5-2014

The Role of Transfer-Appropriate Processing in the Testing Effect

Mary T. Veltre
University at Albany, State University of New York

The Role of Transfer-Appropriate Processing in the Testing Effect

An Honors Thesis presented to the
Department of Psychology,
University at Albany, State University of New York
in partial fulfillment of the requirements
for graduation with Honors in Psychology
and
graduation from the Honors College.

Mary T. Veltre
Research Mentor: Kit Cho
Research Advisor: James H. Neely, Ph.D.
Second Reader: W. Trammell Neill, Ph.D.

May, 2014

Abstract

The testing effect is the finding that taking a review test enhances performance on a final test relative to merely restudying the material. I investigated the role of transfer-appropriate processing in the testing effect using semantic cues to evoke conceptual processing and orthographic cues to evoke data-driven processing. After an initial study phase, subjects either restudied the material or took a cued recall test consisting of half semantic cues and half orthographic cues. Two days later, all of the subjects returned for a final cued recall test. The final test consisted of the exact same cue given for that target in the review phase, or a new cue that matched or mismatched the type of cue used for that target in the review phase. A "far transfer" effect of testing was found, with testing enhancing memory relative to restudying even in conditions in which the review test cue and final test cue involved different processing evoked by the mismatching type of cues. Consistent with transfer-appropriate processing, performance was the best when the review test and final test cues were identical (for the semantic cues), and was better when the type of cues matched than when they mismatched (whether the final test cues were semantic or orthographic). These results suggest that the testing effect is greater to the degree that the type of retrieval processing involved in the final test overlaps with the type of processing done during review.

Acknowledgments

I would like to express my sincere gratitude to my advisor, Dr. James Neely for providing me with this valuable learning opportunity, for the continuous support of my research and writing of my thesis, and for his patience, knowledge, and guidance. I would also like to extend my greatest appreciation to Kit Cho for all of his invaluable assistance and insight. Their contributions have made this research project possible.

I would also like to thank my parents for their hard-work and sacrifices so that I could be where I am today and have this opportunity. Finally, I'd like to acknowledge my sister, Kathryn, and my boyfriend, Ben Rea, for their constant support and encouragement throughout this process.

Table of Contents

Taking a memory test not only assesses one's knowledge of information, but also results in greater long-term retention of information relative to merely restudying the material, a phenomenon known as the testing effect (TE). The TE is robust, having been repeatedly demonstrated across a wide variety of memory tasks and memory materials (Roediger & Karpicke, 2006; McDaniel, Anderson, Derbish, & Morrisette, 2007; Roediger, Putnam, & Smith, 2011; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). The standard experimental paradigm used to investigate the TE consists of a study phase, a review phase, and a final test. During the study phase, subjects typically study either a list of individually presented words (McDaniel & Masson, 1985), a list of word pairs presented one pair at a time (Carrier & Pashler, 1992; Carpenter, 2011), or prose passages (Roediger & Karpicke, 2006). In the review phase, subjects either restudy the information as in the study phase, or they are tested on the information with a cued recall (Carrier & Pashler, 1992), free recall (Roediger & Karpicke, 2006), or recognition test (Chan & McDermott, 2007). Some studies use a distraction task rather than restudy as a control condition, but this results in only the test group receiving an additional presentation of the materials, creating a bias toward obtaining a TE. The restudy control condition ensures that both the test and restudy groups are exposed to the materials an additional time in the review phase, so that any difference in performance must be due to some effect of testing vs. restudying and not merely due to additional exposure to some of the studied items during testing. In fact, when no feedback is provided in the review test and performance is less than 100%, as is almost always the case, the test group is re-exposed to fewer items than is the restudy group. Even when feedback is provided, the restudy presentations are often longer in duration than are feedback presentations for the test group. Consequently, using restudy rather than distraction as the control baseline for evaluating the TE creates a bias against obtaining a

TE. Thus, a TE observed relative to the more conservative restudy control baseline deserves more credence than a TE observed relative to a distraction baseline. After a delay of 5 minutes to 1 week, all subjects receive a final test on the studied items. In most TE studies, the review test and the final test are of the same type, but that is not always the case (Carpenter & DeLosh, 2006; McDaniel, Kowitz, & Dunay, 1989) and will not be the case in the present study. An abundance of studies using this experimental TE paradigm have clearly demonstrated that receiving a review test enhances performance on a delayed memory test relative to restudying the material in the review phase (e.g., Carpenter, Pashler, Wixted, & Vul, 2008; Roediger & Karpicke, 2006).

One possible explanation as to why testing enhances memory on delayed tests comes from the transfer-appropriate processing (TAP) account of memory. According to TAP, performance in a memory test will be good to the degree that the type of processing engaged during retrieval in the memory test matches the type of processing performed in encoding the to-be-remembered material during the study list. Morris, Bransford, & Franks (1977) showed evidence of TAP using semantic and rhyming tasks. They had subjects encode information using rhyme processing or semantic processing and then had subjects take either a semantic recognition test (respond “yes” if the test item is semantically related to a studied item) or a rhyme recognition test (respond “yes” if the test item rhymes with a studied item). They found that initial semantic processing was superior to rhyme processing for a semantic recognition test, but that initial rhyme processing was superior to semantic processing for a subsequent rhyme recognition test. This shows that a match between initial encoding and later retrieval influenced performance. TAP can be used to account for the TE. Specifically, the TE could be produced by TAP in that the episodic retrieval processing engaged by a memory test in the review phase would better match the retrieval processes involved in the final test than would the processes involved in restudying, especially if the review and final tests are identical.

TAP can also be applied to variations in the match between two different types of retrieval tasks given in the review test and the final test in the TE paradigm. A number of studies have shown that a positive TE occurs when the review test and final test are the same, but this leaves open the question of how different the review and final tests can be and still yield a TE. In other words, are the mnemonic benefits of testing confined to the specific conditions of the review test, or do the benefits transfer to new domains? Many studies have shown that testing improves performance relative to restudying on tasks that involve “near transfer” of information within the same domain (Carpenter, Pashler, & Vul, 2006; Jacoby, Wahlheim, & Coane, 2010; Rohrer, Taylor, & Sholar, 2010). Butler (2010) replicated prior findings that testing produces better “near transfer” of information than repeated studying of the material (Experiment 1b), and in addition found that testing also enhances “far transfer” for information to a new domain (Experiment 2). In Butler’s (2010) Experiment 2, the study phase consisted of six different 500 word prose passages on a variety of topics. During the review phase, subjects either restudied each passage one more time or took a test on each of the passages. On the tests, subjects were asked questions about the passage and were required to produce a response. After each question, the correct answer was given as feedback. After a 1-week retention interval, subjects took a final test that consisted of new inferential questions that included some mention of the relevant concept from the initial learning session, but which required the application of the studied concept to a different knowledge domain. Better “far transfer” was found in the tested group than in the group that restudied the passages. This important finding extends the mnemonic benefits of testing to situations in which knowledge must be transferred to a different domain, demonstrating that the TE represents more than the mere reproduction of previous test answers. Butler's (2010) research shows that far transfer can occur in reasoning tasks, but does not say whether this applies to episodic memory for specific words. The present experiment assesses transfer between the review and final tests with episodic memory for specific words, as a function of how similar the review and final tests are.

The role of TAP in the TE has been investigated in experiments reported by McDaniel, Kowitz, and Dunay (1989) and McDaniel and Masson (1985). However, these experiments yielded mixed results. In both of these investigations, the type of processing was manipulated in all three phases by using two types of cues (phonemic vs. semantic). Incidental learning procedures were used in the study phase to control processing during study, with subjects rating how well the cue and target matched phonetically or semantically. The control group received no review at all, and both experiments used a 1-day retention interval between the review and final tests. The only difference between the two experiments was that the type of cue was randomized within subjects in McDaniel and Masson (1985), whereas McDaniel et al. (1989) manipulated the type of cue between subjects. A TE was found relative to the no-review control group, but a reverse TE was found relative to McDaniel and Masson's Experiment 3 restudy group. Evidence for TAP effects were found across the study and review phases as well as the study and final test phases: recall was better when the cues in the two phases matched than when they mismatched. Across the review and final test phases, however, a reverse TAP effect occurred when a phonemic cue was given in both study and final test. Thus, the results were inconsistent and difficult to interpret for a few reasons: (1) There was no positive TE relative to a restudy review (in fact, a reverse TE was found relative to the restudy group in McDaniel and Masson's Experiment 3), even at a 1-day retention interval. (2) These experiments used incidental study procedures rather than the standard intentional study procedures in which subjects are told to study the material in any way they choose. (3) Review performance was low and differed when the cues matched vs. mismatched the study cues, complicating the interpretation of the transfer between the review and final tests.

Carpenter & DeLosh (2006) also investigated the role of TAP in the TE using all possible combinations of cued recall, free recall, and recognition tests in the review and final test, such that the review test and the final test matched for some conditions and mismatched for others. They found that performance in the final test was not enhanced by matching, as opposed to mismatching, review and final tests and that a TE only occurred when the review test was free recall and the final test was cued recall. Clearly, these results run counter to the TAP account of the TE, which predicts that final test performance would be the highest when the review and final test were of the same type. However, there are some limitations to Carpenter & DeLosh's (2006) study. They did not find a TE in all conditions in which the review and final test were the same, most likely due to the short 5-minute retention interval between the review phase and final tests. A longer retention interval would improve the chances of finding a TE, along with any evidence for TAP. Furthermore, recognition, cued recall, and free recall tests all require conceptual processing (Roediger, 1990), and the conceptual processes involved may not have differed enough to produce a TAP effect. Indeed, Finlay & Benjamin (2012) have argued that there is evidence that the conceptual processing demands of free recall and recognition tests are not different enough to invoke qualitatively different encoding strategies.

In the present study, I sought to improve upon prior research by using procedures that might provide a better chance of detecting a TAP effect in the TE if it exists. Specifically, unlike Carpenter and DeLosh (2006), I used a retention interval of 2-days to maximize the chances of obtaining a TE and used tests that evoke qualitatively different processes (semantic/conceptual processing vs. orthographic/data-driven processing, cf. Blaxton, 1989) but yield equivalent performance in the initial review test (as determined by pilot testing). Having equivalent performance in the review phase for the two types of tests, which McDaniel and Masson (1985) and McDaniel et al, (1989) did not observe, will ensure that any difference in final test performance as a function of the type of review test received is not due to the recall differences for these tests during the review phase, but instead is due to the relation of the cues received in the review and final tests.

In the present experiment, the study phase involved procedures that are most typically used in the TE paradigm and that are representative of how information is studied outside the laboratory. That is, subjects studied a list of words under intentional learning instructions and unlike in the experiments done by McDaniel and Masson (1985) and McDaniel et al. (1989) could use any type of processing they chose. In the study phase, the individually presented words were presented three times so as to produce moderately high levels of cued recall in the review test. In the review phase, half of the subjects (i.e., the restudy group) studied the words again, and the other half of the subjects (i.e., the test group) received a cued recall test on the words. In the review test, half of the words were cued semantically (ROAD – STREET) and the other half were cued orthographically (S_R_ET – STREET). On the final cued recall test 2 days later, some of the words were cued with the exact same cue as in the review test, some were cued with a new cue that matched the type of cue in the review test (semantic-semantic or orthographic-orthographic), and others were cued with a new cue that was a different type of cue than that given in the review test (semantic-orthographic or orthographic-semantic). This created six cuing conditions on the final test: Semantic/Identical, Semantic/Match, Semantic/Mismatch, Orthographic/Identical, Orthographic/Match, and Orthographic/Mismatch. The first label (semantic or orthographic) refers to the final test cue and the second label (identical, match, or mismatch) refers to how the final test cue for a studied item related to the review cue for that item. Table 1 provides examples of cues in each condition for each phase. Because subjects in the restudy group did not receive test cues in the review phase, the only distinction was between semantic and orthographic cues during the final test. With a 2-day retention interval, a large TE should occur in the identical cuing conditions. The issue is whether significant TEs will occur in the matching and mismatching cuing conditions. If a TE occurs in the mismatching cuing condition, this would be the first evidence I am aware of that a “far transfer” TE occurs in an episodic memory test for individual words rather than in “reasoning” tasks (Butler, 2010). Further, if the TE is moderated by TAP, the effects of the type of review cue and the type of final test cue should interact. Specifically, the TE should be greatest when the cues in the review test and the final test are identical, next greatest when the cues in the review test and the final test are different but match qualitatively, and smallest in the mismatching cue conditions.

Table 1.

Experimental Procedures

Conditions	Study Phase Each word studied 3x (5 sec/word)	Filler Task (5 mins)	Review Phase (self-paced tests, with 1500 ms feedback)	RI (2-day)	Final Test (self-paced, no feedback)
<u>Test Review</u>					
Semantic/Identical	ABOVE	35+21=?	BEYOND - ____		BEYOND - ____
Semantic/Match	ABOVE	35+21=?	BELOW - ____		BEYOND - ____
Semantic/Mismatch	ABOVE	35+21=?	A _ O V _ - ____		BEYOND - ____
Orthographic/Identical	ABOVE	35+21=?	A _ O V _ - ____		A _ O V _ - ____
Orthographic/Match	ABOVE	35+21=?	A B _ V _ - ____		A _ O V _ - ____
Orthographic/Mismatch	ABOVE	35+21=?	BELOW -____		A _ O V _ - ____
<u>Restudy Review</u>					
Semantic	ABOVE	35+21=?	ABOVE		BELOW - ____ **OR** BEYOND - ____
Orthographic	ABOVE	35+21=?	ABOVE		A _ O V _ - ____ **OR** A B _ V _ - ____

Note. RI = retention interval.

Method

Design

The design was a 2 (Type of Review: restudy or test) x 2 (Final Test Cue: semantic or orthographic) x 3 (Review Test Cue: identical, match or mismatch) mixed design, with Type of Review manipulated between subjects and type of Final Test Cue manipulated within subjects. The type of Review Test Cue was manipulated within subjects and nested under Type of Review. For the Test Review group, there were three Review Test Cue conditions: identical, match, or mismatch. Subjects in the Restudy Review group were not presented cues during the review phase, so the Review Test Cue is a pseudo-variable for this group.

Participants

Ninety-six undergraduate students participated in partial fulfillment of a research participation requirement for psychology courses at the University at Albany, State University of New York. Forty-eight subjects were randomly assigned to the Restudy Review group and forty-eight subjects were randomly assigned to the Test Review group. All subjects reported normal or corrected-to-normal vision and English as their first language.

Materials

The target items were fifty English words, all 3-10 letters in length. Two semantic cues were selected for each of the target items from the University of South Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 1998). Two orthographic cues were also created for each target. Each orthographic cue retained atleast two letters from the target word, one of which was always the first letter. The first letter was always included because the effectiveness of an orthographic cue greatly depends on the inclusion of the first letter (MacLeod & Kampe, 1996; Tenpenny & Shoben, 1992). Forty-eight of the targets served as critical targets and the remaining two served as primacy buffers. No recency buffers were used because a distractor task intervened between the study list and the review test. One of the semantic cues selected for each target had a higher average forward associative strength (from cue to target) and backward associative strength (from target to cue) than the second semantic cue selected. To equate the two semantic cues' associative strengths to the target in the review and final test, two counterbalancing lists were created. The list of targets was arranged in descending order of forward associative strength (FAS) of the first cue. For the first target, the cue with the higher FAS was assigned as S1 (semantic cue 1), and the other corresponding semantic cue for that target as S2 (semantic cue 2). For the next target, the cue with the higher FAS was assigned to S2, and the other corresponding cue for that target to S1. This pattern of alternating assignment of the higher FAS semantic cue as S1 and S2 continued throughout the list. The final list of S1 cues had an average forward associative strength of .317 and backward associative strength of .448. The S2 cues had an average forward associative strength of .274 and backward associative strength of .435. The orthographic cues for each target were also assigned as O1 and O2 cues. The 48 critical words were then divided into 6 groups of 8 words equated for word frequency and length. (The 6 groups of target words and their frequencies are provided in the Appendix, along with the S1, S2, O1, and O2 cues associated with each target and their FAS and BAS values). Twelve counterbalancing lists were created, with each of the 6 groups of words appearing twice in each of the 6 cue conditions: Semantic/Identical, Semantic/Match, Semantic/Mismatch, Orthographic/Identical, Orthographic/Match, Orthographic/Mismatch. In Lists 1-6, the S1/O1 cues were given in the review phase and in Lists 7-12, the S2/O2 cues were given in the review phase. Although all subjects received the same final cued recall test, the distinction between which cue was given in review is a pseudo distinction for the restudy group, since they restudied the list of words and were not exposed to any cues in the review phase. Altogether there were 12 different counterbalancing study lists, each of which was studied by 4 participants in the Test Review group and 4 participants in the Restudy Review group.

Procedure

The experimental procedures are shown in Table 1. Upon arrival on the first day, subjects were randomly assigned to either the Restudy Review or the Test Review group. Each individually tested subject was seated at a computer and told that his or her memory for the studied words would be tested. The items in the study list were 50 English words, including 2 primacy buffers and 48 critical targets. To move performance in the review test off of the floor, the 50-item studied list was studied 3 times, with each word being presented for 5 seconds. Each word was studied N times before it was studied N+1 times, with different random orders in each block. An unrelated filler task was given between the study and the review phases. The filler task was experimenter-paced and consisted of a fixation (+) displayed for 1 s, followed by two 2-digit numbers presented for 5 s (a) above and below or (b) to the left and the right of the fixation. After 5 s, the correct answer was presented as feedback for 1 s. During the review phase, subjects in the Restudy Review group restudied each item one more time at the same 5-second-per-item presentation rate that occurred during the study list. In the review test, the two primacy buffers were presented first, one with an orthographic cue and the other with a semantic cue. Each of the forty-eight critical items was then tested only once, 24 with a semantic cue and 24 with an orthographic cue, in a random order. Subjects responded by typing in each answer and pressing the ENTER key to record the response. If they gave up, they also pressed the ENTER key. The interstimulus interval between the pressing of the ENTER key and the next trial was 500 ms. Subjects were given a maximum of 10 s to respond, after which the computer would automatically advance them to the next trial. Subjects were then given feedback for each item by displaying the correct target on the screen beside the cue. The feedback was followed 1500 ms later by the next cue. After the review phase was completed, all subjects were dismissed and reminded to return two days later for the final cued recall test. They were explicitly told that they would have their memories tested again, but were not told the nature of the test. The procedures for the final test were the same as the review test, except that no feedback was provided on the final test. The final cued recall test was identical for subjects in the Test Review and Restudy Review groups. Each studied word was tested only once, with each subject receiving an equal number of semantic and orthographic cues. For the Test Review group, eight of the targets appeared in each of the six conditions: Semantic/Identical, Semantic/Match, Semantic/Mismatch, Orthographic/Identical, Orthographic/Match, and Orthographic/Mismatch.

Results

Review Phase

In the review test, recall to the semantic cues (61%) and orthographic cues (63%) was virtually identical, showing that the two types of cues were equally effective as had been indicated in pilot research. Thus, if there are differential effects of the type of review cue on recall to the semantic and orthographic cues in the final test, they cannot be due to differences in retrieval success for the two types of cues during review. Rather these differences must be due to the ***relation*** between the type of cue that was given in the review test and the type of cue given for that same target word in the final test, that is, whether the review test cue was identical to the final test cue, or matched or mismatched the type of cue given for that target in the final test.

Final Cued Recall Test

Table 2.

Percent Correct Recall in the Final Cued Recall Test

Final Cue	Review Cue	Test (n = 48)	Restudy (n = 48)	Testing Effect
Semantic BEYOND	Identical BEYOND	79%	43%	36% ± 9%**
Semantic BEYOND	Match BELOW	67%	43%	24% ± 9%**
Semantic BEYOND	Mismatch A _O V _	55%	43%	12% ± 9%*
Orthographic A _ O V _	Identical A _ O V _	64%	52%	12% ± 6%**
Orthographic A _ O V _	Match A B _ V _	65%	52%	13% ± 6%**
Orthographic A _ O V _	Mismatch BELOW	59%	52%	7% ± 6%^

Note. ** $p < .001$; * $p < .01$; ^ $p = .017$, two-tailed. The value next to each ± sign is the 95% confidence interval for that mean.

Table 2 presents the mean percent correct in each of the six cuing conditions on the final cued recall test for both the Test Review and the Restudy Review groups. Keep in mind that the distinction between identical, match, and mismatch cues in the final test is a pseudo distinction for subjects in the restudy group who had no prior test with the cues in the review phase. In the final cued recall test, subjects in the Restudy Review group displayed better recall for the orthographic cues (52%) than for the semantic cues (43%), $t(47) = 2.54$, $p < .05$. Despite the equivalent effectiveness of orthographic and semantic cues in the review phase for the Test Review group, the orthographic cues were 9% more effective after a 2-day retention interval. This difference is seemingly anomalous, but one possibility is that when studying the words for a fourth time, subjects put less effort into semantic processing and perhaps attended more to the surface characteristics of the words.

The TEs are displayed in the last column of Table 2 along with their 95% confidence intervals and the associated *p* values. As shown in Table 2, a statistically significant positive TE occurred in all of the cuing conditions. Most notably, a TE occurred even in the two mismatch conditions. This shows that testing improved “far transfer” even when the transfer is between two types of tests in tasks involving episodic memory for specific words, just as transfer occurs between two types of tests in reasoning tasks (Butler, 2010).

The results in the Test Review group were submitted to a 2 x 3 analysis of variance (ANOVA) with Review Cue (semantic or orthographic) and Final Test Cue (identical, match, or mismatch) as factors. This ANOVA showed that the main effect of Review Cue was significant, $F(1,47) = 18.287$, $MSE = .396$, $p < .001$. Final test recall was better for words that were tested with semantic cues in review (M = 67.3%) than for words that were tested with orthographic cues in review (M = 62.7%). This is likely due to a levels-of-processing effect, with deep semantic processing resulting in a more durable memory trace than shallow orthographic processing (Craik & Lockhart, 1972). Most important for present purposes, the main effect of Final Test Cue was also significant, $F(2,94) = 16.959$, $MSE = .515$, $p < .001$. This shows that the magnitude of the TE depends on TAP. However, as indicated by the significant Review Cue x Final Test Cue interaction, $F(2,94) = 4.934$, $MSE = .125$, $p < .01$, the magnitude of the TAP effect depended on whether the review cue was a semantic cue or an orthographic cue, an issue to which I now turn.

When the final test cues were semantic, the evidence for TAP was strong. Tests were conducted using Bonferroni adjusted alpha levels of .017 per test (.05/3). Recall was significantly better in the Semantic/Identical condition (79%), for which the retrieval processes used in the review test and the final test would be most similar, than in the Semantic/Match condition (67%), $t(47) = 3.66$, $p = .001$. The Semantic/Match condition (67%) in turn yielded significantly better recall than the Semantic/Mismatch condition (55%), $t(47) = 3.86$, $p < .001$., which presumably would have invoked the most dissimilar retrieval processes in the review test and the final test.

The orthographic Final Test Cue conditions, however, did not show obvious evidence of TAP. Most notably, performance in the Orthographic/Identical condition (64%) was virtually identical to that in the Orthographic/Match condition (65%), $t(47) = 0.23$, $p = .818$, rather than being greater as was so for the semantic review cues. However, because the identical orthographic cue and the matching orthographic cue always shared the same first letter and because the first letter in a word is by far the most important letter for successful recall to occur to an orthographic cue (MacLeod & Kampe, 1996; Tenpenny & Shoben, 1992), this finding could be due to the fact that the processing similarity between the identical orthographic cue and the matching orthographic cues was much greater than the processing similarity between the identical semantic and the matching semantic cues. Therefore, the retrieval processes in Orthographic/Identical and Orthographic/Match conditions might not be different enough to produce a TAP effect (cf. Finlay & Benjamin, 2012).

However, if one takes into account the levels-of-processing effect that was observed in the Identical cuing conditions, there may be evidence of TAP for the orthographic cues in the final test. Subjects in the restudy group who had no prior exposure to the cues performed 9% better on the orthographic cues (52%) than on the semantic cues (43%) in the final test, indicating that the orthographic cues are more effective than the semantic cues at a 2-day retention interval. Despite this, when subjects in the test review group received the same exact cue in the review phase and on the final test, final test recall was better when the cue was semantic (79%) compared to when the cue was orthographic (64%), $t(47) = 5.14$, $p < .001$. Based on the effectiveness of the cues determined by the restudy group performance, the Orthographic/Identical condition would be expected to yield 9% better performance than the Semantic/Identical condition. However, the Orthographic/Identical condition showed 15% ***worse*** performance than the Semantic/Identical condition. This shows that there is a 24% levels-of-processing effect in the identical cue conditions. If the differences in performance in the conditions were due to a pure levels-of processing effect, we would expect the Orthographic/Mismatch condition, in which a semantic cue was given in review, to yield considerably better performance than the Orthographic/Match condition, in which an orthographic cue was given in review. However, the opposite was found. Performance in the Orthographic/Mismatch condition (59%) was 6% ***worse*** than performance in the Orthographic/Match condition (65%), but not significantly so, $t(47) = 1.61$, $p = .114$. However, this TAP effect would definitely be significant if a levels-of-processing effect “correction” were applied. Thus, after a correction is made for the levels-of-processing effect, there is strong evidence for a TAP effect in the Orthographic/Match vs. Orthographic/Mismatch conditions.

Discussion

The present experiment yielded several key results. First, the results provide support for a TAP explanation of the TE, which holds that the benefits of testing are due to compatible retrieval processes being used in the review and final tests. In addition to a levels-of-processing effect (deep semantic processing produced better recall than shallow orthographic processing), final test performance was impacted by the relation of the review test and final test cues. This was clearly shown in the semantic final test cue conditions, in which the TE was greater in magnitude in the conditions in which the review test and final test cues matched than in those in which they mismatched. At first glance, the evidence for TAP does not seem to be as clear for the orthographic final test cue conditions. However, if the differences in performance were only due to the levels-of-processing effect which was shown in the identical conditions, one would expect the Orthographic/Mismatch condition, which consisted of deep semantic processing in the review phase, to yield better performance than the Orthographic/Match condition, which consisted of shallow orthographic processing in the review phase. However, the opposite was found. Even though the Orthographic/Mismatch condition had the advantage of deeper review phase processing, the shallower review phase processing of the Orthographic/Match condition yielded better recall in the final test, presumably because it was congruent with the final test processing. This suggests that the magnitude of the TE may be linked to the relation between the processing operations used for the review test and final test retrieval cues, so that testing benefits memory to the degree that the retrieval processes used in the final test match the retrieval processes used during the review test. The present results may differ from the findings of Carpenter and DeLosh (2006), McDaniel and Masson (1985), and McDaniel et al. (1989) due to important methodological differences. The 2-day retention interval between the review and final tests in the present experiment, as opposed to the 5-min (Carpenter & DeLosh, 2006), and 1-day (McDaniel & Masson, 1985; McDaniel et al., 1989) retention intervals in the prior experiments, was long enough to obtain a TE in all conditions. Furthermore, whereas Carpenter and DeLosh (2006) failed to find a TAP effect using different types of tests which all required conceptual processing, I found that semantic/conceptual processing and orthographic/structural processing proved to be distinct enough to reveal TAP at work in the TE. While my results show a TAP effect contributing to the TE with my specific procedures (2-day retention interval and cued recall tests that require either conceptual or data-driven processing and include feedback), other mechanisms could be contributing to the TE in other situations. Further research is needed to examine the constraints on the conditions under which the TAP mediates the TE and in which conditions the TE is mediated by other mechanisms.

Second, the fact that a positive TE was found in every condition (even the conditions in which the review test cue and final test cue mismatched) suggests that although compatible review test processing and final test processing enhances the TE, the mnemonic benefits of testing are not confined to the exact review test conditions. In other words, the TE generalizes: semantic processing in review produces a TE for orthographic processing in the final test, and orthographic processing in review produces a TE for semantic processing in the final test. This is important because it shows that a “far transfer” TE occurs, not only in reasoning tasks (Butler, 2010), but also in an episodic memory test for individual words.

This finding has important implications for educational theory and practice, providing further support for learning strategies that include retrieval practice. The beneficial effects of review testing not only occur when the retrieval processes evoked by the review test “match” the retrieval processes evoked by the final test, but also occur when those retrieval processes overlap very little if at all. One possible reason why far transfer occurs is that any kind of testing strengthens the target's memory trace, thereby making it more accessible to any type of retrieval cue. Alternatively, it may be the case that semantic and orthographic cues evoke overlapping retrieval processes, that is, the retrieval of the target's phonological code (Weldon, 1991, suggests that priming requires lexical access of the studied words), which is what mediates retrieval of the target.

In general, my data have advanced knowledge of the TE by providing support for transfer-appropriate processing as a possible explanation for the TE. My data show that final test performance depends on a match in retrieval processes across review and final tests, which suggests that testing does not merely enhance the strength of a target's memory trace. Rather, review testing provides mnemonic benefits through the use of compatible review-test episodic retrieval processing. The concept of transfer-appropriate processing may help educators understand how taking tests can be beneficial to learning, that is, that testing leads subjects to engage in retrieval processes that transfer in the long-term to different retrieval environments. TAP also has implications for how educators should train students to study. TAP suggests that the efficacy of a study task depends on the processing it involves and its congruency with the processing required on the criterion test.

References

Blaxton, T. A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 657-668.

Brainerd, C. J., & Reyna, V. F. (2004). Fuzzy-trace theory and memory development. *Developmental Review, 24*, 396-439.

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.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 1547-1552.

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.

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.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of learning on testing and forgetting. *Memory & Cognition, 36*, 438-448.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*, 633-642.

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 431-437.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684.

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

Finlay, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: Evidence from the test-expectancy paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 632-652.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1441-1451.

MacLeod, C. M., & Kampe, K. E. (1996). Word frequency effects on recall, recognition, and word fragment completion tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 132-142.

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.

McDaniel, M. A., Kowitz, M. D., & Dunay, P. K. (1989). Altering memory through recall: The effects of cue-guided retrieval processing. *Memory & Cognition*, *17*, 423-434.

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 371-385.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 519-533.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://w3.usf.edu/FreeAssociation/AppendixB/index.html.

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist, 45*, 1043-1056.

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

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mestre & B. Ross (Eds.), *Psychology of learning and motivation: Cognition in education* (pp. 1-36). Oxford: Elsevier.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 233-239.

Tenpenny, P. L., & Shoben, E. J. (1992). Component processes and the utility of the conceptually-driven/data-driven distinction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 25-42.

Weldon, M. S. (1991). Mechanisms underlying priming on perceptual tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 526-541.

Appendix.

Target	Frequency	S1	FAS	BAS	S2	FAS	BAS	O1	O2
SUBTRACT	7.506	ADD	.692	.685	MINUS	.178	.243	S _ B _ R _ C T	S U _ T _ A _ T
COW	8.890	MILK	.380	.352	CALF	.690	.194	C _ W	C O _
BEE	8.098	STING	.600	.362	HONEY	.230	.220	B _ E	B E _
ARM	9.925	HAND	.100	.145	LEG	.500	.673	A _ M	A R _
TIME	13.578	CLOCK	.372	.652	WATCH	.154	.370	T _ M _	T I _ _
OLD	12.381	NEW	.473	.725	YOUNG	.236	.595	O L _	O _ D
THIN	9.565	SKINNY	.134	.264	FAT	.563	.201	T _ _ N	T H _ N
STREET	11.240	ROAD	.314	.348	AVENUE	.096	.678	S _ R _ E T	S T _ E _ T
	10.148		**.383**	**.438**		**.331**	**.397**		
CRACKER	7.102	SALTINE	.830	.106	CHEESE	.170	.141	C _ A _ K _ R	C R _ C _ E _
SCARE	8.365	AFRAID	.124	.605	FRIGHT	.464	.636	S _ A _ E	S C _ R _
SHORT	11.467	TALL	.417	.696	LONG	.222	.536	S _ O _ T	S H _ R _
TAKE	12.709	STEAL	.129	.328	GIVE	412	.444	T _ K _	T A _ _
SIT	10.153	STAND	.348	.534	CHAIR	.183	.212	S I _	S _ T
SEE	13.272	EYE	.250	.362	LOOK	.239	.678	S E _	S _ E
PIG	8.760	HOG	.200	.741	PORK	.106	.594	P I _	P _ G
PEN	8.975	PAPER	.109	.300	INK	.152	.695	P _ N	P E _
	10.100		**.301**	**.459**		**.244**	**.492**		
SAME	12.903	DIFFERENT	.534	.480	ALIKE	.142	.415	S _ M _	S _ _ E
VOLCANO	8.024	LAVA	.180	.388	ERUPT	.525	.641	V _ L _ A _ O	V O _ C _ N _
CUP	10.233	SAUCER	.520	.418	GLASS	.130	.148	C _ P	C U _
ORANGE	9.509	FRUIT	.194	.174	JUICE	.235	.655	O _ A _ G _	O R _ N _ _
SHOE	8.558	SOCK	.212	.617	FOOT	.321	.337	S _ _ E	S H _ E
WORLD	12.597	UNIVERSE	.152	.385	GLOBE	.182	.679	W _ R _ D	W O _ L _
COMB	7.386	BRUSH	.160	.636	HAIR	.150	.315	C _ M _	C _ _ B
REMEMBER	11.868	FORGET	.493	.527	RECALL	.094	.528	R _ M _ M _ _ R	R E _ E _ B E _
	10.135		**.306**	**.453**		**.222**	**.465**		

STOP	11.478	GO	.615	.539	SIGN	.112	.348	S _ O _	S _ _ P
SLOW	10.686	TURTLE	.115	.372	FAST	.527	.598	S L _ W	S _ O _
SYRUP	7.558	PANCAKES	.503	.417	MAPLE	.143	.357	S _ R _ P	S Y _ U _
WHY	12.798	HOW	.103	.412	BECAUSE	.455	.466	W _ Y	W H _
SALAD	7.820	DRESSING	.197	.428	LETTUCE	.299	.278	S _ L _ D	S A _ A _
SMART	9.786	DUMB	.409	.210	INTELLIGENT	.195	.711	S M _ R _	S _ A _ T
RUN	12.316	JOG	.140	.783	WALK	.493	.465	R U _	R _ N
ROOF	8.635	CEILING	.173	.350	SHINGLE	.118	.611	R _ O _	R O _ _
	10.135		**.282**	**.439**		**.293**	**.479**		
PRECISE	8.821	EXACT	.514	.386	ACCURATE	.167	.134	P R _ C _ S _	P _ E C _ _ E
HEAR	11.411	EAR	.260	.189	LISTEN	.500	.322	H _ A _ R	H _ _ R
NEEDLE	8.527	THREAD	.424	.758	PIN	.212	.289	N _ E _ L _	N E _ D _ _
RIGHT	12.870	CORRECT	.122	.236	WRONG	.392	.723	R _ G _ T	R I _ H _
TOOTHPASTE	6.571	TOOTHBRUSH	.296	.493	CREST	.086	.345	T _ O T_ P _ S T _	T _ O _ H P _ S T _
SKY	9.893	BLUE	.522	.284	CLOUD	.228	.346	S _ Y	S K _
PULL	10.247	TUG	.133	.580	PUSH	.594	.397	P _ _ L	P U _ _
SMALL	11.975	BIG	.351	.635	TINY	.088	.650	S _ A _ L	S M _ L _
	10.039		**.328**	**.445**		**.283**	**.401**		
MIRROR	10.208	REFLECTION	.710	.383	IMAGE	.140	.224	M I _ R _ R	M _ R _ O _
BOW	8.923	TIE	.100	.124	ARROW	.530	.400	B _ W	B O _
ABOVE	12.089	BELOW	.500	.564	BEYOND	.110	.114	A B _ V _	A _ O V _
SQUARE	9.852	ROUND	.128	.394	CIRCLE	.473	.635	S Q _ A _ E	S _ U _ R _
DARK	11.174	LIGHT	.370	.428	NIGHT	.190	.211	D _ R _	D _ _ K
WANT	13.138	NEED	.289	.599	DESIRE	.278	.610	W _ N _	W _ _ T
SPAGHETTI	7.057	MEATBALLS	.240	.675	SAUCE	.247	.468	S P _ _ H _ T T _	S _ A _ H _ T _ I
THIEF	8.361	CROOK	.091	.459	ROBBER	.224	.361	T _ I _ F	T H _ E _
	10.100		**.304**	**.453**		**.274**	**.378**		
	10.110		**.317**	**.448**		**.275**	**.435**		