustrative purposes. During the final test, one test item from each triplet was shown per block, and all 36 triplets appeared once per block. All three test items per triplet were tested over three blocks. *Final-test condition* indicates whether a test item was previously presented in the first session: *Tested-identical* items had been trained using testing with feedback, *tested-inverted* items had not been presented, and the complete triplet corresponding to each *restudied* item was restudied during training. The same final-test block design was also used in Experiments 2 and 3, with the sole difference being different numbers of tested-identical and tested-inverted items per block

Final-Test Block 1		Final-Test Block 2		Final-Test Block 3	
Triplet Number(s)	Final-Test Condition	Triplet Number(s)	Final-Test Condition	Triplet Number(s)	Final-Test Condition
1	*Tested-identical*	1	*Tested-inverted*	1	*Tested-inverted*
2	*Tested-identical*	2	*Tested-inverted*	2	*Tested-inverted*
3	*Tested-identical*	3	*Tested-inverted*	3	*Tested-inverted*
4	*Tested-identical*	4	*Tested-inverted*	4	*Tested-inverted*
5	*Tested-identical*	5	*Tested-inverted*	5	*Tested-inverted*
6	*Tested-identical*	6	*Tested-inverted*	6	*Tested-inverted*
7	*Tested-inverted*	7	*Tested-identical*	7	*Tested-inverted*
8	*Tested-inverted*	8	*Tested-identical*	8	*Tested-inverted*
9	*Tested-inverted*	9	*Tested-identical*	9	*Tested-inverted*
10	*Tested-inverted*	10	*Tested-identical*	10	*Tested-inverted*
11	*Tested-inverted*	11	*Tested-identical*	11	*Tested-inverted*
12	*Tested-inverted*	12	*Tested-identical*	12	*Tested-inverted*
13	*Tested-inverted*	13	*Tested-inverted*	13	*Tested-identical*
14	*Tested-inverted*	14	*Tested-inverted*	14	*Tested-identical*
15	*Tested-inverted*	15	*Tested-inverted*	15	*Tested-identical*
16	*Tested-inverted*	16	*Tested-inverted*	16	*Tested-identical*
17	*Tested-inverted*	17	*Tested-inverted*	17	*Tested-identical*
18	*Tested-inverted*	18	*Tested-inverted*	18	*Tested-identical*
19–36	Restudied	19–36	Restudied	19–36	Restudied

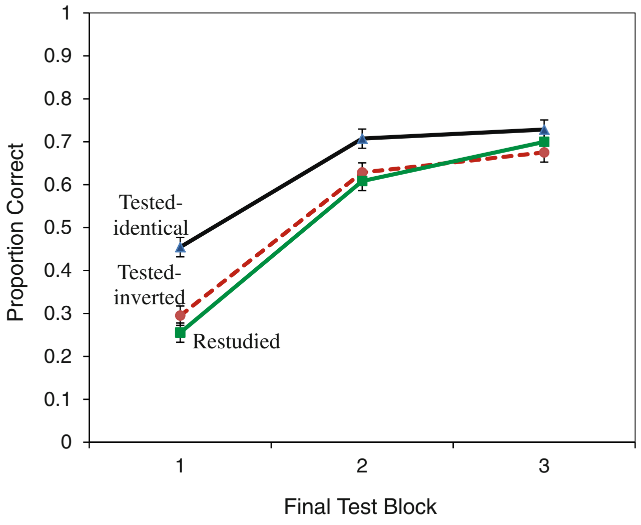

Fig. 1 Mean accuracy scores (proportions correct) on the final test of Experiment 1 as a function of final-test condition and block. Error bars indicate standard errors based on the interaction error term of a within-subjects analysis of variance on the subject mean accuracy scores (based on the method outlined by Loftus & Masson, 1994)

third test item. It may be that testing on two different test items yields triplet representations that are more flexibly accessible and can support transfer to the third. For example, training on a second test item of a triplet may reactivate memory for the first test item that was presented on the preceding block, which may result in a more integrated, or holistic, representation that can promote transfer to a third test item on the final test.

Method

Subjects The sample size was increased in this experiment. Sixty-one University of California, San Diego, undergraduate students participated for course credit, and all but three subjects completed both sessions of the experiment. On the basis of the standard deviation observed in Experiment 1, a sample size of 58 in this experiment yields statistical power greater than .90 to detect a proportion correct advantage of .05 or larger in the tested-inverted versus the restudy condition (based on a one-tailed test on the difference scores averaged over final test blocks, $\alpha = .05$).

Materials, design, and procedure The design of this experiment was nearly identical to that of Experiment 1, with the primary exception that a second 36-trial training block was added to the training phase. Eighteen of the trials in the second block involved a second restudy opportunity for the 18 triplets restudied in the first block. The remaining 18 trials in the second block involved testing on a different test item taken from the same triplet that had been tested in the first block (i.e., if *wine, gift, ?* was the test item in the first block, then either *rose, gift, ?* or *wine, rose, ?* was the test item in the second block). The position of each word (and *???* for test trials) remained randomized on each trial.

As in Experiment 1, there were three final-test blocks, each with 36 trials, such that all 108 possible test items were tested once over blocks. As a consequence of the changes to the training phase, however, each final-test block included 12 tested-identical items, six tested-inverted items, and 18 restudied items. Of the 12 tested-identical items per block, half each were from the first and second training blocks. The order of presentation remained randomized within each block.

The final test was also modified to feature no imposed time limit on each trial. This change avoided instances of subjects being in the midst of typing when a trial advanced, which had occurred on a few occasions in Experiment 1. Subjects were instructed to press the Enter key after they had finished typing their answer on each trial. To prevent skipped trials, the responsiveness of the Enter key on each trial was conditional on either (1) subjects having typed a minimum of three letters, or (2) 3 s of time having elapsed.

Results and discussion

Training The mean accuracies for tested items were .73, SE = .059 (first block) and .75, SE = .057 (second block). No significant difference in performance was apparent on the two blocks, $t(57) = 1.15$, $p = .25$, $d = 0.15$, suggesting minimal transfer of learning across training blocks from one test item to another from the same triplet. The apparent priming effect over the final-test blocks of Experiment 1 was not evident across the two training blocks of this experiment. That result, however, is not necessarily problematic for the priming account; it may be that all elements were already primed during the immediately preceding study phase, just as we hypothesized was the case after the first final-test block in Experiment 1.

Final test An ANOVA identical to that from Experiment 1 was performed (see Fig. 2), yielding highly significant main effects of final-test condition, $F(2, 114) = 32.20$, $p < .001$, $MSE = 0.024$, $\eta_p^2 = .36$, and block, $F(2, 114) = 196.80$, $p < .001$, $MSE = 0.026$, $\eta_p^2 = .78$, as well as a significant interaction, $F(4, 228) = 5.92$, $p < .001$, $MSE = 0.019$, $\eta_p^2 = .094$. These results closely replicated the findings of Experiment 1.

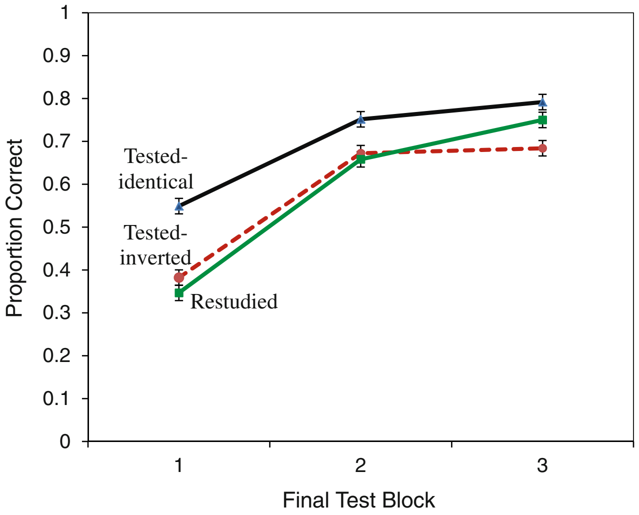

Fig. 2 Mean accuracy scores (proportions correct) on the final test of Experiment 2 as a function of final-test condition and block. Error bars indicate standard errors based on the interaction error term of a within-subjects analysis of variance on the subject mean accuracy scores

In the follow-up ANOVA limited to tested-inverted and restudied items, we found no main effect of final-test condition, $F(1, 57) = 0.076$, $p = .78$, $MSE = 0.034$, $\eta_p^2 = .0013$, as had been observed in Experiment 1. There was, however, a significant final-test condition by block interaction, $F(2, 114) = 4.12$, $p = .019$, $MSE = 0.020$, $\eta_p^2 = .067$. That interaction mirrors the same (but nonsignificant) crossover pattern observed in Experiment 1. Post-hoc t tests revealed a statistically significant performance advantage for restudied over tested-inverted items on the third block, $t(57) = 2.47$, $p = .017$, $d = 0.32$, but no significant difference between those conditions on the first or second block.

Overall, despite testing on two of the three possible test items during training, no evidence remained of positive transfer from tested-identical to tested-inverted items. These results diverge from the prior finding that encoding variability facilitates the transfer of test-enhanced learning (e.g., Goode et al., 2008), although that previous design involved anagram solving and can arguably be viewed as a skill-learning rather than a memory recall paradigm. Moreover, the results of this experiment again contrast sharply with prior work showing excellent transfer for paired associates.

Experiment 3

In the third experiment, we examined the effects of training on all three possible test items from each triplet involving two stimulus elements and one response. During the training phase, either a triplet was restudied three times or each of its three test items was tested once. Therefore, this experiment included no transfer manipulation. Rather, it served to evaluate the efficacy of testing, relative to restudy, when all possible test items were tested once during training, and it promoted cross-experiment analyses that provided insight into whether testing yields increasing or decreasing enhancement of learning as a function of the number of items tested.

In the training phase, we also examined the effects of adding reminders for subjects to use interactive images, as well as increasing the number of unique interactive images that subjects were instructed to form for each triplet. This manipulation was motivated by prior evidence that instructing subjects to form additional memory associations during training can enhance transfer performance (e.g., Chan et al., 2006, Exp. 3).

Method

Subjects Sixty-four University of California, San Diego, undergraduate students participated for course credit. All but four subjects completed both sessions of the experiment.

Materials, design, and procedure This experiment's design was identical to that of Experiment 2, with the exception of two changes to the training phase. First, training featured three blocks of 36 trials instead of two. For 18 of the triplets, restudy occurred on each of the three blocks. For the remaining 18 triplets, each of the three test items from the triplet was tested once across blocks. Thus, six tested-identical items per final-test block had been presented, respectively, in the first, second, and third blocks of training (constituting the 18 tested-identical items).

Second, we also explored whether the type of interactive imagery instruction might moderate either overall performance or the testing effect. There were three levels of instruction type during training, manipulated between subjects with random assignment: (1) standard interactive imagery, in the exact manner as in the preceding two experiments; (2) reinforced interactive imagery, in which subjects were reminded of the image-forming process at the start of the second phase of the training session; and (3) multiple interactive imagery, in which subjects were told to form new interactive images at the start of each training block.

Results and discussion

Training The mean accuracies (and *SE*s) for test items were .68 (.060), .78 (.054), and .72 (.058) on the first, second, and third blocks, respectively. A one-way within-subjects ANOVA revealed a significant difference in performance over training blocks, $F(2,\ 118) = 9.46$, $p < .001$, $MSE = 0.013$, $\eta_p^2 = .14$. Thus, in contrast to the training results for Experiment 2, accuracy in this experiment depended on the training block. The effect was nonmonotonic over blocks, however, and the difference in proportions correct between the first and third blocks was small (.04).

Final test A mixed ANOVA on subject-level mean accuracy scores with a between-subjects factor of Interactive Imagery (standard vs. reinforced vs. multiple) and within-subject factors of Final-Test Condition (tested-identical vs. restudied) and Block (1 vs. 2. vs. 3) was performed. In that analysis, no significant main effects or interactions involving the interactive imagery manipulation were observed (all *p*s ≥ .57). Neither reminders to use interactive imagery nor instructions to construct a new image for the test items presented on each block appreciably affected performance. Given prior work supporting the effectiveness of interactive imagery (Bower, 1970), those null results are surprising. One plausible account is that our triplet stimuli, which were designed to facilitate imagery, resulted in spontaneous use of imagery-based representation regardless of the instructional condition. Alternatively, the imagery manipulations may not have been heeded by subjects.

We therefore removed Imagery Instruction as a factor and performed a within-subjects ANOVA on data combined from subjects in all three imagery groups (Fig. 3). Analogously to Experiments 1 and 2, we found highly significant main effects of final test condition, $F(1,\ 59) = 67.060$, $p < .001$, $MSE = 0.021$, $\eta_p^2 = .53$, and block, $F(2,\ 118) = 187.10$, $p < .001$, $MSE = 0.012$, $\eta_p^2 = .76$, as well as a significant final-test condition by block interaction, $F(2, 118) = 52.23$, $p < .001$, $MSE = 0.0095$, $\eta_p^2 = .47$. These results confirm and extend the results of the prior experiments: A large testing effect was observed for triplets, regardless of the number of unique test items per triplet that were presented during training.

Cross-experiment analyses

To assess the cross-experiment patterns for the testing effect, we performed a mixed factorial ANOVA on the mean accuracy scores from the first final-test block of each experiment, with the between-subjects factor Experiment (1, 2, and 3) and the within-subjects factor Final-Test Condition (combined items vs. restudied items). In this analysis, *combined* refers to both tested-identical and tested-inverted items. By combining these two types of test items into the same condition and comparing it against the restudied condition, in this analysis we focused on a question of potential educational importance, namely: Does test-enhanced learning for a triplet in its entirety accelerate or decelerate, relative to restudy, when one, two, or all three test items per triplet are trained? In other words, what is the effect of taking tests on an increasing number of test items per triplet on performance for all three test items per triplet? For Experiments 1 and 2, the combined condition included data averaged over both tested-identical and tested-inverted items for each tested triplet. All three test items from each triplet were tested in Experiment 3, and hence all test items were classified into the combined condition.

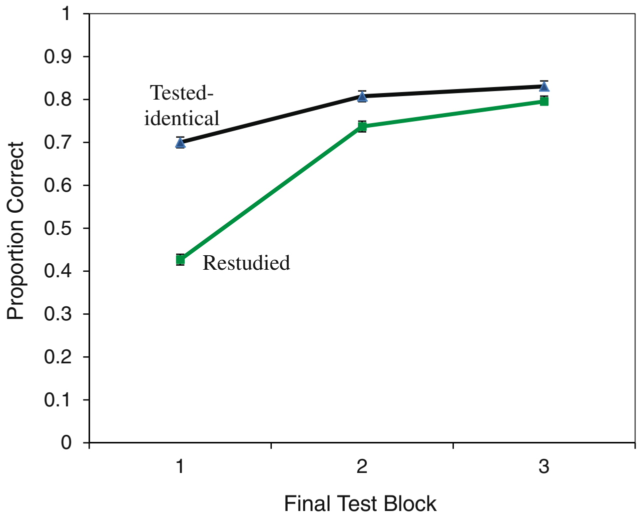

Fig. 3 Mean accuracy scores (proportions correct) on the final test of Experiment 3 as a function of final-test condition and block. Error bars indicate standard errors based on the interaction error term of a within-subjects analysis of variance on the subject mean accuracy scores.

The results are shown in Fig. 4. We observed highly significant main effects of experiment, $F(2, 157) = 26.15$, $p < .001$, $MSE = 0.068$, $\eta_p^2 = .25$, and final-test condition, $F(1, 157) = 143.37$, $p < .001$, $MSE = 0.018$, $\eta_p^2 = .48$, as well as a significant experiment by final-test condition interaction, $F(2, 157) = 13.31$, $p < .001$, $MSE = 0.018$, $\eta_p^2 = .14$. The main effect of experiment corresponds to the increased overall performance levels for both combined and restudied triplets as the number of training exposures increased in Experiments 1 to 3. The main effect of final-test condition corresponds to the overall advantage for combined versus restudied triplets. Inspection of Fig. 4 shows that the experiment by final-test condition interaction reflects an increasing difference between combined and restudied triplets over the progression from Experiments 1 to 3. Thus, the testing effect for overall triplet learning does not appear to be diminished (at least under the present conditions, in which ceiling effects are unlikely to constitute a confound), and may be accelerated as the number of test items per triplet increases. These results should be interpreted with caution, however, given that the assignment of subjects to experiments was not random.

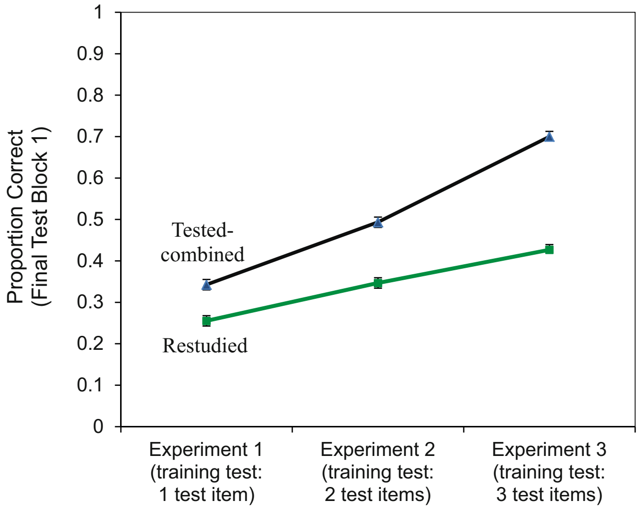

Fig. 4 Results of the cross-experiment analysis showing mean accuracy scores (proportions correct) on the first final-test blocks of Experiments 1, 2, and 3 for tested-identical and tested-inverted items (tested-combined) versus restudied items

General discussion

We explored test-enhanced learning relative to restudy for triple associates, as well as transfer of that learning to inverted test items. A robust testing effect was observed across all experiments, extending the effect to triple associates, and likely to the larger class of multi-element stimuli. Moreover, the testing effect occurred despite randomization of word order and spatial position between training and the final test, which eliminates a superficial perceptual-learning account. In stark contrast to the robust testing effect, we found no evidence for positive transfer from tested-identical to tested-inverted items relative to restudy, neither when one nor two of the three possible test items were tested during training.

The absence of transfer persisted across variations in the preexperimental associative strength over triplets. This was evident in supplementary analyses using the forward and backward associative strengths from available data (Nelson, McEvoy, & Schreiber, 1998) on all possible word pair combinations per triplet (i.e., *A–B*, *B–A*, *A–C*, *C–A*, *B–C*, and *C–B*), thus allowing us to calculate an average associative strength score for each triplet. A split-half analysis comparing triplets that scored relatively high versus low in associative strength (averages of .036 vs. .00079 in forward and backward associative strength across all possible word pair combinations per triplet) did not alter the transfer results observed in Experiments 1 and 2. We observed large testing effects for both low- and high-associative-strength items in both experiments, as well as a large main effect of associative strength in Experiment 1 (but not Exp. 2). Most importantly, in neither experiment did we see a trend toward an interaction between associative strength (high vs. low) and condition (restudy vs. transfer; tested condition was excluded), *ps* > .50.

Although the finding of no transfer relative to restudy was not anticipated by the testing-effect transfer literature, it is consistent with prior work on word triplets and arithmetic facts focused on RTs, as we discussed in the introduction. Given the wide range of stimulus types (both verbal and numerical), populations (both adults and children), skill levels (both single-trial testing in the present experiments vs. extended retrieval practice for adult arithmetic), prior learning (none in the present experiments vs. many hours for adult arithmetic), and dependent variables (proportion correct vs. RT) over which a high learning specificity for triplets following recall has been demonstrated (Rickard & Bajic, 2006; Walker, Bajic, Mickes, Kwak, & Rickard, 2014), it appears that a substantial degree of learning specificity due to testing is universal over various types of triple-associate stimuli.

Theoretical implications for triple associates

The present results suggest that the IE model, which was developed as an account of the specificity of RT learning for multi-element stimuli in the context of extended retrieval practice, may extend to the case of a single-retrieval trial and the dependent variable of proportion correct. As we noted in the introduction, that model assumes that correct retrieval creates a new recall representation that is independent of the study representation and can later be accessed on the final test only for the tested-identical item. The present work suggests that even a single correct retrieval trial can establish an independent recall representation that can support subsequent performance on the final test.

Because transfer in the present study was evaluated relative to restudy, we cannot conclude that entirely no transfer of learning to tested-inverted items took place in Experiments 1 and 2. Restudy presumably yielded some degree of additional learning (Roediger & Karpicke, 2006b). Thus, positive transfer of testing relative to a hypothetical no-training condition likely did occur in Experiments 1 and 2. The IE model does not anticipate that transfer. A slightly modified version of that model might nevertheless be consistent with the present results, given the possibility that feedback during training played less of a role in the previously described RT studies than in the present study. In the RT studies, accuracy was high throughout training, and feedback was typically provided only on incorrect trials. Feedback thus may have played little role in learning, and the learning due to correct retrieval may have been fully asymmetrical in the tested direction. In the present context of accuracy learning, on the other hand, feedback on incorrect trials is likely to have played an important role in enhancing the study memory that is assumed in the IE model. That enhancement might in turn support transfer to tested-inverted items, under the assumption that feedback processing yields roughly symmetrical triplet learning (much as initial study presumably does) that can produce transfer.

According to that account, the nearly equivalent overall performance in the tested-inverted and restudy conditions in Experiments 1 and 2 reflects a balance of minimal or no transfer (even relative to a hypothetical no-training control) following correct training test trials, and positive transfer following incorrect training test trials. Here, feedback is assumed to be critical for learning on incorrect trials but inconsequential for learning on correct trials (for supporting evidence, see Fazio, Huelser, Johnson, & Marsh, 2010; Pashler, Cepeda, Wixted, & Rohrer, 2005). There is also independent evidence that incorrect initial test trials with feedback yield more learning than does restudy (Butler & Roediger, 2008; Hays, Kornell, & Bjork, 2013).

Transfer as a function of initial test performance To gain insight into the viability of the hypothesis outlined above, we conducted post-hoc conditional analyses of the first final-test block data of Experiment 1. Specifically, we explored transfer on that block separately for tested items that had been answered correctly versus incorrectly on the initial test. Thirty-five subjects who had both correct and incorrect initial test observations for both tested-identical and tested-inverted items on the final test were included. The results, shown in Table 2, are consistent with the IE model outlined above. When the response for a tested item was incorrect on the initial test, proportions correct were nearly identical for the tested-identical and tested-inverted items on the final test, consistent with the hypothesis that feedback after an incorrect test trial produces symmetrical learning that supports transfer. In contrast, when the response was correct on the initial test, we found a very large specificity of learning effect, consistent with the IE model predictions.

In summary, the results of the conditional analysis support the hypothesis that, for triplets at least, feedback on an incorrect test trial promotes the transfer of learning, whereas correct answer retrieval (with or without feedback) promotes specificity of learning. A caveat to that conclusion is the possibility of selection bias. Test items that were answered correctly on the initial test presumably correspond, on average, to triplets that are relatively easy to learn, whereas test items incorrectly answered correspond to triplets that are more difficult to learn. That fact raises the possibility that the transfer results are causally related to the intrinsic difficulty of the test items. It is conceivable that triplets that are difficult to learn intrinsically yield strong transfer, regardless of whether they are answered correctly on the initial test, whereas triplets that are easy to learn intrinsically yield weak transfer. Although that possibility cannot be ruled out on the basis of the present data, there is no apparent mechanistic reason to expect it.

Also relevant to the present analyses is performance on restudy test items on the first final-test block. The mean for those test items overall ($M = .23$, $SE = .023$) was approximately the same as that for both tested-identical and tested-inverted items from triplets that were incorrectly solved on the initial test, as well as for tested-inverted items from triplets that had been correctly solved on the initial test (see Table 1). Those comparisons should be interpreted with caution, however, because restudied triplets cannot be grouped into those that would or would not have been correctly answered on a training phase test. If such grouping could be done, it would almost certainly yield a lower proportion correct (i.e., less than the .23 for overall restudied items) for restudied triplets that would have been answered incorrectly on an initial test, and a higher proportion correct (>.23) for restudied triplets that would have been answered correctly on an initial test. Using that logic, we

Table 2 Conditional analyses of mean proportions correct (with *SE*s) for tested-identical and tested-inverted items on Block 1 of the final test, by training phase test results in Experiment 1

Training Phase Test Correct		Training Phase Test Incorrect	
Tested-Identical Items	Tested-Inverted Items	Tested-Identical Items	Tested-Inverted Items
.61 (.050)	.30 (.034)	.26 (.051)	.23 (.050)

can draw the following tentative conclusions for word triplets: (1) An incorrect initial test followed by correct answer feedback produces roughly symmetric learning that is accessible for both tested-identical and tested-inverted (transfer) items, and (2) correct retrieval (with or without feedback) produces substantial learning for the tested items but minimal transfer of that learning to tested-inverted items.

Transfer for triple associates versus paired associates

The present findings for triple-associate word stimuli contrast strikingly with prior findings of robust transfer for cue–response reversals in the case of paired associates, provided that training did not yield a high level of criterion learning (e.g., Carpenter et al., 2006; Vaughn & Rawson, 2014). Given the design and procedural similarities to the Carpenter et al. study, it appears that the contrasting results stem from the properties of the stimuli themselves (pairs vs. triplets), rather than from other factors.

Why does presenting both a prior stimulus element and a prior response element as retrieval cues—as was the case for tested-inverted items in this study—prevent access to the learning that occurred on the initial test, whereas that learning can be accessed in the case of pure stimulus–response reversal for paired associates? The IE model as it was originally developed and described above does not explain that difference. However, a revised version of the IE model that was proposed to explain positive RT transfer results between multiplication and factoring (Rickard, 2005) provides one candidate account. For those two tasks, the numerical elements of the cues (e.g., *4 × 7 = ?* or *28* = __ ×__) are fully reversed for multiplication versus factoring. The transfer item in that case can be considered a pure reversal of the critical numerical stimulus and response elements (the only elements that varied over items), much like paired associates. To account for the multiplication–factoring transfer, the revised IE model incorporated bidirectional associative links between the stimulus and response elements of a tested item. Thus, in the case of a pure reversal such as multiplication and factoring, positive transfer is expected. That version of the IE model maintained the critical property of the original IE model, however, in that the stimulus elements on the transfer test must be an exact match to either the previously tested stimulus (or, in the revised model, the response) for transfer to occur. Hence, training on, for example, __ × __ = *28* should transfer to a "pure numerical reversal"—for example, *4 × 7* = __—as was observed. However, training on any of the following problems (e.g., *4 × 7* = _, *28 ÷ 4* = _, *28 ÷ 7* = _) should transfer to neither of the other two problems in that set, as was observed in several studies noted in the introduction. That version of the IE model can thus account for both the strong transfer for paired associates in the testing-effect paradigm and the poor transfer for triplets observed here. Nevertheless, that model leaves open the important question of why an exact match of the transfer stimulus elements to either the previously encountered stimulus elements or the previously encountered response element appears to be necessary for positive transfer (relative to restudy) to occur. It also does not explain the recent finding of Vaughn and Rawson (2014) that the extent of transfer for paired associates appears to decrease with increasing levels of mastery.

The difference in the observed transfer patterns for triple associates versus paired associates can also be viewed from the perspective of at least two other theoretical accounts of the testing effect: the elaborative retrieval model (Carpenter, 2009) and the mediator effectiveness model (Pyc & Rawson, 2010). According to the elaborative retrieval model, memory retrieval during testing activates semantic associative paths between the two cue words (in the present experiments) to the response word, potentially resulting in multiple retrieval pathways and facilitating subsequent improved retrieval, relative to restudy. From the perspective of the mediator effectiveness model, a similar process occurs during training tests: Mediators that link cues to targets are more effective on training trials involving a test than on trials involving restudy, thus enhancing later recall of tested items (Vaughn & Rawson, 2014). Neither of those theories in their current forms make strong claims about transfer for triplets. They do, however, explain testing effects at an associative level that lends itself to the development of transfer accounts, and the mediator effectiveness hypothesis has been discussed as a candidate account of the finding for paired associates that asymmetry holds at higher mastery levels (Vaughn & Rawson, 2014).

Educational implications

Following the recent progression from laboratory materials to applied materials in other testing-effect domains, it will be important to determine whether the present results for triple associates generalize to educationally relevant multi-element factual materials in domains such as history (e.g., *who*-, *what*-, *where*-, *when*-, and *how*-type facts) and biology, among many others. Although triplet stimuli of the type used in this study do not typically appear in classroom contexts, commonly used short answer and fill-in-the-blank test questions often have an analogous multi-element structure (e.g., in the question *Winston*

Churchill was the Prime Minister of what country during World War II?, the terms *Winston Churchill*, *Prime Minister*, *United Kingdom*, and *World War II* form a quartet). Such questions are often included in textbooks as practice problems and are a natural application of test-enhanced learning in future educational interventions. It remains to be determined whether transfer of test-enhanced learning from one question to another requiring a different response would occur in that case.

Author Note Steven C. Pan, Department of Psychology, University of California, San Diego; Carol M. Wong, Department of Psychology, University of California, San Diego; Zachary E. Potter, Department of Psychology, University of California, San Diego; Jonathan Mejia, Department of Psychology, University of California, San Diego; Timothy C. Rickard, Department of Psychology, University of California, San Diego.

This study was completed in partial fulfillment of the requirements for S. C. Pan's M.A. degree in Experimental Psychology at the University of California, San Diego. S. C. Pan is supported by a National Science Foundation (NSF) Graduate Research Fellowship. The authors sincerely appreciate helpful comments from Andrew Butler, Shana Carpenter, and an anonymous reviewer. Thanks also to Christine Diao, Arpita Gopal, and Annie Vasishta for assistance with data collection.

Appendix

Table 3 Triple associate word lists used in Experiments 1, 2, and 3

ants	spray	trash
bark	dog	stick
bat	dark	cave
beat	drum	march
bleach	stain	shirt
boy	run	field
bus	coins	line
cheer	game	score
clock	rush	late
cloth	soap	sink
cow	grass	milk
doctor	pills	note
farm	sun	sweat
gift	rose	wine
girl	smile	flower
ground	snail	mist
honk	cab	traffic
knight	castle	sword
lion	hunt	meat
map	hike	water
mug	hand	tea
paper	ink	desk
phone	ear	sound
plane	nap	blanket
room	key	door
ship	horn	sea
skate	fall	knee
sky	bird	rain
sofa	laugh	friends
street	bike	car
teeth	bite	wolf
towel	swim	pool
tree	child	swing
voice	sing	guitar
wall	paint	frame
warm	bread	coffee

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 database. (UMI No. 3468823)

Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 9, pp. 89–132). New York: Academic Press.

Bajic, D., Kwak, J., & Rickard, T. C. (2011). Specificity of learning through memory retrieval practice: The case of addition and subtraction. *Psychonomic Bulletin & Review, 18,* 1148–1155. doi:10.3758/s13423-011-0151-4

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

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Josslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale: Erlbaum.

Bower, G. H. (1970). Imagery as a relational organizer in associative learning. *Journal of Verbal Learning and Verbal Behavior, 9,* 529–533.

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. doi:10.1037/a0019902

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36,* 604–616. doi:10.3758/MC.36.3.604

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1563–1569. doi:10.1037/a0017021

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21,* 279–283. doi: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,* 268–276. doi:10.3758/BF03193405

Carpenter, S. K., & Kelly, J. W. (2012). Tests enhance retention and transfer of spatial learning. *Psychonomic Bulletin & Review, 19,* 443–448. doi:10.3758/s13423-012-0221-2

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. doi:10.3758/BF03194004

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. doi:10.1016/j.jml.2009.04.004

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

Chan, J. C., McDermott, K. B., & Roediger, H. L., III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135,* 553–571. doi:10.1037/0096-3445.135.4.553

Fazio, L. K., Huelser, B. J., Johnson, A., & Marsh, E. J. (2010). Receiving right/wrong feedback: Consequences for learning. *Memory, 18,* 335–350. doi:10.1080/09658211003652491

Goode, M. K., Geraci, L., & Roediger, H. L., III. (2008). Superiority of variable to repeated practice in transfer on anagram solution. *Psychonomic Bulletin & Review, 15,* 662–666. doi:10.3758/PBR.15.3.662

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 290–296. doi:10.1037/a0028468

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory, 19,* 290–304. doi: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,* 151–164. doi:10.1016/j.jml.2013.03.002

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101,* 621–629. doi:10.1037/a0015183

Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory & Cognition, 30,* 823–840. doi:10.3758/BF03195769

Kahana, M. J., & Caplan, J. B. (2002). Associative asymmetry in probed recall of serial lists. *Memory & Cognition, 30,* 841–849. doi:10.3758/BF03195770

Kang, S. H. K., McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review, 18,* 998–1005. doi:10.3758/s13423-011-0113-x

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborate studying with concept mapping. *Science, 331,* 772–775. doi:10.1126/science.1199327

Larsen, D. P., Butler, A. C., Lawson, A. L., & Roediger, H. L. (2013). The importance of seeing the patient: test-enhanced learning with standardized patients and written tests improves clinical application of knowledge. *Advances in Health Sciences Education, 18,* 409–425. doi:10.1007/s10459-012-9379-7

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1,* 476–490. doi: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,* 494–513. doi: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. doi:10.1111/j.1467-9280.2009.02325.x

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. doi:10.1002/acp.2914

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. Available at www.usf.edu/FreeAssociation

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale: Erlbaum.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 3–8. doi:10.1037/0278-7393.31.1.3

Postman, L., & Stark, K. (1967). Studies of learning to learn: IV. Transfer from serial to paired-associate learning. *Journal of Verbal Learning and Verbal Behavior, 6,* 339–353.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330,* 335. doi:10.1126/science.1191465

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,* 523–548. doi:10.1007/s10648-013-9240-4

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General, 126,* 288–311. doi:10.1037/0096-3445.126.3.288

Rickard, T. C. (2005). A revised identical elements model of arithmetic fact representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 250–257. doi:10.1037/0278-7393.31.2.250

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. doi:10.1037/0278-7393.32.4.734

Rickard, 1. 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. doi: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. doi:10.1037/0278-7393.20.5.1139

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

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

Roediger, H. L., III, & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249–255. doi:10.1111/j.1467-9280.2006.01693.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. doi:10.1037/a0017678

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

Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion-level effects on memory: What aspects of memory are enhanced by repeated retrieval? *Psychological Science, 22,* 1127–1131. doi:10.1177/0956797611417724

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: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:10.1016/j.jecp.2013.11.018