Europe's Journal of Psychology
ejop.psychopen.eu | 1841-0413

Research Reports

The Effects of Test Trial and Processing Level on Immediate and Delayed Retention

Sau Hou Chang*[a]

[a] School of Education, Indiana University Southeast, New Albany, IN, USA.

Abstract

The purpose of the present study was to investigate the effects of test trial and processing level on immediate and delayed retention. A 2 × 2 × 2 mixed ANOVAs was used with two between-subject factors of test trial (single test, repeated test) and processing level (shallow, deep), and one within-subject factor of final recall (immediate, delayed). Seventy-six college students were randomly assigned first to the single test (studied the stimulus words three times and took one free-recall test) and the repeated test trials (studied the stimulus words once and took three consecutive free-recall tests), and then to the shallow processing level (asked whether each stimulus word was presented in capital letter or in small letter) and the deep processing level (whether each stimulus word belonged to a particular category) to study forty stimulus words. The immediate test was administered five minutes after the trials, whereas the delayed test was administered one week later. Results showed that single test trial recalled more words than repeated test trial in immediate final free-recall test, participants in deep processing performed better than those in shallow processing in both immediate and delayed retention. However, the dominance of single test trial and deep processing did not happen in delayed retention. Additional study trials did not further enhance the delayed retention of words encoded in deep processing, but did enhance the delayed retention of words encoded in shallow processing.

Keywords: single test trial, repeated test trial, shallow processing level, deep processing level, immediate retention, delayed retention

Europe's Journal of Psychology, 2017, Vol. 13(1), 129–142, doi:10.5964/ejop.v13i1.1131

Received: 2016-02-09. Accepted: 2016-11-22. Published (VoR): 2017-03-03.

Handling Editor: Rhian Worth, University of South Wales, Newport, United Kingdom

*Corresponding author at: School of Education, Indiana University Southeast at 4201 Grant Line Road, New Albany, IN 47150, United States. Telephone: 812-941-2606. Fax: 812-941-2667. E-mail: sauchang@ius.edu

Testing is usually viewed as a way of assessing how much students know, but is seldom seen as a way of enhancing students' learning. However, Roediger and Karpicke (2006a, 2006b) argued that taking a test had a greater positive effect than studying the material on future retention. Such an improved performance from taking a test is known as the *testing effect*.

The research design of testing effect usually includes a study phase, an intervening phase, and a test phase (e.g., Chan & McDermott, 2007; Roediger & Karpicke, 2006a). During the study phase, participants take study trials to study some set of material varying from word lists to prose passages. During the intervening phase, participants may take study trials again to study the material or take test trials to test how much they retain the material. During the test phase, participants are given a final retention test of the material. The typical finding is that those participants who take test trials outperform those who take study trials during the intervening phase.

Evidence for the testing effect in promoting learning comes from laboratory studies (e.g., Wheeler, Ewers, & Buonanno, 2003), educationally related studies (e.g., Nungester & Duchastel, 1982) and classroom studies (e.g., Leeming, 2002). Laboratory studies typically use word lists as material, and free recall as test. For example, Wheeler et al. asked participants to study a 40-word list presented at a rate of one word every 3 seconds. After the first presentation, participants in the repeated test conditions were told to take a recall test to write down as many of the words as they could recall from the list, and this process was repeated four times with 1 minute break after each recall test. On the other hand, after the first presentation, participants in the repeated study conditions were told to study the words presented at the same rate, and this process was repeated four times with 1 minute break after each study. No matter whether participants were in the repeated test or repeated study conditions, participants in the 5-min delay conditions took a recall test for the study list after five minutes, and those in the 7-day delay conditions took the recall test after 7 days. Results revealed a huge advantage for repeated study trials on the immediate free-recall test, but repeated test trials were found to be favorable on the final free-recall test given a week later.

Other laboratory studies showed how the number of test trials at retrieval affects retention. Roediger and Karpicke (2006a) had participants either study a passage three times and take one test or study a passage once and take three tests. Results showed that those who had one test trial recalled more than those who had three test trials in immediate retention, but the opposite happened in delayed retention. Wheeler and Roediger (1992) also reported that taking three tests immediately after studying a list of pictures greatly improved retention on a final test relative to taking a single test.

Dempster (1997) identified two hypotheses to account for the positive effects of test trials on learning. The first hypothesis stated that the testing effect was a result of additional exposure to material and overlearning of the material during the test trials (e.g., Thompson, Wenger, & Bartling, 1978). However, when Roediger and Karpicke (2006b) reviewed experiments with equal exposure to the material in the study trials when participants were asked to study the material several times, and in the test trials when participants were given a test several times, they still found testing effects. In addition, Wheeler et al. found that overlearning of the material with additional studying only produced better retention in the short term than repeated testing did, even though testing produced better long-term retention. If additional exposure and overlearning cannot explain the testing effect, an alternative is needed.

The second hypothesis stated that the testing effect was a result of the retrieval processes that increased the elaboration of a memory trace and multiplied retrieval routes (e.g., Bjork, 1975; Jacoby, 1978). Since recall tests that required production led to greater testing effects than recognition tests that involved identification, Bjork argued that recall tests required greater retrieval effort than recognition tests. The effortful retrieval increased the elaboration of the memory trace and enhanced the testing effect. In addition, McDaniel and Masson (1985) manipulated whether studied words were processed with semantic or phonemic encoding tasks. The testing group was given the first cued-recall tests with semantic or phonemic cues matched or mismatched the type of encoding, and the control group was dismissed. All subjects took a final cued-recall test the next day. They found that the testing group performed better on the final test when the cues for the first test mismatched the original encoding than when the cues on the first test matched the type of encoding. The effortful retrieval increased the types of retrieval routes to the memory trace and enhanced the testing effect. Therefore, effortful retrieval processes that increased the elaboration of a memory trace and multiplied retrieval routes are better able to account for the testing effect.

With a sizable research on the testing effect, several variables have been investigated: The material to be learned (e.g., Roediger & Karpicke, 2006a) the format of the test trial and final retention test (e.g., Carpenter & DeLosh, 2006), the feedback received on the test trial (Karpicke & Roediger, 2007), the time interval between study and test trials (e.g., Carpenter & DeLosh, 2005), and the interval before the final retention test (e.g., Wheeler et al., 2003). However, how to study the material or the encoding of the material receive less attention.

One way to study the material is to encode it in different levels of processing (LOP). Craik and Tulving (1975) conducted a series of experiments to explore the LOP effect on memory. To process words at different depths, they asked participants to answer various questions about the words. For example, questions about typescript encouraged shallow encodings, questions about rhymes encouraged intermediate encoding, and questions about category encouraged deep encoding. After the encoding phase was completed, participants were given a recall or recognition test for the words. Results showed that deeper encodings took longer to accomplish and were associated with higher levels of performance on the subsequent memory test.

To further investigate the shallow and deep processing, Morris, Bransford, and Franks (1977) had participants encode words phonemically (shallow processing) or semantically (deep processing). They found that semantic encoding led to greater recognition than phonemic encoding in standard recognition test. However, phonemic encoding was superior to semantic encoding given a rhyming recognition test. Encoding manipulations that directed subjects to attend to the rhymes of inputs resulted in better performance on a rhyming test than did encoding activities that prompted subjects to process the semantic meaning of inputs.

In a separate study, Kuo and Hirshman (1997) manipulated the LOP (semantic vs. letter) by asking subjects to say aloud a word that was either related in meaning to the initial word (deep processing) or to share the first letter of the initial word (shallow processing). Subjects studied a list consisting of 48 context words (exception words or pronounceable nonsense words) and 19 regular words. A free recall test was given after five minutes. Results showed that the mean proportions of regular words correctly recalled were significantly higher in the deep processing condition than those in the shallow processing condition. The LOP effect was approximately equal in the nonsense and exception word context.

In addition, Jacoby, Shimizu, Daniels, and Rhodes (2005) investigated if recognition memory was based on trace strength or familiarity, or depth of processing. In Phase 1, subjects made pleasantness judgments for 36 words in one list (deep processing) and vowel judgments (whether a word included an O or U) for 36 words in another list (shallow processing). In Phase 2, subjects received deep and shallow recognition memory tests. For the deep recognition memory test, words whose pleasantness had been judged were mixed with an equal number of new words (i.e., foils). Subjects were correctly informed that all of the “old” words in the test list were from the pleasantness-judged list. For a separate, shallow recognition memory test, the subjects were correctly informed that all “old” words in the test list were presented in the vowel judgment list. In Phase 3, three types of words appeared in a recognition memory test of foils: 36 deep foils (presented as new items in the deep recognition memory test); 36 shallow foils (presented as new items in the shallow recognition memory test); and 72 new foils (words that were not presented earlier). The subjects were instructed to judge a word as “old” if it has been presented earlier during any phase of the experiment, and to respond “new” only if the word had not been presented earlier. Results showed that attempting to recognize old items that were deeply processed during study resulted in greater depth of processing at retrieval and thus better memory for foils than did attempting to recognize items that were shallowly processed during study. In contrast to formal models of

recognition memory that highlighted the importance of quantitative criteria (e.g., strength of global familiarity in the model), specifying the source of old items or source-constrained retrieval could produce a qualitative change in the type of information used for memory judgments.

One study was found to study deep processing and testing effect. Karpicke and Smith (2012) investigated if another type of deep processing (elaboration) at encoding contributed to the testing effect of repeated retrieval. Elaboration is the process of encoding more features or attributes of an event, producing distinctive representations and multiple retrieval routes for later retrieval. They asked participants to learn word pairs across alternating study and test trials. In elaborative study conditions, participants used an imagery-based keyword method, a verbal elaboration method, or a semantic elaboration method to encode items during study trials. In the imagery-based keyword method, a mental image of a meaningful interaction (e.g., an ant drinks poison) between the keyword (*ant*) and the definition of the vocabulary word (*antiar* means poison) was produced. In the verbal elaboration method, subjects were shown a word pair (e.g., wingu-cloud) and were told to type a word (e.g., bird or sky) that would help them relate the word and English word (e.g., a bird flying in the sky). In the semantic elaboration method, the word pairs were identical (e.g., castle-castle) so that the production of verbal elaborations relating the identical word pairs would be restricted or prevented. On a criterial test one week after the learning phase, repeated test trials produced better long-term retention than repeated study trials regardless of the elaborative encoding conditions.

Karpicke and Smith (2012) did not find any type of elaborative encoding to be accountable for the testing effect of repeated retrieval. Without comparing shallow processing to deep processing, it is still unclear if LOP would be a factor affecting the benefits of retrieval practice. In addition, when McDaniel and Masson (1985) asked subjects to process words with semantic or phonemic encoding, they found the semantic or phonemic cues on the first test could affect how much subjects remembered on the second test. With additional time exposed to the material in the first test, the testing effect could not be attributed to LOP. Further study is needed to examine whether LOP would be a factor mediating the testing effect.

Therefore, the purpose of the present study was to investigate the effects of test trial and processing level on immediate and delayed retention. Research questions included (a) Was there any difference between single test trial and repeated test trial on immediate and delayed retention? The testing effect expected that single test trial enhanced immediate retention but repeated test trial enhanced delayed retention (e.g., Wheeler & Roediger, 1992). (b) Was there any difference between shallow and deep processing on immediate and delayed retention? The level of processing effect expected that deep processing enhanced immediate and delayed retention (e.g., Craik & Tulving, 1975). (c) Was there any difference between test trials and processing level on final recall? Previous studies expected that there was an interaction among test trial, processing level and final recall (e.g., Karpicke & Smith, 2012; McDaniel & Masson, 1985).

Method

Participants

Seventy-six college students (mean age = 21.3 years old; range = 19 – 27 years old; Male = 8; Female = 68) completed the immediate and delayed tests in partial fulfillment of a psychology course requirement. At the beginning, ninety-one college students were invited to participate in the present study. Data of fifteen

participants were discarded because nine of them were over 27 years-old (to keep the age range within 10), two of them did not show up for the delayed free-recall test, and four of them did not follow instruction to provide complete data. The procedures met all American Psychological Association (APA) ethical principles for use of human subjects (APA, 2002), and participants were provided informed consent in accordance with guidelines set by the Institutional Review Board of the university.

Materials

Forty stimulus words were taken from the words used by Craik and Tulving (1975, Experiment 9, see Table 1). From the MRC Psycholinguistic Database (Wilson, 1998), several properties of the stimulus words were obtained. The average number of letters was 4.75 (*SD* = .74), the average number of syllables was 1.23 (*SD* = .42), the average printed Kucera-Francis word frequency was 16.3 per million (*SD* = 14.45), the average concreteness rating was 571.91 (*SD* = 40.7), and the average familiarity rating was 507.73 (*SD* = 54.06).

Table 1

Stimulus Words and Category Questions

Word	Category Question	Word	Category Question
Bear	a wild animal	Lamp	a type of furniture
Brake	a part of a car	Lane	a type of road
Brush	used for cleaning	Lark	a type of bird
Cart	a type of vehicle	Mast	a part of a ship
Chapel	a type of building	Monk	a type of clergy
Cheek	a part of the body	Nurse	associated with medicine
Cherry	a type of fruit	Pail	a type of container
Clip	a type of office supply	Pond	a body of water
Copper	a type of metal	Rice	a type of grain
Drill	a type of tool	Roach	a type of insect
Earl	a type of nobility	Robber	a type of criminal
Fence	found in the garden	Sheep	a type of farm animal
Fiddle	a musical instrument	Soap	a type of toiletry
Flame	something hot	Sonnet	a written form of art
Flour	used for cooking	Speech	a form of communication
Glove	something to wear	Tire	a round object
Gram	a type of measurement	Tribe	a group of people
Grin	a human expression	Trout	a type of fish
Honey	a type of food	Witch	associated with magic
Juice	a type of beverage	Wool	a type of material

Design

A 2 × 2 × 2 mixed ANOVAs was used with two between-subject factors of test trial (single test, repeated test) and processing level (shallow, deep), and one within-subject factor of final recall (immediate, delayed). Participants were randomly assigned to the single test and the repeated test trials. They were then randomly assigned again to the shallow processing level and the deep processing level. Therefore, there were 38 participants in each test trial (single test and repeated test) and each processing level (shallow and deep).

The design for the test trial was based on that in Roediger and Karpicke (2006a) and Wheeler and Roediger (1992). In the single test trial, participants studied the stimulus words three times and took one free-recall test in

each cycle. In the repeated test trial, participants studied the stimulus words once and took three consecutive free-recall tests in each cycle. There were three cycles of study/test trials (either SSST or STTT) for 12 trials total. There were nine study and three test trials in the single test trial, and there were three study and nine test trials in the repeated test trial.

In the shallow processing level, participants were asked whether each stimulus word was presented in capital letter or in small letter. In the deep processing level, participants were asked whether each stimulus word belonged to a particular category (see Table 1). The final immediate free-recall test was administered five minutes after the 12 study and test trials, whereas the delayed free-recall test was administered one week later.

Procedure

Participants were tested in groups of five or fewer. They were told to study and recall a list of words, and answer some questions to help them remember the words. The task was programmed by E-prime experimental software (Version 1.1; Schneider, Eschman, & Zuccolotto, 2002). Before the word list was presented, participants were given a practice list of two words to familiarize themselves with the task and the presentation rate, and a practice recall test to familiarize themselves with the testing procedure.

There were a learning phase and a testing phase after the practice. The learning phase consisted of 12 study and test trials and took about 30 minutes. At the beginning of each study trial, participants were asked to rest their hands on a key labeled “yes” and the other on a key labeled “no” on the computer keyboard. First, a “Ready” prompt was shown on the computer screen for 1 s. The typescript question or category question was then shown for 1 s, and participants were asked to answer the question by pressing the appropriate key. The typescript question was asked in the form, “Is the word in capital letter?” or “Is the word in small letter?” The category question was asked in the form, “Is the word (a category)?” Both typescript and category questions were counterbalanced, so that half of the answers to the questions was “yes” and half was “no.”

The purpose of the question was to induce the participant to process the word at a relatively shallow level (typescript questions) or at a relatively deep level (category questions). No matter if participants answered the typescript or category questions, stimuli words were presented on a computer at 2 s per word and the screen proceeded to the next word after 2 s. To present 40 stimuli words, the total time for one study trial was 80 s. Participants who were not able to answer the questions correctly over 80% were discarded from the analysis.

The beginning of each test trial was indicated by a tone (presented over headphones for 0.5 s) and a “Recall” prompt that remained on the computer screen throughout the test. During each test trial, participants were given 80 s to write down as many of the words as possible, in any order, on a response booklet. Therefore, the time of exposure to materials on study trials and test trials was equated (both are 80 s). The transition from one test trial to another (in the repeated test condition) was indicated by a tone as well as a change in the background color on the computer screen: The background was blue during the first test, green during the second test, and red during the third test. At the end of each test trial, participants were instructed to turn to the next page on their response booklets and not to look back at any of their previous responses at any time during the learning phase.

After the learning phase of three cycles of 12 study and test trials, participants proceeded to the testing phase and were asked to complete mazes for five minutes. Participants were then given an immediate free-recall test to write down as many of the words as they could recall in 10 minutes, and were instructed to draw a line on

their recall sheet to mark their progress at one minute intervals. This procedure ensured that participants had exhausted their knowledge by the end of the 10 minutes recall test and allowed the researcher to measure the number of words recalled.

All participants, except two, returned for the delayed free-recall test one week later. They were given 10 minutes to write down as many of the words as they could recall, and were instructed to draw a line on their recall sheet to mark their progress at one minute intervals. Finally, participants were asked whether they expected to be given a test in the second session and whether they consciously rehearsed the test items after the first session. At the end of the delayed free-recall test, participants were debriefed and thanked for their participation.

Results

The mean number of correct words recalled out of 40 words on the immediate and delayed free-recall tests is presented in Figure 1, as a function of test trial (single test, repeated test) and processing level (shallow, deep). A significance level of .05 is used for all analyses in this study. A 2 test trial (single test vs. repeated test) × 2 processing levels (shallow vs. deep) × 2 final recall (immediate vs. delay) mixed Analysis of Variance (ANOVA) revealed a main effect of final recall, $F(1, 72) = 400.446$, $p < .001$, partial $\eta^2 = .848$. Effect size indicated a high proportion of variance accounted for by final recall. Further pairwise comparisons using a Bonferroni correction showed that the mean number of words recalled in five minutes (23.868) was significantly higher than words recalled in one week (14.395), $p < .001$.

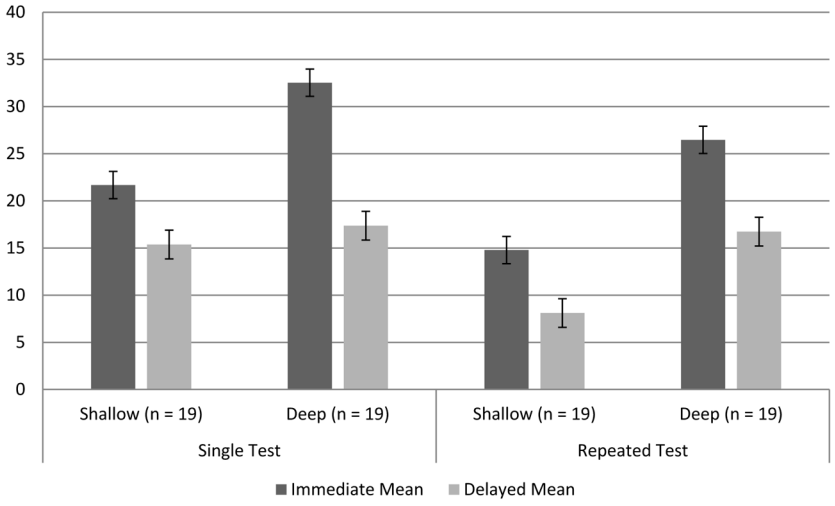

Figure 1. Means and standard error of the number of words (Total = 40) recalled in immediate and delayed final recall by test trial and processing level (N = 76).

Results showed a main effect of test trial, $F(1, 72) = 13.7$, $p < .001$, partial $\eta^2 = .160$. Effect size revealed low strength in associations. Further pairwise comparisons using a Bonferroni correction showed that the mean number of words recalled in single test trial (21.737) was significantly higher than those recalled in repeated test trial (16.526), $p < .001$. There was also a main effect of processing level, $F(1, 72) = 34.676$, $p < .001$, partial $\eta^2 = .325$. Further pairwise comparisons using a Bonferroni correction showed that the mean number of words

recalled in deep processing level (23.276) was significantly higher than those recalled in shallow processing level (14.987), $p < .001$.

There was an interaction between final recall and test trial, $F(1, 72) = 7.119$, $p = .009$, partial $\eta^2 = .09$. There was an interaction between final recall and processing level, $F(1, 72) = 39.454$, $p < .001$, partial $\eta^2 = .354$. No interaction was found between test trial and processing level, $F(1, 72) = 1.762$, $p = .189$, partial $\eta^2 = .024$. There was an interaction between final recall, test trial and processing level, $F(1, 72) = 9.347$, $p = .003$, partial $\eta^2 = .115$.

Simple main effects analysis was then conducted to investigate the interaction between final recall and test trial (Table 2). In immediate recall, the number of words recalled at the single test trial ($M = 27.11$, $SD = 8.611$) was significantly higher than those recalled in repeated test trial [$M = 20.63$, $SD = 8.274$; $F(1, 72) = 20.027$, $p < .001$]. In delayed recall, the number of words recalled at the single test trial ($M = 16.37$, $SD = 7.134$) was significantly higher than those recalled in repeated test trial [$M = 12.42$, $SD = 7.417$; $F(1, 72) = 6.719$, $p = .012$]. In single test trial, the number of words in immediate recall ($M = 27.11$, $SD = 8.611$) was significantly higher than those in delayed recall [$M = 16.37$, $SD = 7.134$; $F(1, 72) = 257.176$, $p < .001$]. In repeated test trial, the number of words in immediate recall ($M = 20.63$, $SD = 8.274$) was significantly higher than those in delayed recall [$M = 12.42$, $SD = 7.417$; $F(1, 72) = 150.39$, $p < .001$].

Table 2

Means and Standard Deviations of the Number of Words (Total = 40) Recalled in Immediate and Delayed Final Recall by Test Trial (N = 76)

	Immediate		Delay	
Test Trial	***M***	***SD***	***M***	***SD***
Single Test (*N* = 38)	27.11	8.611	16.37	7.134
Repeated Test (*N* = 38)	20.63	8.274	12.42	7.417

Simple main effects analysis was also conducted to investigate the interaction between final recall and processing level (Table 3). In immediate recall, the number of words recalled at deep processing ($M = 29.5$, $SD = 6.185$) was significantly higher than those recalled in shallow processing [$M = 18.24$, $SD = 7.793$; $F(1, 72) = 60.621$, $p < .001$]. In delayed recall, the number of words recalled at deep processing ($M = 17.05$, $SD = 6.363$) was significantly higher than those recalled in shallow processing [$M = 11.74$, $SD = 7.675$; $F(1, 72) = 12.186$, $p = .001$]. In shallow processing, the number of words in immediate recall ($M = 18.24$, $SD = 7.793$) was significantly higher than those in delayed recall [$M = 11.74$, $SD = 7.675$; $F(1, 72) = 94.255$, $p < .001$]. In deep processing, the number of words in immediate recall ($M = 29.5$, $SD = 6.185$) was significantly higher than those in delayed recall [$M = 17.05$, $SD = 6.363$; $F(1, 72) = 345.646$, $p < .001$].

Table 3

Means and Standard Deviations of the Number of Words (Total = 40) Recalled in Immediate and Delayed Final Recall by Processing Level (N = 76)

	Immediate		Delay	
Processing Level	***M***	***SD***	***M***	***SD***
Shallow (*N* = 38)	18.24	7.793	11.74	7.675
Deep (*N* = 38)	29.50	6.185	17.05	6.363

Another simple main effects analysis was conducted to investigate the interaction among final recall, test trial and processing level (Table 4). In shallow processing and immediate recall, the number of words recalled at the single test trial (M = 21.684, SD = 8.226) was significantly higher than those recalled in repeated test trial [M = 14.789, SD = 5.663; $F(1, 72) = 11.358$, $p = .001$]. In shallow processing and delayed recall, the number of words recalled at the single test trial (M = 15.37, SD = 8.348) was significantly higher than those recalled in repeated test trial [M = 8.11, SD = 4.852; $F(1, 72) = 11.374$, $p = .001$]. In deep processing and immediate recall, the number of words recalled at the single test trial (M = 32.526, SD = 4.765) was significantly higher than those recalled in repeated test trial [M = 26.474, SD = 6.05; $F(1, 72) = 8.753$, $p = .004$]. In deep processing and delayed recall, no difference was found between the number of words recalled at the single test trial (M = 17.37, SD = 5.727) and those recalled in repeated test trial [M = 16.74, SD = 7.086; $F(1, 72) = .086$, $p = .77$].

Table 4

Means and Standard Deviations of the Number of Words (Total = 40) Recalled in Immediate and Delayed Final Recall by Test Trial and Processing Level (N = 76)

	Immediate				Delayed			
	Single		Repeated		Single		Repeated	
Processing Level	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Shallow (*n* = 38)	21.68	8.226	14.79	5.663	15.37	8.348	8.11	4.852
Deep (*n* = 38)	32.53	4.765	26.47	6.050	17.37	5.727	16.74	7.086

In single test trial and immediate recall, the number of words in deep processing (M = 32.526, SD = 4.765) was significantly higher than those in shallow processing [M = 21.684, SD = 8.226; $F(1, 72) = 28.087$, $p < .001$]. In single test trial and delayed recall, no difference was found between the number of words in deep processing (M = 17.37, SD = 5.727) and those in shallow processing [M = 15.37, SD = 8.348; $F(1, 72) = .862$, $p = .356$]. In repeated test trial and immediate recall, the number of words in deep processing (M = 26.474, SD = 6.05) was significantly higher than those in shallow processing [M = 14.789, SD = 5.663; $F(1, 72) = 32.619$, $p < .001$]. In repeated test trial and delayed recall, the number of words in deep processing (M = 16.74, SD = 7.086) was significantly higher than those in shallow processing [M = 8.11, SD = 4.852; $F(1, 72) = 16.064$, $p < .001$].

In single test trial and shallow processing, the number of words in immediate recall (M = 21.684, SD = 8.226) was significantly higher than those in delayed recall [M = 15.37, SD = 8.348; $F(1, 72) = 44.494$, $p < .001$]. In single test trial and deep processing, the number of words in immediate recall (M = 32.526, SD = 4.765) was significantly higher than those in delayed recall [M = 17.37, SD = 5.727; $F(1, 72) = 256.286$, $p < .001$]. In repeated test trial and shallow processing, the number of words in immediate recall (M = 14.789, SD = 5.663)

was significantly higher than those in delayed recall [M = 8.11, SD = 4.852; F(1, 72) = 49.836, p < .001]. In repeated test trial and deep processing, the number of words in immediate recall (M = 26.474, SD = 6.05) was significantly higher than those in delayed recall [M = 16.74, SD = 7.086; F(1, 72) = 105.751, p < .001].

The free-recall tests in the learning phase were also analyzed to see if results were similar to the free-recall tests in the testing phase (immediate and delay). Another 2 test trial (single test vs. repeated test) × 2 processing levels (shallow vs. deep) × 3 recall (recall 1 vs. recall 2 vs. recall 3) mixed Analysis of Variance (ANOVA) was conducted. Similar results were found with main effects in test trial, F(1, 72) = 28.704, p < .001, partial η^2 = .285; processing level, F(1, 72) = 32.054, p < .001, partial η^2 = .308; and recall, F(2, 144) = 124.216, p < .001, partial η^2 = .633. Table 5 shows that the mean number of words recalled in single test trial (15.588) was significantly higher than those recalled in repeated test trial (10.64); those in deep processing level (15.728) was significantly higher than those in shallow processing level (10.50), those in the third recall test (16.618) was significantly higher than those in the second (13.697) and first recall tests (9.026) in the learning phase. However, no interactions were found between recall and test trial, F(2, 144) = .362, p = .697, partial η^2 = .005, recall and processing level, F(2, 144) = 1.593, p = .207, partial η^2 = .022; test trial and processing level, F(1, 72) = .159, p = .691, partial η2 = .002; and among recall, test trial and processing level, F(2, 144) = .356, p = .701, partial η2 = .005.

Table 5

Means and Standard Deviations of the Number of Words (Total = 40) Recalled in Test Trial by Processing Level in the Learning Phase (N = 76)

Factor	*M*	*SD*
Test Trial		
Single	15.59	5.528
Repeated	10.64	5.044
Processing Level		
Shallow	10.50	5.191
Deep	15.73	5.283
Recall		
Recall 1	9.03	5.317
Recall 2	13.69	5.767
Recall 3	16.62	6.511

Discussion

The present study investigated the effects of test trial and processing level on immediate and delayed retention. There was an interaction between final recall and test trial; between final recall and processing level; and among final recall, test trial and processing level. However, no interaction was found between test trial and processing level.

The finding that participants in single test trial recalled more words than repeated test trial in immediate final free-recall test was consistent with previous studies that single test trial produced more short-term benefits than repeated test trial (Roediger & Karpicke, 2006a; Wheeler et al., 2003). However, the dominance of single test

trial over repeated test trial in delayed retention was different from previous studies that repeated test trial produced more long-term benefits than single test trial (Roediger & Karpicke, 2006a; Wheeler et al., 2003).

Participants in the single test trial were exposed to the words in 9 study trials and those in the repeated test trial were exposed to the words in 3 study trials. The additional exposure to the words in single test trial may lead to overlearning of the words and better retention on immediate and delayed test.

On the other hand, participants in the repeated test trial took 9 test trials and those in the single test trial took 3 test trials. The additional test trials were supposed to give participants in the repeated test trial more retrieval practice. Bjork (1975) stated that the retrieval process increased the elaboration of memory trace and enhanced the testing effect. However, Duchastel (1981) noted that the free-recall test contained no cues to assist participants in answering the test and might therefore result in recall of only part of the contents, or a lesser testing effect. With less exposure time to the words in the study trials and no cues to assist free recall in the test trials, participants in the repeated test trial failed to produce greater benefits on the delayed recall test.

The finding that participants in deep processing performed better than those in shallow processing in both immediate and delayed retention was consistent with previous studies that deeper encodings led to higher levels of performance on subsequent retention test (Craik & Tulving, 1975; Jacoby et al., 2005). The effort participants put forth to differentiate if each stimulus word belonged to a particular category promoted a deep processing of the words whereas the effort to differentiate if the words were presented in capital letters encouraged a shallow processing. Craik and Tulving explained that memory performance depended on the elaborateness of the encoding, and retention was enhanced when the encoding context was more fully descriptive.

Even though no interaction was found between test trial and processing level, there was an interaction among test trial, processing level and final recall. No matter whether it was in shallow or deep processing, participants in single test trial performed better than those in repeated test trial in immediate retention. The advantage of single test trial over repeated test trial carried over to delayed retention in shallow processing, but not in deep processing. Similarly, no matter whether it was in immediate or delayed retention, participants in deep processing performed better than shallow processing in repeated test trial. The advantage of deep processing over shallow processing carried over to single test trial in immediate retention, but not delayed retention.

The most interesting finding was the delayed retention of participants in the deep processing and single test trial. The current study showed a dominance of single test trial over repeated test trial and deep processing over shallow processing in retention. However, the dominance of single test trial and deep processing did not happen in delayed retention.

Even though participants in the repeated test trial were only exposed to the words in 3 study trials, their delayed retention was the same as those in the single test trial who were exposed to the words in 9 study trials. When participants studied the words in deep processing, the number of study and test trials did not matter. No testing effect was found because the repeated test trial still did not outperform the single test trial in delayed retention. Kang, McDermott, and Roediger (2007) pointed that testing could be of little help when very few items were successfully retrieved on test trials. A further look at the recall performance in the learning phase found that participants in the repeated test trial did not recall more items than those in the single test trial. With a disadvantage of the fewer items retrieved in the learning phase, the repeated test trials managed to perform the

same as the single test trial in delayed retention, but did not perform well enough to bring the testing effect. Retrieval practice was only beneficial to memory when retrieval was successful.

The advantage of deep processing over shallow processing prevailed when participants only studied the words 3 times in repeated study trial. However, when participants studied the words 9 times in single test trial, the deep processing advantage disappeared in delayed retention. When participants studied more times, the depth of processing did not mediate the delayed retention. Craik (2002) stated that initial encoding determined the potential for later retrieval, while retrieval environment determined the degree to which that potential will be realized. Deep processing has the potential for assisting later performance but the retrieval environment makes the potential possible. Even though shallow processing does not have the potential for greater retrieval, the number of study trials may have increased the odds of the environment for greater retrieval.

Conclusion

The present study found the level of processing effect or the superiority of deep processing over shallow processing on subsequent retention tests, but did not find any testing effect or the superiority of repeated testing over simple testing on subsequent retention tests. Even though testing effect was not found in delayed retention, the depth of processing did mediate the delayed retention.

In deep processing, participants managed to perform the same no matter whether they were in single or repeated test trial. It showed that the number of study and test trials did not affect the delayed retention when participants studied the words in deep processing. Once participants established a connection of the word to the category in 3 study trials, the additional 6 study trials did not further enhance retention.

In single test trial when participants studied the words 9 times, they performed the same no matter whether they processed the words in shallow or deep encoding. It showed that the number of study and test trials affected the level of processing effect. Even when participants processed the words in shallow encoding, they could perform as well as those in deep processing when both studied the words 9 times.

In conclusion, additional study trials did not further enhance the delayed retention of words encoded in deep processing, but did enhance the delayed retention of words encoded in shallow processing.

Funding

The author has no funding to report.

Competing Interests

The author has declared that no competing interests exist.

Acknowledgments

The author has no support to report.

References

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *The American Psychologist, 57*, 1060-1073. doi:10.1037/0003-066X.57.12.1060

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

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology, 19*, 619-636. doi:10.1002/acp.1101

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*, 268-276. doi:10.3758/BF03193405

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. doi:10.1037/0278-7393.33.2.431

Craik, F. I. M. (2002). Levels of processing: Past, present…and future? In M. A. Conway (Ed.), *Levels of processing 30 years on: Special issue of memory* (pp. 305-318). Hove, United Kingdom: Psychology Press.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268-294. doi:10.1037/0096-3445.104.3.268

Dempster, F. N. (1997). Using tests to promote classroom learning. In R. F. Dillon (Ed.), *Handbook on testing* (pp. 332-346). Westport, CT, USA: Greenwood Press.

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

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17*, 649-667. doi:10.1016/S0022-5371(78)90393-6

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review, 12*, 852-857. doi:10.3758/BF03196776

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

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 33*, 704-719. doi:10.1037/0278-7393.33.4.704

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67*, 17-29. doi:10.1016/j.jml.2012.02.004

Kuo, T.-M., & Hirshman, E. (1997). The role of distinctive perceptual information in memory: Studies of the testing effect. *Journal of Memory and Language, 36*, 188-201. doi:10.1006/jmla.1996.2486

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29*, 210-212. doi:10.1207/S15328023TOP2903_06

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 371-385. doi:10.1037/0278-7393.11.2.371

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. doi:10.1016/S0022-5371(77)80016-9

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology, 74*, 18-22. doi:10.1037/0022-0663.74.1.18

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

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

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime user's guide*. Pittsburgh, PA. USA: Psychology Software Tools.

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 210-221. doi:10.1037/0278-7393.4.3.210

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*, 571-580. doi:10.1080/09658210244000414

Wheeler, M. A., & Roediger, H. L., III. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3*, 240-245. doi:10.1111/j.1467-9280.1992.tb00036.x

Wilson, M. D. (1998). MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers, 20*, 6-10. doi:10.3758/BF03202594

About the Author

Sau Hou Chang is an Associate Professor in Educational Psychology at School of Education at Indiana University Southeast, USA. She teaches Educational Psychology, Psychology of Teaching, Classroom Assessments, and Assessment in Schools. She also supervises clinical experiences and practice of the Elementary Education students.