Memory & Cognition
https://doi.org/10.3758/s13421-020-01048-y

Test-enhanced learning for pairs and triplets: When and why does transfer occur?

Timothy C. Rickard [1] · Steven C. Pan [1,2]

Abstract

In four experiments, we explored conditions under which learning due to retrieval practice (i.e., testing) transfers to the case in which the cue and response words are rearranged (e.g., a training test on *gift, rose, ?*, wherein the target is *wine*, and a final test on *gift, ?, wine*, wherein the answer is *rose*). In both Experiment 1 and a supplementary experiment, we observed divergent results for pairs and triplets: Relative to a restudy control condition, strong transfer was observed for pairs, but none for triplets. In Experiments 2 and 3, the theoretical basis of the specificity of learning for triplets was explored. The results rule out the possibilities that transfer is wholly absent for triplets and that transfer occurs only for the case of exact cue–response reversal on the final test. Rather, it appears that, for both pairs and triplets, transfer *will* occur unless both of the following conditions hold: (1) two or more independent cues are presented on the training test, and (2) the correct responses on the training and final tests are different. We show that the majority of the results can be explained by combining the dual-memory theory of the testing effect with an inclusive-OR representation that forms when two or more cues are presented on the training test. Follow-up analyses that were conditionalized on training test accuracy suggest that specificity of learning is greater on a correct than on an incorrect training test trial, although selection confounds and contradictory experimental results preclude a strong conclusion.

Keywords Memory · Testing effect · Retrieval practice · Transfer · Paired associates · Triplets

Attempting to retrieve material from memory, as occurs when taking a practice test, generally improves subsequent recall relative to both a no reexposure and a restudy control condition. Known as the *testing effect, test-enhanced learning*, and the *retrieval practice effect,* this memory benefit has been demonstrated for materials ranging from vocabulary to photographs (Rawson & Dunlosky, 2011; for reviews see Delaney, Verkoeijen, & Spirgel, 2010; Kornell & Vaughn, 2016; Rickard & Pan, 2018; Roediger & Butler, 2011; Roediger & Karpicke, 2006). Many cognitive and educational psychologists consider retrieval practice to be one of the most potent and effective evidence-based learning techniques known to date, with potential applications for different grade levels and across a wide range of topics (e.g., Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Pashler, Bain, et al., 2007).

For the case of cued recall, which constitutes about 40% of studies in the testing effect literature (Rickard & Pan, 2018; Rowland, 2014) and is the focus of the current study, one basic question is whether test-enhanced learning exhibits transfer, relative to a restudy control, when cue and response words are rearranged (henceforth, *CR rearrangement*) between the training and final tests. There is strong empirical support for such transfer for word pair sets (henceforth, the term *set* refers to a particular pair or triplet of words). In Carpenter, Pashler, and Vul (2006), subjects first studied word pairs (e.g., *beach, blanket*). In the subsequent *training phase*, half of those sets and were restudied and the remaining half were tested with correct answer feedback (henceforth, feedback; e.g., *beach, ?*). On a *final test* that took place an average of 33 hrs later, there was 83% to 100% transfer of learning to the *tested-reverse* condition (e.g., *blanket, ?*); that is, performance (measured as proportion correct) was nearly identical in the *tested-same* (e.g., *beach, ?*) and tested-reverse conditions, and in both cases was much better than performance in the restudy control condition.

Data and materials for this study are accessible via the Open Science Framework (https://osf.io/95b6r/).

✉ Timothy C. Rickard
trickard@ucsd.edu

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

[2] Present address: Department of Psychology, University of California Los Angeles, Los Angeles, CA, USA

Published online: 03 June 2020

Similar transfer for pairs has been reported by Vaughn and Rawson (2014) for the case in which subjects were trained to a criterion level of one correct trial per set, although they also concluded that the extent of that transfer may decrease at higher learning criterion levels. Those results suggest that bi-directional (although not necessarily symmetrical) associative strengthening occurs on the training phase test for paired associates. See Kahana (2002) for evidence that a paired association is symmetric after study only.

In contrast, for materials such as word triplets and facts, there is evidence for minimal or no transfer of test-enhanced learning to CR rearranged sets relative to restudy. For example, in Pan, Wong, Potter, Mejia, and Rickard (2016, Experiment 1), a set of word triplets was first studied (e.g., *gift, wine, rose*). Each triplet was then either restudied (e.g., *gift, wine, rose*) or tested with feedback for recall of one of the words (e.g., *gift, wine, ?*). On a final cued recall test 7 days later, performance on CR rearranged sets involving two word cues, one of which was the prior response (e.g., *?, wine, rose*), was indistinguishable from performance in the restudy condition despite a large standard testing effect (i.e., in the tested-same vs. restudy conditions). That complete lack of transfer relative to the restudy control has also been demonstrated across several experiments for history and biology facts. For example, in Pan, Gopal, and Rickard (2015, Experiment 1), testing with feedback on multiterm history facts (e.g., *"Thomas Jefferson purchased the Louisiana territory from the ____?"*, for which the answer is *French*) yielded a large testing effect on a final cued recall test for short-answer questions assessing the same response (e.g., the answer, *"French"*). However, as was the case for triplets, performance on final test questions assessing different responses (e.g., the answer, *"Louisiana"*) was no better than that in the restudied condition (for related work, see Hinze & Wiley, 2011; Pan, Hutter, D'Andrea, Unwalla, & Rickard, 2018; Pan & Rickard, 2017; cf. McDaniel, Anderson, Derbish, & Morrisette, 2007; McDaniel, Bugg, Liu, & Brick, 2015).

The contrasting transfer results for pairs on one hand, and triplets and facts on the other, represent a potentially important divergence point in the testing effect literature and may reflect fundamental properties of memory. Strong inference along those lines is not yet possible, however, because to date no randomized studies exploring transfer for paired versus multielement sets have been conducted, and because alternative forms of CR rearrangement for triplets have not been explored. In Experiment 1 of this manuscript (and in a separate experiment described in the Appendix), the results of experimentally controlled comparisons of pairs and triplets are reported, confirming the expected transfer divergence. In Experiments 2 and 3, alternative forms of CR rearrangement for triplets were used to explore several candidate theoretical accounts of test-enhanced learning and transfer for those materials (see Fig. 1 for CR rearrangements between practice and final tests that were used in the current experiments). The results inform our proposal of a unified model of transfer of test-enhanced learning for pairs, triplets, and facts.

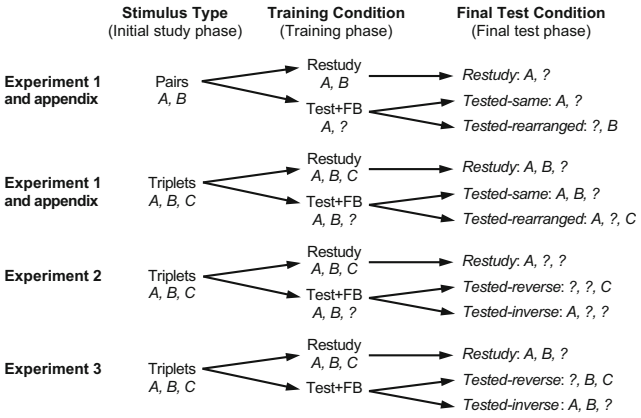

Fig. 1 Schematic of stimulus types examined across Experiments 1–3. Pairs versus triplets were investigated in Experiment 1 and in a supplemental experiment, and triplets only in Experiments 2–3. Italicized letters (*A, B, C*) represent stimulus words; Test+FB refers to testing with feedback

Experiment 1

In the first experiment, transfer for word pairs and triplets was investigated using a fully randomized design and with the same words (counterbalanced) for both tasks. As in prior work, one word was to be retrieved on both the training and final test for both pairs and triplets. The delay interval between the practice and final tests (24 hrs or 1 week) was also manipulated to investigate whether transfer effects following retrieval practice are consistent across different retention intervals.

Method

Subjects One hundred and twenty-seven undergraduate students participated for course credit. Data from eight were excluded due to noncompletion of the experiment or experimenter error. Among the 119 remaining subjects, 61 were in one of the triplets groups (24-hr delay: $n = 32$; 1-week delay: $n = 29$) and 58 were in one of the pairs groups (24-hr delay: $n = 33$; 1-week delay: $n = 25$). For triplets, at least 32 subjects in each group should yield statistical power of about 0.8 to detect a final test proportion correct difference between the tested-inverted condition and the restudy condition of at least .05. That estimate, which was generated using G*Power (Version 3.1.9.2; Faul, Erdfelder, Lang, & Buchner, 2007), is based on the data variability that was observed for that comparison in Pan et al. (2016, Experiment 1), using an upper-tail critical t test on the difference scores at $\alpha = 0.05$. Similar power is expected for paired associates.

Design and procedure The experimental design involved three phases, reflecting the most commonly used retrieval practice paradigm: a study phase in Session 1, a training phase in Session 1, in which restudy versus testing with feedback was manipulated, and a final test phase in Session 2. On the final test, there were three crossed independent variables: stimulus type (pairs or triplets; between subjects), the delay between sessions (24-hr or 1 week; between subjects), and final test condition (*tested-same* vs. *tested-rearranged* vs. *restudy*; within subjects). Within each level of delay interval, subjects were randomly assigned to either the pairs group or the triplets group. Subjects in the 24-hr and 1-week delay groups, although sampled from the same subject pool, were not run concurrently. In all other respects, subjects in the two delay groups received identical treatment.

The relative spatial positions of the words for each set were varied randomly across the initial study, training, and final test phases. Thus, spatial position of presented words was not a valid retrieval cue on the final test. Example images of the pairs and triplets conditions across each of the three phases are presented in Fig. 2. The three phases occurred as follows.

Study phase In the study phase, subjects read instructions stating that they were to memorize the set of words (i.e., a pair or a triplet) presented on each trial, and to promote learning, to link the concepts represented by those words via interactive imagery. They were then shown all 36 sets, one at a time, for 8 s each. On each trial here and throughout the experiment, all words for each set were presented simultaneously. The ordering of sets over trials was random, and there was no delay between trials. Each pair or triplet appeared in columnar fashion and in large serif font (40-pt. Times New Roman) at the center of the screen (see Fig. 2). The columnar position of the words (i.e., top, middle, or bottom of the column) was determined randomly on each trial.

Training phase In the training phase, subjects were tested with feedback on 18 of the sets (pairs or triplets) and restudied the remaining 18 sets, all in random trial order, for a total of 36 trials within one uninterrupted block. Trials for both tested and restudied sets lasted for 8 s. On restudy trials, the stimuli had a columnar format identical to that of the study phase. On test trials, a columnar format was also used, but one word of each pair (or one word of each triplet) was replaced by *???*, indicating the answer to be retrieved. An empty text box appeared directly underneath the presented words (see Fig. 2), and subjects had 6 s to type their answer into it, after which no new input was accepted, and "*???*" was replaced by the correct response word for 2 s, constituting feedback. During that feedback period, the full stimulus and any typed characters continued to be displayed. For both restudied and tested sets, columnar word order was randomly determined anew on each trial.

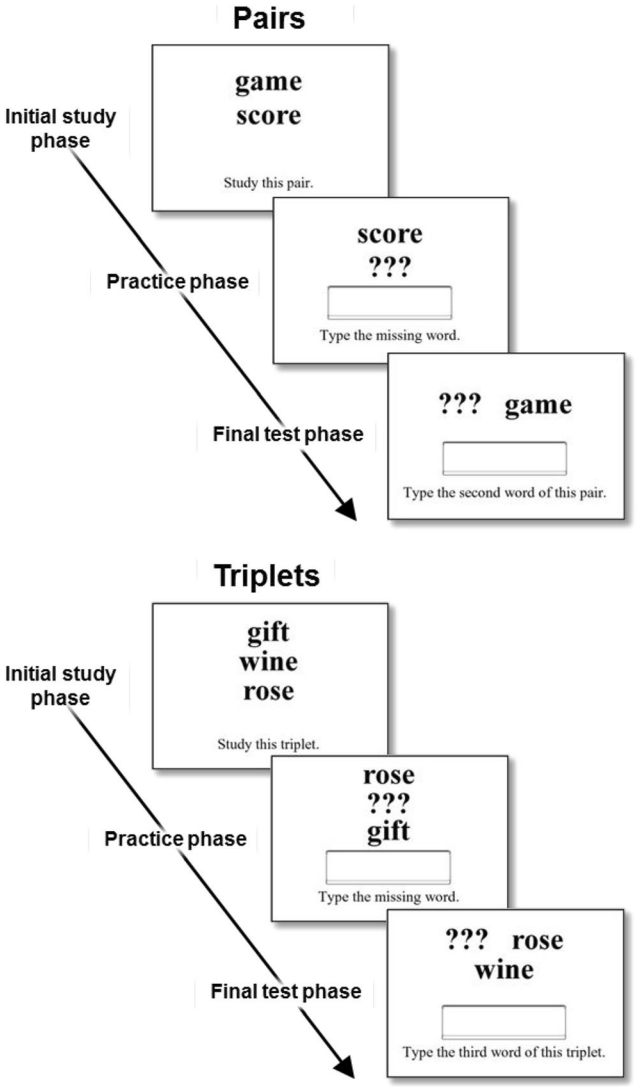

Fig. 2 Example stimulus presentation for a single word pair or triplet across all three phases of Experiment 1 (with intervening trials on other word pairs or triplets omitted). The tested-rearranged condition is shown for both pairs and triplets. Spatial order of stimulus elements (e.g., a given word from a pair or triplet might appear at top or bottom, left or right) was randomized on every trial in this experiment. Top: word pairs. Bottom: triplets

Final test phase On the final cued recall test, subjects' memory for the previously presented word pairs or triplets were assessed on two 36-trial blocks, with each pair or triplet assessed once per block, in random order. For pairs, each trial involved horizontal presentation of one word and *"???,"* with word order newly randomized. A text box appeared immediately below, in which subjects typed their answer. For triplets, each trial involved presentation of three elements: two words and *"???,"* of which two elements were horizontally presented and the third centered immediately below, again with word order newly randomized. Examples of both are shown in Fig. 2. The switch in spatial arrangement of words from training to the final test phases was a further measure to discourage an

attempted strategy (which could not be successful) of using spatial order of the elements as presented during initial study and training as a retrieval cue. Subjects had unlimited time to respond on each trial. No feedback was provided on the final test, and there were no breaks between blocks.

In the first final test block, half of the presented pairs or triplets (18) had been restudied during practice (restudy condition), while the remaining half (18) had been tested. Of those tested, half (9) were tested for the same response as during practice (tested-same condition), while the other half (9) were tested for a different response (tested-rearranged condition). In the second block, all 36 stimuli were tested again, but with a different missing word to be retrieved; tested sets that were presented in the tested-same direction on the first block were presented in the tested-rearranged direction in the second block, and vice versa.

Materials, counterbalancing, and scoring One hundred and eight common words of three to seven letters and one to two syllables in length, drawn from Pan et al. (2016), were used. Thirty-six paired associates were also created from those words by random selection of two words per triplet. The cue and response words for each triplet and pair were fully counterbalanced over subjects.

For the triplet groups, six training phase lists were created, each containing all 36 triplets. For each list, half of the triplets (randomly determined) were presented for testing (with one missing word to be retrieved) during training, and half were presented (complete) to be restudied. That assignment of triplets was counterbalanced such that the 18 triplets presented for testing in Lists 1–3 were presented for restudy in Lists 4–6, and vice versa. Across Lists 1–3, and also across Lists 4–6, each tested triplet was presented with a different missing word on the training test (three words per triplet, therefore three possible missing words). Thus, the six lists encompassed all three iterations of one missing word from each triplet when assigned to be tested during training, as well as the corresponding assignment of the other half of triplets assigned to be restudied. One training list was assigned to each subject randomly, and each list was used once per sequential group of six subjects. For the pairs groups, six practice lists were also created, each containing 36 word pairs. Assignment of lists to testing or restudy, as well as the missing word per pair on the training test, was counterbalanced and randomized in a manner analogous to that used for the triplet lists. For both pairs and triplets, training and final test trials were scored as correct only if the missing word was typed with no errors.

Results and discussion

Training phase test performance Mean proportion correct on the training test ranged between 0.56 and 0.70 across the four groups (see Table 1). A factorial between-subjects analysis of variance (ANOVA) yielded a significant effect (at $\alpha = 0.05$) of stimulus type (pairs vs. triplets), $F(1, 114) = 5.21$, $p = 0.024$, $\eta_p^2 = 0.042$, suggesting that triplets were somewhat easier to learn during the study phase than were pairs. There were no significant effects of delay (24 hr vs. 1 week) or the Delay Interval × Stimulus Type interaction ($ps > .28$), as was expected, given that the training phase occurred prior to the delay manipulation.

Final test performance Preliminary analysis showed minimal differences in relative condition performance across the two final test blocks. Given that the first final test block is the purest measure of memory and transfer after the delay interval, all analyses here and below were performed on first block data only. Results are shown in Fig. 3.

An ANOVA with the factors delay (between subjects), stimulus type (between subjects), and final test condition (tested-same vs. tested-rearranged vs. restudy; within subjects) confirmed large main effects of delay, $F(1, 115) = 59.02$, $p < .0001$, $\eta_p^2 = 0.33$, indicating more forgetting in the 1-week than in the 24-hr delay group, and final test condition, $F(2, 230) = 48.54$, $p < .0001$, $\eta_p^2 = 0.30$, but no main effect of stimulus type, $F(1, 115) = 0.04$, $p = .84$, $\eta_p^2 < 0.001$. The latter result shows that overall retrieval accuracy for pairs and triplets was statistically indistinguishable on the final test.

There were also no significant interactions involving delay ($ps > .11$), including no three-way interaction, suggesting that the transfer effects are robust across the two retention intervals. Nevertheless, the relatively large numerical difference in the extent of transfer for pairs in the 24-hr and 1-week delay groups (see Fig. 3) raises the possibility that transfer for pairs may in fact decrease with increasing delay interval. That hypothesis, however, is weakened by the incomplete transfer for pairs in the 24-hr delay experiment reported in the Appendix, transfer which is numerically similar to that of the 1-week delay group in Experiment 1. The simplest account of those results may thus be sampling variability over experiments.

Most critically, the main effect of final test condition was qualified by a significant interaction with stimulus type, $F(2, 230) = 6.26$, $p = .002$, $\eta_p^2 = 0.05$. That result confirms, under randomized conditions, the sharply contrasting extent of transfer of test-enhanced learning for pairs versus triplets. For pairs, there was substantial transfer relative to restudy for the tested-rearranged condition. For triplets, there was no evidence of transfer to the tested-inverted condition relative to restudy.

With respect to our goals and conclusions in this paper, the finding of incomplete transfer for pairs is not critical. Rather, the critical findings that motivate the remainder of this paper are that (a) there is substantially more transfer for pairs than for triplets, and (b) for triplets there appears to be performance equivalence for the tested-rearranged and restudy conditions, and hence little or no transfer of learning relative to restudy.

In a supplementary experiment described in the Appendix, we found that the results of Experiment 1 replicated almost

Table 1 Experiments 1–3 first block training and final test mean proportion correct (*SE*) results

Experiment	Group	Training test performance	Final test performance		
			Tested same or reverse	Tested rearranged or inverse	Restudy
1	Pairs 24-hr	0.65 (0.035)	0.69 (0.035)	0.67 (0.046)	0.49 (0.034)
	Triplets 24-hr	0.70 (0.041)	0.64 (0.050)	0.51 (0.043)	0.50 (0.041)
	Pairs 1-week	0.56 (0.039)	0.39 (0.047)	0.28 (0.043)	0.19 (0.032)
	Triplets 1-week	0.60 (0.045)	0.45 (0.045)	0.29 (0.039)	0.27 (0.026)
2	Triplets	0.65 (0.035)	0.25 (0.031)	0.40 (0.033)	0.27 (0.035)
3	Triplets	0.41 (0.033)	0.67 (0.026)	0.69 (0.027)	0.59 (0.030)

exactly under the condition of consistent spatial element arrangement for each set throughout all experimental phases, and a 24-hr delay. Words for each triplet or pair were presented in columnar form during all phases, and spatial order of words within the column was fixed for each set across all phases rather than varying randomly. Cross-experiment analyses involving that experiment and the 24-hr groups of Experiment 1 yielded no statistically significant differences. Those findings suggest that learning in the current paradigm occurs at a level that does not encode spatial word position, or

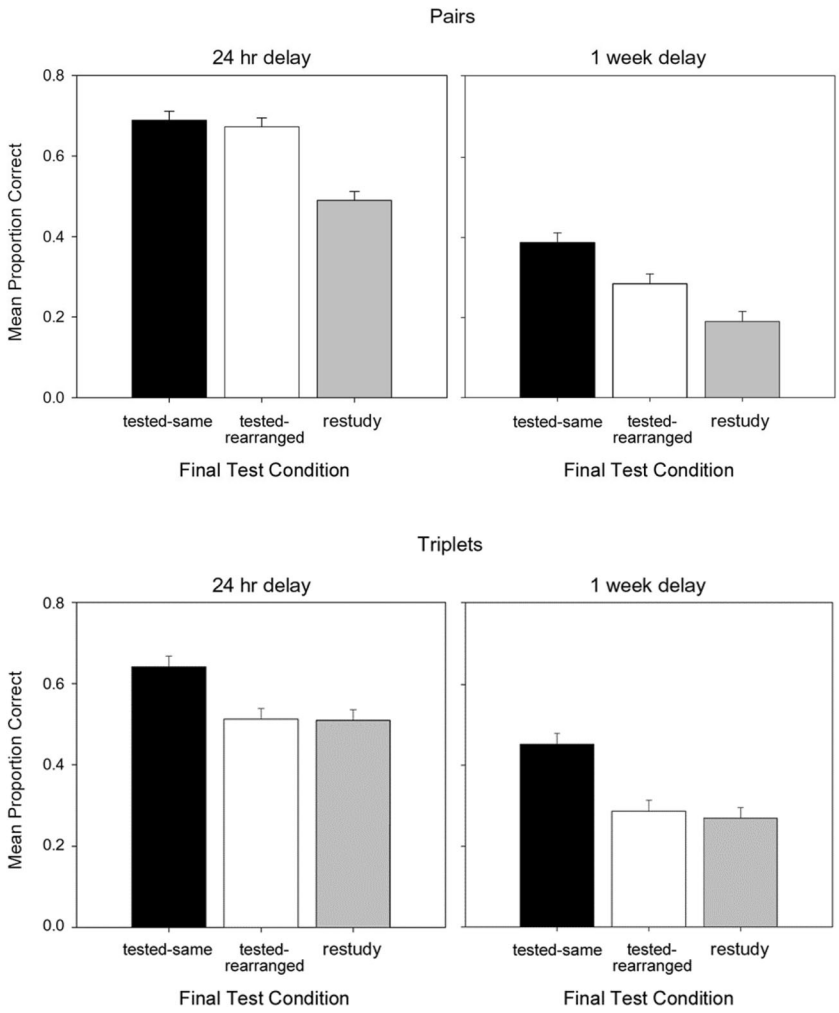

Fig. 3 Mean proportion correct in the three final test conditions for the pairs and triplets groups and the 24-hr and 1-week delay groups in Experiment 1. Standard error bars were calculated separately for each group based on within-subject ANOVA error term for the final test condition factor (Loftus & Masson, 1994). The error bars thus show the expected standard error of the relative mean values across test conditions

at least that any memory for spatial position has negligible impact on final test performance.

Specificity of learning for triplets as a function of training test accuracy

It is natural to ask whether the learning specificity for triplets that was observed in Experiment 1 and in prior experiments holds both for sets that were correctly answered and incorrectly answered on the training test. One approach that avoids selection bias (for discussion, see Kornell, Hays, & Bjork, 2009) is to compare the degree of learning specificity over a set of experiments that varied in the observed mean proportion correct on the training test. In their review of transfer of test-enhanced learning, Pan and Rickard (2018) tabulated the necessary data for 17 such experiments (see their Table 1) that involved facts and triplet materials, and in which the tested-rearranged condition was structurally equivalent to that of Experiment 1. Among those experiments (encompassing 775 subjects), the grand mean proportion correct difference score (tested-rearranged minus restudy proportion correct) was virtually zero (.003). Most importantly, there was no trend toward a systematic change in that difference score over the substantial range in experimental mean training test proportion correct (0.22 to 0.93; see Fig. 4). Thus, in the context of a large data set, there is no evidence for different degrees of specificity of learning as a function of training test accuracy, in turn suggesting that the same mechanism underlies the specificity of test-based learning on both correct and incorrect (with feedback) training test trials. The alternative possibility, that specificity of learning relative to the restudy control condition occurs only on correct training test trials, is inconsistent with this analysis. If that alternative possibility were correct, then the difference scores on the left side of Fig. 4 (wherein most training test trials were incorrect trials) should be greater than zero (indicating positive transfer relative to restudy), and the difference scores should approach zero only on the right side of the graph, wherein most training test trials were incorrect trials.

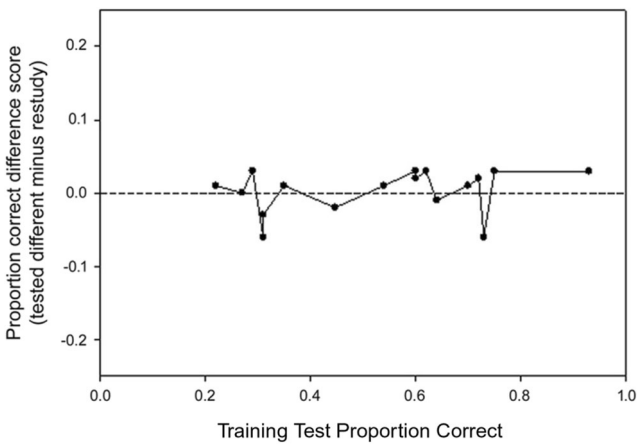

Fig. 4 Learning specificity that results from retrieval practice as a function of training test performance. Results from 17 experiments in the literature (each point represents a different experiment)

Theoretical implications thus far

The contrasting transfer results for pairs and triplets in Experiment 1 and in the Appendix appear to falsify any testing effect theory in which testing modifies and enhances memory in the same manner as does restudy, but only to a greater extent. Rather, it appears that some unique property of learning through testing allows for positive transfer relative to restudy in some circumstances, but precludes it in others.

Broadly speaking, at least two hypotheses for triplets seem viable. One possibility is that there is categorically no transfer to CR rearranged triplets relative to a restudy control under any transfer circumstances. There may be a fundamental property of memory for triplets and other multielement materials that precludes such transfer. A second hypothesis is that the learning specificity that we have observed for triplets is not a global property of memory for such materials, but rather is a property of the particular CR rearrangement that has been explored to date. In particular, it may be that, regardless of material type, transfer occurs only when the stimulus–response roles of words are exactly reversed on the final test. As Experiment 1 shows, that hypothesis holds for pairs. There was no exact reversal condition for triplets in Experiment 1, nor in prior experiments, but it is included as one of the two CR rearranged conditions in Experiment 2.

Experiment 2

Experiments 2 and 3 explored transfer effects for triplets only. For Experiment 2, the design and procedure for both the initial study and training phases were identical to those for the triplet groups of Experiment 1. The final test stimuli, however, differed in one important respect: In all conditions, only one cue word was presented, and two other words were to be retrieved. The three final test conditions included two CR rearranged conditions (*tested-reverse* and *tested-inverse*) and the *restudy* condition (see Fig. 1). The tested-reverse condition constitutes an exact CR reversal, analogous to that for paired associates (e.g., *gift, wine, ?* on the training test and *?, rose, ?* on the final test). For tested-inverse sets, the presented cue on the final test was one of the cues that was presented on the training test (henceforth, a *prior cue*), randomly selected, and the correct responses were the other prior cue and the prior correct response (e.g., *gift, wine, ?* on the training test and *gift, ?, ?* on the final test).

There was no tested-same final test condition in either this experiment or Experiment 3. In Experiment 2, a tested-same condition would require that two cues be presented for each triplet on the final test, as opposed to only one cue in the other

three conditions. Given that two cues should (all else held constant) result in higher proportion correct than would one cue, it would not have been possible to make strong inferences about performance in a tested-same condition relative to the other three conditions.

If the absence of transfer relative to restudy in Experiment 1 is a fundamental property of triplet materials, then there should be no performance differences among the three final test conditions of Experiment 2 (i.e., no transfer of training test learning to either the tested-reverse or tested-inverse conditions, relative to the restudy condition). Alternatively, if positive transfer to CR rearranged sets occurs only for exact CR reversals, then final test performance should be better in the tested-reverse condition than in either the restudy or the tested-inverse condition, and, by analogy to prior triplet results, performance in those latter two conditions may be equivalent.

Finally, having fully described the three final test conditions, a third transfer hypothesis is evident: transfer relative to restudy may be strongest in the tested-inverse condition. Uniquely for that condition, the final test cue was a prior cue, and one of the correct responses on the final test was the prior response (see Fig. 1). If retrieval of the prior response given a prior cue is easier than retrieval of the prior cue given the prior response, then more transfer may be observed in the tested-inverse condition—at least for retrieval of the prior response—than in the tested-reverse condition.

Method

Subjects Forty-one undergraduate students participated for course credit, and all completed both sessions. One subject's data were excluded because of computer error, leaving 40 subjects for data analysis.

Design, materials, and procedure The design for the study and training phases was identical to that for triplets in the 24-hr condition of Experiment 1. Only the final test differed. It entailed one block of 36 trials. On each final test trial, one word was displayed, while two were absent, and each replaced by a *"???";* subjects were instructed to type the two missing words, in any order, and to press Enter after typing each word. No further editing was permitted after the Enter button had been pressed. Of the triplets assessed on the final test, one third (12) had been restudied during training (restudy condition), while two thirds (24) had been tested during training. Of the previously tested triplets, half (12) featured two missing words, which were the two cue words during training (tested-reverse condition). For the other half (12) of previously tested triplets, one of the two missing words was a stimulus during training (tested-inverse condition). Note that the number of sets assigned to be trained using testing with feedback was increased from the prior experiments, from 18 to 24. This change equated the number of sets in the three final test conditions (12), potentially increasing sensitivity to effect differences between the tested-reverse and tested-inverse conditions.

Results

Training phase Mean proportion correct on the training test was 0.65 (SE = 0.035).

Final test phase Results, wherein a trial was scored as correct only if both responses were typed correctly, are depicted in Fig. 5 and listed in Table 1. That dependent measure was also used in the ANOVAs described below. The same relative proportions correct across conditions were also observed (albeit with better overall performance) using a more lenient accuracy criterion in which only one response had to be correct.

A one-way ANOVA on mean proportion correct, with a factor of final test condition (restudy vs. tested-reversed vs. tested-inverse; within subjects) yielded a significant main effect, $F(2, 78) = 14.22$, $p < .0001$, $\eta_p^2 = 0.27$. To further explore that result, each of the three possible pairwise comparisons was performed. The first comparison, between the tested-reverse and tested-inverse conditions, was statistically significant, $t(39) = 4.82$, $p < .0001$, $d = .76$, as was the second comparison, between the tested-inverse and restudy conditions, $t(39) = 4.65$, $p < .0001$, $d = .73$. The third comparison, between tested-reverse and restudy conditions, did not approach significance $t(39) = .60$, $p = .56$, $d = .09$.

In the tested-inverse condition there were two types of required responses: the prior response and a prior cue (see Fig. 1). Proportion correct was significantly higher for the prior response (0.58) than for the prior cue (0.44), $t(39) = 6.17$, $p <$

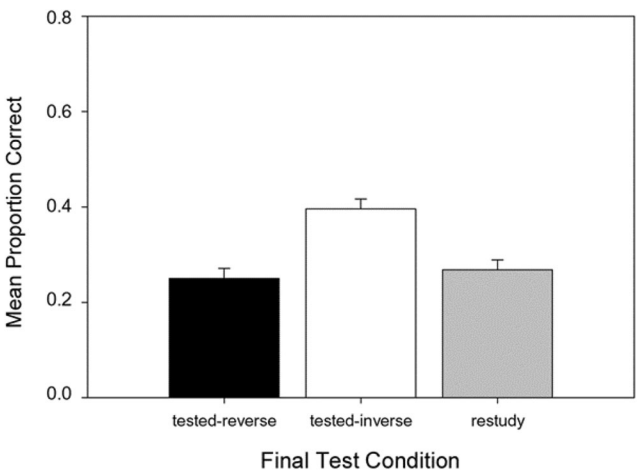

Fig. 5 Mean proportion correct for the tested-reverse, tested-inverse, and restudy final test conditions of Experiment 2. Error bars are standard errors based on the error term of a within-subjects ANOVA on final test proportion correct data (Loftus & Masson, 1994)

.0001, $d = 0.98$. However, proportion correct for the prior cue was only slightly and nonsignificantly higher than for either of the two required responses selected randomly in either the restudy or tested-reverse condition: In both of those conditions, the average proportion correct for one randomly selected response was 0.40. Thus, it appears that the higher overall proportion correct in the tested-inverse condition is driven primarily, if not exclusively, by a higher rate of retrieval for the prior response.

Discussion

Two candidate transfer hypotheses for triplets appear to have been falsified by this experiment. The hypothesis that no transfer of learning relative to restudy occurs for CR rearranged triplets under any circumstances is inconsistent with the robust main effect of the ANOVA. The hypothesis that transfer for triplets occurs only for exact CR reversals is not supported based on the nonsignificant pairwise effect involving the tested-reverse and restudy conditions.

The third hypothesis, that of selective transfer to the tested-inverse condition, was strongly supported by the restudy versus tested-inverse contrast. However, that "transfer" was observed only for the prior response given the prior cue, and thus could reasonably be framed as a testing effect rather than a transfer effect. The lack of transfer for retrieval of prior cues in this experiment (i.e., both required responses in the tested-reverse condition and the prior cue response in the tested-inverse condition) is analogous to the lack of transfer to the prior cue for triplets in Experiment 1 and in the Appendix. Thus, a clear pattern has been observed so far for triplets: there is no transfer of test-enhanced learning (relative to restudy) to CR rearranged sets when a required response is a prior cue.

The combined results from Experiments 1 and 2 indicate that, although bidirectional associations form for pairs in the retrieval practice paradigm, supporting transfer, under at least some circumstances they apparently do not form for triplets. Perhaps the most compelling evidence for that conclusion can be seen in the tested-reversed condition of Experiment 2. If word-to-word bidirectional associations were an important part of triplet learning on the training test, then positive transfer relative to restudy would have been expected in that condition. To the contrary, no such transfer was observed, despite that condition being the direct analog of the tested-rearranged (i.e., tested-reverse) condition for pairs in Experiment 1. For potentially related evidence that associative symmetry does not always hold for triplets in a serial memory paradigm, see Caplan, Glaholt, and McIntosh (2006) and Kahana and Caplan (2002); although that work yields new insights into associative processes, the multiple differences in experiment design and theoretical emphasis in their experiments versus our current experiments make a more detailed comparison beyond the scope of the current article.

A joint conditions hypothesis of transfer to stimulus–response rearranged triplets and pairs

The results of the experiments described thus far led us to a descriptive hypothesis—the *joint conditions hypothesis*—that accommodates the full pattern of results, including the contrasting results for pairs and triplets. According to that hypothesis, transfer relative to restudy will occur unless each of two conditions hold: (1) two (or more) cues are presented on the training test, *and* (2) the response on the final test is a prior cue. For a summary of the results and predictions of the joint conditions hypothesis, see Table 2.

The joint conditions hypothesis makes a novel prediction: positive transfer for triplets will be observed for the reverse of the CR rearrangement in Experiment 2, wherein *only one cue* is presented on the training test and *two cues* are presented on the final test (see Fig. 1). The presence of only one cue on the training test violates Condition 1 of the joint conditions hypothesis, and thus positive transfer is predicted by that hypothesis in both CR rearranged conditions. That prediction was tested in Experiment 3.

Experiment 3

This experiment was nearly identical to Experiment 2, the primary exception being that the cue–response configurations in the training and final test phases were reversed: on the training test, a single word cue was presented (and hence Condition 1 of the joint conditions hypothesis does not hold), with the other two words to be retrieved. In all conditions of the final test, however, two words were presented, with the third word (the prior cue) to be retrieved (Fig. 1). The final test again involved three conditions: *tested-reverse*, *tested-inverse*, and *restudy*. In the tested-reverse condition, the two cues on the final test were the two prior responses, and the required response was the prior cue. In the tested-inverse condition, the two cues on the final test included one of the prior cues and the prior response, and the required response was the other prior cue. If the joint conditions hypothesis is correct, then positive transfer relative to restudy should be observed in both the tested-reverse and tested-inverse conditions.

Method

Subjects Fifty-four undergraduate students participated for course credit. Analyses are limited to the 50 subjects that returned to complete both sessions.

Design, materials, and procedure The design was based on that of Experiment 2, with modifications as follows. Training test trials entailed the display of one word, with two absent and each replaced by empty text boxes; subjects were instructed to type the

Table 2 Experiments 1–3 final test results versus predictions of the joint conditions hypothesis and the dual memory theory of the testing effect

Experiment	Stimulus type	Final test condition	Results: observed testing effect or transfer relative to restudy	Predictions: joint conditions hypothesis and dual memory theory plus inclusive-OR account
1 / Appendix	Pairs	Tested-same	Yes	Yes
		Tested-rearranged	Yes	Yes
1 / Appendix	Triplets	Tested-same	Yes	Yes
		Tested-rearranged	No	No
2	Triplets	Tested-reverse	No	No
		Tested-inverse	Yes (prior response only)	Yes (prior response only)
3	Triplets	Tested-reverse	Yes	Yes
		Tested-inverse	Yes	Yes

two missing words in either order, and to press Enter after typing the first word. The spatial arrangement of the single word and two text boxes was columnar. After 12 s had elapsed, no new input was accepted, and the two correct words appeared below the text boxes (and side by side in relation to one another) for 2 s, which constituted feedback. Restudy trials also lasted for 12 s; this training trial duration (4 s longer than in the previous experiments) was designed to allow subjects sufficient time to type two answers on training test trials. On the final test, two stimulus elements were present on every trial, with the third missing element to be retrieved; one third of the triplets (12) had been restudied during training (restudy condition), while two thirds of the triplets (24) had been previously tested.

Results and discussion

Training phase Mean proportion correct on the training test, defined as trials on which both responses were correct, was 0.41 (SE = 0.033).

Final test phase Results are depicted in Fig. 6 and listed in Table 1. A within-subjects ANOVA with a single factor of final test condition yielded a robust main effect, $F(2, 98) = 8.37$, $p = .0004$, $\eta_p^2 = 0.14$. Two pairwise comparisons were performed to test whether the joint conditions prediction of positive transfer relative to restudy holds for both the tested-inverse and tested-reverse conditions. The first comparison, between the restudy and tested-reverse conditions, was statistically significant, $t(49) = 2.90$, $p = .006$, $d = .41$, as was the second comparison, between the restudy and tested-inverse conditions, $t(49) = 3.91$, $p = .0003$, $d = .55$.

General discussion

In four experiments, we investigated transfer of test-enhanced learning for the case of CR rearrangement between training and the final test. Experiment 1 confirmed strong positive transfer for pairs relative to restudy, but no such transfer for triplets when two of the three words of a set were presented as cues on both the training and final tests. Those results held for both varied spatial arrangement of stimulus elements for each set over experimental phases (Experiment 1) and consistent spatial arrangement (the experiment described in the Appendix).

In Experiments 2 and 3, transfer for triplets across previously untested CR rearrangements was explored. In Experiment 2, two words were presented as cues on the training test, just as in the prior triplet experiments, whereas on the final test one word was presented as the cue and the other two words were to be retrieved. Positive transfer relative to restudy was not observed, with the exception of retrieval of the prior response from a prior cue in the tested-inverse condition. In Experiment 3, one word was presented as a cue on the training test and two words were presented as cues on the final test. Positive transfer was observed both when the two cue words on the final test were both prior responses (tested-reverse) and when one of the two cue words on the final test was a prior

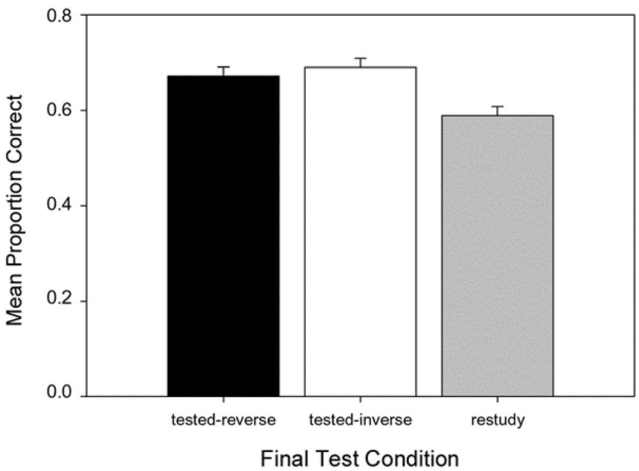

Fig. 6 Mean proportion correct for the tested-reverse, tested-inverse, and restudy final test conditions of Experiment 3. Error bars are standard errors based the error term of a within-subjects ANOVA on final test proportion correct data (Loftus & Masson, 1994)

correct response and the other cue word was the prior cue (tested-inverse).

The current results allow us to reject two transfer hypotheses that seemed plausible a priori: (a) that transfer categorically does not occur for SR rearranged triplets and (b) that transfer only occurs when the SR elements are exactly reversed on the final test. In contrast, the joint conditions hypothesis successfully describes the major transfer results that have been investigated to date for both pairs and triplets using the testing effect paradigm, and it predicted the results of Experiment 3. The joint conditions hypothesis is not intended as a process model, however, and we thus turn to other sources for a candidate account of the psychological basis of the study-wide pattern of results.

The dual memory model plus inclusive-OR gate

Here, we introduce a candidate process account that is consistent with the joint conditions hypothesis, and which may also explain the highly reliable phenomenon of equivalent performance in the restudy and tested-rearranged conditions when the joint conditions hypothesis holds. The account draws on a recently proposed quantitative model of the testing effect, the dual memory model (Rickard & Pan, 2018). The dual memory model assumes that restudy trials strengthen the episodic *study memory* that was formed during the study phase. Testing with feedback both strengthens study memory and encodes a new and separate *test memory*, which can be understood as a new episodic memory of the testing event. Study memory strengthening occurs on a training test trial because either (a) on correct trials study memory must be accessed to retrieve the correct answer (provided that there is no prior knowledge that could support accurate performance, as in the current experiments), and that access strengthens study memory just as restudy does, or (b) on incorrect test trials, study memory may be accessible after correct answer feedback is provided, and is hence strengthened, even if it could not be accessed when only the test cue(s) were available (and that accessibility should be similar to that on restudy trials since all elements are available for study memory retrieval once feedback has been provided). The testing effect is observed because there are two routes to retrieval for tested sets (through study and test memory), but only one route to retrieval for restudied sets (through study memory), and because study memory strength is assumed to be roughly equivalent for restudied and tested sets. Thus, provided that retrieval probability through study and test memory on a final test trial is at least partially independent, final test performance is predicted to be better in the test than in the restudy condition.

A simplest case, parameter-free quantitative model derived from this theory has provided, to date, the only quantitative account for multiple core phenomena in the testing effect literature for cued recall, including testing effect magnitude as a function of proportion correct in the restudy condition, the effect of correct answer feedback, and the testing effect retention function for the cases of both feedback and no feedback, among others (Rickard & Pan, 2018).

The dual memory theory was developed to explain the testing effect for cued recall. It makes only minimal assumptions, however, about basic properties of study and test memory. Study memory is assumed to be accessible—on a probabilistic basis—for any to-be-retrieved response on a test trial, including the case in which the studied set involves three (or more) elements. That is, associations between words in study memory are assumed to be strongly bidirectional, although not necessarily symmetric for triplets (for pairs, there is evidence that associations after study are symmetric; Kahana, 2002). That assumption is in principle consistent with work indicating that cued recall from episodic memory occurs through a pattern completion process (Horner, Bisby, Bush, Lin, & Burgess, 2015; Horner & Burgess, 2013, 2014).

Test memory is defined in terms of a stimulus-to-response mapping. When a cue is presented (or cues) on the training test under a task goal to retrieve the associated response, a new episodic memory is assumed to be created (cue memory). When the response is retrieved through study memory, or is provided through feedback, an association forms between cue memory and the response. Cue memory plus the association to the response constitutes *test memory*.

Here, we propose an extension of the dual memory theory to accommodate the current transfer results, still with no free parameters. We propose that, in the default case, test memory for both pairs and triplets can be understood as involving bidirectional associative links between cues and the response (analogous to the structure of study memory), links that can support positive transfer. The critical exception, however, is the case in which the joint conditions hold. In that case, we hypothesize that the test memory is fully *asymmetrical*, having a feedforward structure from the cues to a joint cue representation, and from that joint cue representation to the response, but no associations in the reverse direction. In the implementation described here, the joint cue representation exists independently of the cue representations, and it serves, in effect, as an inclusive-OR gate (see Fig. 7). We also assume for simplicity that cue-to-cue associations do not form on the training test, or that if they do form, they have negligible influence on final test performance. That assumption is consistent with the results in the tested-inverse condition of Experiment 2, wherein proportion correct for retrieval of one prior cue when the other prior cue was presented on the final, was statistically equivalent proportion correct for a single randomly selected response in the restudy condition. Hence, in both cases, retrieval appears (in our modeling framework) to occur only through study memory, and there is no evidence of cue-to-cue learning on the training trial. The assumption is also consistent with the finding of Starns and Hicks (2005,

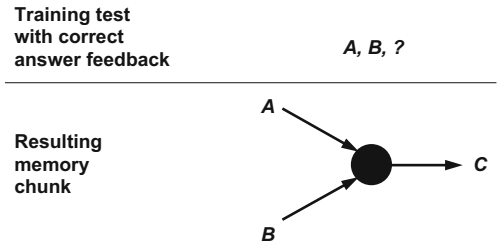

Fig. 7 Graphical depiction of the asymmetric associative representation that we propose forms on the training test when two or more cues are presented. The black circle represents the gating of activation from the presented cue(s) to the response. On the final test, that representation can only support retrieval when at least a subset of the originally presented cues are presented (A or B, or both) and when the correct response is the same as the correct response on the training test (C). Further, in this proposal, presentation of one of the cues on the final test (e.g., A) does not support retrieval of the other cue (e.g., B). Hence, training test learning given two or more retrieval cues effectively yields an inclusive-OR gate for later response retrieval

2008; see also Meiser & Bröder, 2002) that incidental context features of a set (e.g., location, color) are not associatively linked during study.

Accounting for the transfer results of Experiments 1–3 and the Appendix

The dual memory plus inclusive-OR account is consistent with the joint conditions hypothesis and it accommodates all of the experimental effects at the ordinal level for pairs and triplets, as well as the performance equivalence in the restudy and tested-rearranged conditions when the joint conditions hold.

For pairs in Experiment 1 and the Appendix, and for triplets in Experiment 3, the inclusive-OR gate does not form on the training test, because in those cases there is a single cue on the training test. The observed positive transfer is thus predicted by our modeling framework because (a) test memory can contribute to final test performance in those transfer conditions, and (b) the dual memory model predicts that positive transfer relative to restudy will occur when both study and test memory can contribute to performance.

For triplets in Experiment 1, the inclusive-OR test memory is formed on the training test (i.e., there are two training test cues). Because in the *tested-rearranged* final test condition the correct response is a prior cue, the inclusive-OR test gate blocks retrieval of the correct response. The same reasoning holds for the *tested-reverse* condition of Experiment 2. Hence, in the proposed model, retrieval in the tested-rearranged and tested-reverse conditions of those experiments must occur only through study memory, just as occurs in the restudy condition. As a result, performance in those restudy and tested-transfer conditions is predicted by the model to be equivalent, as was observed to a close approximation.

In Experiment 2, positive transfer relative to restudy *was observed* in the tested-inverse condition for the prior response only, a finding that can also be explained by our model. In that condition, the presented cue on the final test was a prior cue (see Fig. 1), and one of the responses on the final test was a prior response. Thus, the presented cue in the tested-inverse condition can access the inclusive-OR representation that was formed on the training test, supporting retrieval of the prior response through test memory, and yielding in our model the observed positive transfer relative to restudy. However, the same inclusive-OR representation blocks retrieval of the prior cue form test memory in that condition. Retrieval of the prior cue can occur only through study memory, again yielding the observed near equivalent proportion correct for (1) the prior cue in the tested-inverse condition and (2) either of the required responses in the restudy condition.

The interpretation above for the tested-inverse condition in Experiment 2 implicitly assumes that there is no resampling of memory after successful retrieval of the prior response. Such resampling is not specified for the dual memory model (Rickard & Pan, 2018), but it would be plausible in that case. After the prior response is retrieved, there are effectively two cues available for resampling of study memory, as opposed to just the one cue that was presented for retrieval at the outset of the trial. Since the number of cues is doubled, one would expect that the probability of retrieval of the single remaining element from study memory would be greater than on the first retrieval attempt, wherein only the presented cue was available. That dynamic would yield higher proportion correct for retrieval of the prior cue in the tested-inverse condition than for a given response in the restudy condition. However, there was only a small and nonsignificant trend in that direction (0.44 vs. 0.40). That results suggests that there was minimal or no resampling. One possibility is that resampling in such contexts is strategic and that subjects chose not to engage in it.

Analyses conditionalized on training test accuracy

Although the focus of this paper is empirical and theoretical development based on randomized experimental evidence, we also conducted a supplementary, and in essence correlational, analyses of the Experiment 1 data in which final test proportion correct was calculated separately for sets that were answered correctly versus incorrectly on the training test. This analysis may yield new insight into whether transfer of test-enhanced learning is dependent on training test accuracy in the context of correct answer feedback after each trial. The earlier analysis in which the transfer effect was plotted against training test proportion correct (see Fig. 4) already suggests that the specificity of learning for triplets (and facts) holds both when mean training test accuracy is low and when it is high, implying that the learning specificity holds on both correct and

incorrect test trials with feedback. However, that analysis is also correlational.

For the conditional analysis described here, it is important to keep in mind that inference may be limited by selection bias. Incorrectly answered sets on the training test are almost certainly more difficult to learn, on average, than are correctly answered sets, and various factors that introduce noise (lapses of attention; typing errors; interference from the preceding set) are far more likely to yield an incorrect response for a set that has a high memory strength after the study phase than to yield a correct response for a set that has a low memory strength after study. Further, matched comparisons between conditionalized test and restudy performance are not possible this analysis, because such comparisons would require foreknowledge of which sets in the restudy condition would have been answered correctly versus incorrectly in the training phase, had they been tested.

For inclusion in this analysis, each subject in each of the four experimental groups had to have final test observations in each of the four cells generated by crossing the training test accuracy factor (correct or incorrect) with the final test condition (tested vs. transfer). That criterion resulted in elimination of 21 of the 119 subjects (although unbalanced analyses of the full set of 119 subjects yielded very similar results).

Results are shown in Fig. 8, with error bars representing standard errors of the mean. Overall restudy performance is indicated for each of the four groups. For triplets in both the 24-hr and 1-week delay groups (upper panel), the pattern appears to be straightforwardly interpretable: for training test correct (TT-C) sets—but not for training test incorrect (TT-I) sets—learning specificity was observed. That is, for TT-C sets only, proportion correct was much higher in the tested-same condition than in the tested-rearranged condition. For pairs, the TT-C sets show a smaller proportion correct decrease from tested-same case to tested-rearranged case, in-line with the overall proportion correct findings, and with the conclusion that associations for pairs following a training test are bidirectional. Pairs results for TT-I sets are similar to those for triplets, exhibiting a trend toward an *increased* proportion correct in the tested-rearranged condition relative to the tested-same condition.

The conditionalized results for triplets are, on their face, at least, only partially consistent with the dual memory plus exclusive-OR account of the overall experimental findings. In line with that account, test-enhanced learning for triplets appears to be highly specific to the trained CR arrangement. But in contrast to that account, that specificity of learning appears to occur primarily on correct training test trials. If correct, that interpretation would require modification of the proposed modeling framework. It would be premature to draw a strong conclusion along those lines, however, because the aforementioned selection confounds may be at play. Consider two example confounding factors. First, restudy sets that are

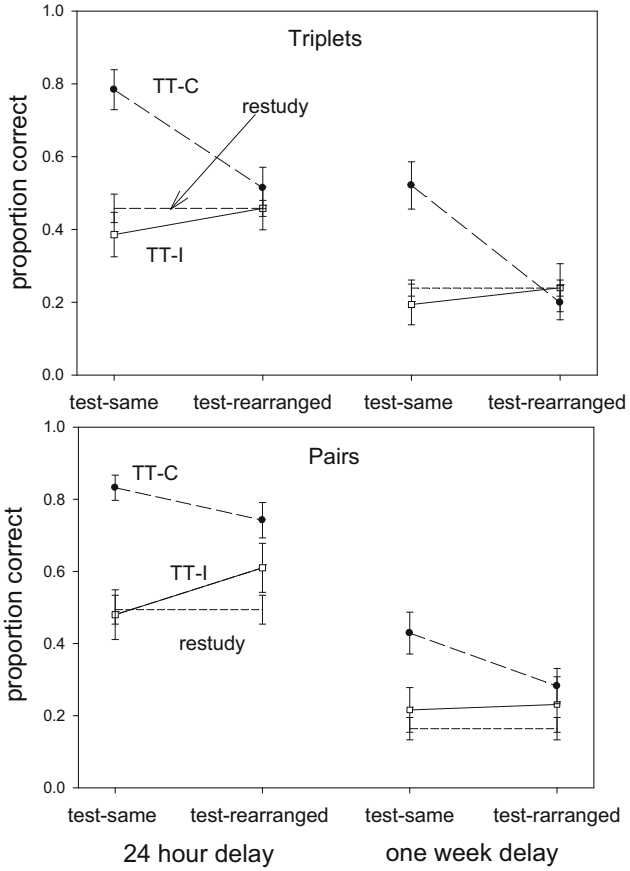

Fig. 8 Results of the conditional analysis of the Experiment 1 data. TT-C = training test correct; TT-I = training test incorrect

difficulty-matched to the TT-C sets (if they could be identified) would almost certainly show a higher mean proportion correct (if they had been tested) than would the restudy sets overall (because those sets should be easier to learn). Conversely, hypothetical restudy sets that are difficulty-matched to the TT-I sets would almost certainly have a lower mean proportion correct than would be the case for restudied items overall. Thus, for the tested-same case, it is possible that there *is* test-enhanced learning, relative to matched restudy, for both TT-I sets and TT-C sets (consider the tested-same data for triplets in Fig. 8).

A second confounding factor may explain the tested-rearranged data for triplets. Specifically, it seems likely that associative strengths after initial study vary not just from set to set, but also to some extent from word to word within each set, and in the forward versus reverse directions. If so, then part of the reason a set was answered correctly on the training test could be because the learned associations from the presented cue words to the response word was a bit stronger (on average over sets) than for other possible cue–response arrangements for those sets. Correspondingly, for incorrectly answered sets on the training test, associative strength between the cue words and the response words could have been a bit weaker than would be expected on average. In other words, training

test accuracy is likely to depend on both overall set-level associative strength and on variation in strength between the words within a set. If so, then there should be a reverse consequence for conditionalized performance when cues and responses are rearranged on the final test. Specifically, for TT-I sets, there may be some degree of boost in proportion correct in the tested-rearranged condition relative to the tested-same condition (because of the purely statistical expectation that the cue-to-response strengths on the final test would on average a bit higher for rearranged for TT-I sets than for the arrangement that was presented on the training test, and in the tested-same condition of the final test). As noted above, a trend toward that outcome is evident for both triplet and pair TT-I sets in both the 24-hr and 1-week groups; proportion correct is actually higher in the tested-rearranged than in the tested-same condition. Outside of the mechanism proposed here regarding random variation in associative strengths over words within a set, it is unclear to us why that pattern would be consistently observed. If the unknown (and untestable using the current data) magnitude of that bias effect is large enough, it could conceivably mask true results for TT-I sets (i.e., absent that bias effect, the performance could actually be better in the tested-same than in the tested rearranged condition). That scenario is more in line with the dual memory plus chunking account, and with the results of the cross-experiment training test proportion correct analysis that were discussed earlier (see Fig. 4).

For TT-C triplets and pairs, the reverse effect would be expected, suggesting that the reduction in proportion correct that was observed in the tested-rearranged compared with the tested-same conditions is greater than would be expected absent that source of bias. It should be noted that none of those complicating factors are at play in the primary results for each experiment, which were not conditionalized on training test accuracy.

Despite the substantial interpretational complications, the results of these conditional analyses provide useful information in our view, and may indeed herald differences in learning specificity on correct versus incorrect training test trials. In addition, some types of bias effects in this type of analysis may be of psychological interest in themselves. A complete model of testing effect and transfer phenomena should ideally integrate various types of conditionalized and nonconditionalized results. It is evident, however, that achievement of that goal will require a more complex modeling approach than currently exists.

Conclusions

In addition to revealing diverse patterns of retrieval practice and transfer for pairs and triplets, the current work allows us to rule out the possibility that the previously observed high specificity of cued recall-based learning for triplets is a global property. It also allows for confident rejection of the possibility that such transfer only occurs for the case of pure SR reversals (to the contrary, there is no transfer relative to restudy in that case). Rather, the results of Experiment 2 led to the proposal of a straightforward rule (the joint conditions hypothesis), which describes the conditions under which transfer does and does not occur across experiments, and which successfully predicted the results of Experiment 3. Those transfer findings raise the possibility of a new principle of retrieval-based associative learning—namely, asymmetric learning when two or more cues are presented—that when applicable may supersede the pairwise bidirectional association principle that appears to adequately describe associative learning in other contexts.

We also showed that a psychological process theory, the dual memory theory of the testing effect combined with the hypothesized inclusive-OR representation, can explain at least the nonconditionalized transfer pattern for both pairs and triplets. The repeated finding of performance equivalence between the restudy condition and the various tested-transfer conditions when the inclusive-OR representation is expected to form, further supports the dual memory theory, without which that performance equivalence would have no obvious mechanistic basis.

The data and materials for all experiments are available through the Open Science Framework and can be accessed at the following URL: https://osf.io/95b6r/

Acknowledgements The authors thank John Barry, Maxim Deinitchenko, Kayla Hartman, Yangyang Liu, Jonathan Mejia, and Thomas Ting for assistance with data collection. Thanks also to Dina Rodgers for expert assistance with subject pool management.

Appendix

Replication of Experiment 1 under conditions of consistent spatial element arrangement

In a supplementary experiment, we explored whether the results of Experiment 1 hold under conditions of consistent spatial arrangement of the words for each set across all phases of the experiment. This experiment, which was designed to address potential comparisons between the retrieval practice paradigm employed in the present experiments versus those used in other literatures (e.g., work on associative symmetry), replicated the design of Experiment 1 using a 24-hr delay (the 1-week delay was dropped), with the exception that the spatial position of the stimulus elements for each pair and triplet was no

longer randomized across the three experimental phases. Specifically, the stimulus words for both pairs and triplets were presented in columnar form throughout all phases of the experiment (whereas in Experiment 1 such a format was used for only the first two phases, as illustrated in Fig. 1), and each word for each set always filled the same columnar position (for both pairs and triplets, the missing word during both training and the final test in the tested-same condition was always the bottom-most word of the column; the missing word in the tested-rearranged condition of the final test was always the highest word of the column; in the restudy condition, half of the missing words were in the lowest position, and half were in the highest position). This provided an implicit cue to subjects that columnar word order was held constant throughout all phases. Thus, in this experiment, memory for relative word location can in principle be a driver of final test performance, including the magnitude of the testing and transfer effects.

Seventy-four undergraduate students participated for course credit. Six subjects were dropped due to not returning for Session 2 or computer errors; analysis was limited to the 68 subjects (33 in the pairs condition, and 35 in the triplets condition) that completed both sessions.

Results and discussion

In the training phase, mean proportion correct on the training test was 0.60 (SE = 0.025) and 0.63 (SE = 0.041), in the pairs and triplets conditions, respectively. Those mean differences were not statistically significant, $t(103) = 0.9$, $p = .37$, $d = 0.088$.

Final test results are depicted in Fig. 9. A factorial ANOVA with the factors stimulus type (between subjects), and final test condition (tested identical vs. tested rearranged vs. restudy; within subjects) indicated no significant main effect of stimulus type, $F(1, 66) = 1.79$, $p = .19$, $\eta_p^2 = 0.026$, replicating Experiment 1. Also as observed in Experiment 1, there was a significant main effect of final test condition, $F(2, 132) = 32.4$, $p < .0001$, $\eta_p^2 = 0.329$. Of most interest is the Stimulus Type × Final Test Condition interaction, $F(2, 132) = 3.73$, $p = 0.026$, $\eta_p^2 = 0.054$; as in Experiment 1, for triplets there was no trend toward positive transfer relative to restudy in the tested-rearranged condition, whereas for pairs there was substantial transfer.

A cross-experiment analysis of the 24-hr delay groups of Experiments 1 and this experiment was performed to more formally investigate the effects of consistent versus varied word location over experimental phases. In an ANOVA with the factors experiment (between subjects), stimulus type (between subjects), and final test condition (within subjects), there was again a significant interaction between final test condition and stimulus type, $F(2, 258) = 9.7$, $p < .0001$, $\eta_p^2 = 0.07$. There were, however, no main or interaction effects involving experiment (all ps > .07). Thus, there is no statistical evidence that consistency of spatial word order affected any aspect of final test performance. It appears that either the spatial position of words was weakly encoded during training or was not retained over the delay between training and the final test. In any case, that factor appears to play a minimal role in testing and transfer effects in this paradigm.

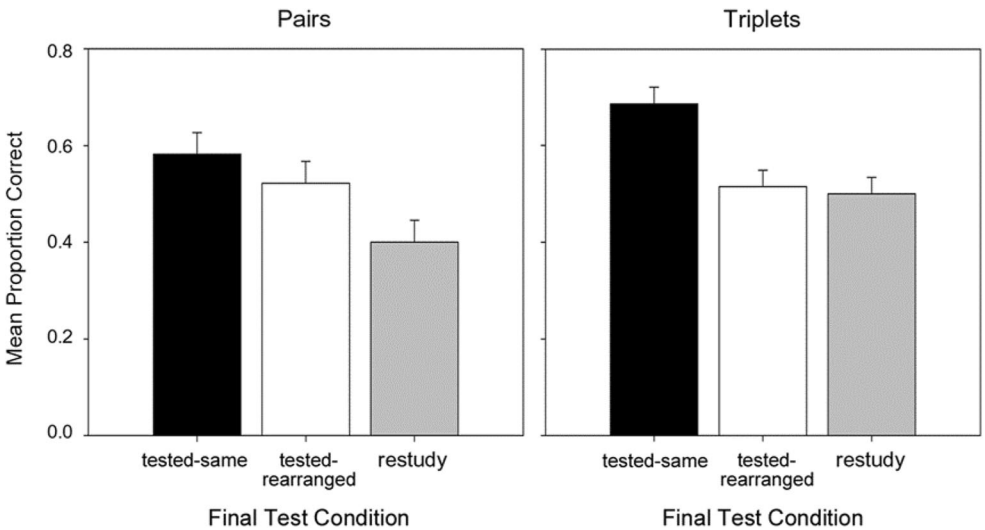

Fig. 9 Mean proportion correct for pairs and triplets in the tested-same, tested-rearranged, and restudy final test conditions of the supplemental experiment. Error bars are standard errors based on the error term of a within-subjects ANOVA conducted separately for pairs and triplets (Loftus & Masson, 1994)

References

Caplan, J. B., Glaholt, M. G., & McIntosh, A. R. (2006). Linking associative and serial list memory: Pairs versus triples. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(6), 1244–1265. https://doi.org/10.1037/0278-7393.32.6.1244

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. https://doi.org/10.3758/BF03194004

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory Vol. 53* (pp. 63–147). San Diego, CA: Academic Press. https://doi.org/10.1016/S0079-7421(10)53003-2

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. https://doi.org/10.1177/1529100612453266.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

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

Horner, A. J., & Burgess, N. (2013). The associative structure of memory for multi-element events. *Journal of Experimental Psychology: General*, *142*(4), 1370.

Horner, A. J., & Burgess, N. (2014). Pattern completion in multielement event engrams. *Current Biology*, *24*(9), 988–992.

Horner, A. J., Bisby, J. A., Bush, D., Lin, W. J., & Burgess, N. (2015). Evidence for holistic episodic recollection via hippocampal pattern completion. *Nature Communications*, *6*, 7462.

Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory & Cognition, 30*(6), 823–840. https://doi.org/10.3758/BF03195769

Kahana, M. J., & Caplan, J. B. (2002). Associative asymmetry in probed recall of serial lists. *Memory & Cognition, 30*(6), 841–849.

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. In B. H. Ross (Ed.), *The psychology of learning and motivation; the psychology of learning and motivation* (pp. 183–215). San Diego, CA: Elsevier Academic Press.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*(4), 476–490.

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. https://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. https://doi.org/10.1037/xap0000063

Meiser, T., & Bröder, A. (2002). Memory for multidimensional source information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(1), 116–137.

Pan, S. C., & Rickard, T. C. (2017). Does retrieval practice enhance learning and transfer for term-definition facts? *Journal of Experimental Psychology: Applied, 23*(3). https://doi.org/10.1037/xap0000124

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710–756. https://doi.org/10.1037/bul0000151

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

Pan, S. C., Wong, C. M., Potter, Z. E., Mejia, J., & Rickard, T. C. (2016). Does test-enhanced learning transfer for triple associates? *Memory & Cognition* https://doi.org/10.3758/s13421-015-0547-x

Pan, S. C., Hutter, S., D'Andrea, D., Unwalla, D., & Rickard, T. C. (2018). In search of transfer following cued recall practice: The case of process-based biology concepts. *Applied Cognitive Psychology*. https://doi.org/10.1002/acp.3506

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., et al. (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.

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. https://doi.org/10.1037/a0023956

Rickard, T. C., & Pan, S. C. (2018). A dual memory theory of the retrieval practice effect. *Psychonomic Bulletin & Review* https://doi.org/10.3758/s13423-017-1298-4

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012

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. https://doi.org/10.1037/a0037559

Starns, J. J., & Hicks, J. L. (2005). Source dimensions are retrieved independently in multidimensional monitoring tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1213.

Starns, J. J., & Hicks, J. L. (2008). Context attributes in memory are bound to item information, but not to one another. *Psychonomic Bulletin & Review*, *15*(2), 309–314.

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. https://doi.org/10.1016/j.jml.2014.04.004