Memory & Cognition
2010, 38 (8), 995-1008
doi:10.3758/MC.38.8.995

The testing effect in free recall is associated with enhanced organizational processes

Franklin M. Zaromb and Henry L. Roediger III
Washington University, St. Louis, Missouri

In two experiments with categorized lists, we asked whether the testing effect in free recall is related to enhancements in organizational processing. During a first phase in Experiment 1, subjects studied one list over eight consecutive trials, they studied another list six times while taking two interspersed recall tests, and they learned a third list in four alternating study and test trials. On a test 2 days later, recall was directly related to the number of tests and inversely related to the number of study trials. In addition, increased testing enhanced both the number of categories accessed and the number of items recalled from within those categories. One measure of organization also increased with the number of tests. In a second experiment, different groups of subjects studied a list either once or twice before a final criterial test, or they studied the list once and took an initial recall test before the final test. Prior testing again enhanced recall, relative to studying on the final test a day later, and also improved category clustering. The results suggest that the benefit of testing in free recall learning arises because testing creates retrieval schemas that guide recall.

A robust finding is that testing a person's memory for previously learned material enhances long-term retention, relative to restudying the material for an equivalent amount of time (e.g., Carrier & Pashler, 1992; for a review, see Roediger & Karpicke, 2006a). This finding, known as the *testing effect*, has been demonstrated using a wide range of study materials and types of tests, in both laboratory and classroom settings and in various subject populations (e.g., Butler & Roediger, 2007; Gates, 1917; Kang, McDermott, & Roediger, 2007; McDaniel, Anderson, Derbish, & Morrisette, 2007; Roediger & Karpicke, 2006b; Spitzer, 1939; Tse, Balota, & Roediger, in press). Recent years have seen renewed interest among researchers investigating the potential benefits of testing for learning as a means to improving learning in educational settings (McDaniel, Roediger, & McDermott, 2007; Pashler, Rohrer, Cepeda, & Carpenter, 2007).

One limitation with this work is that testing effects typically report improvements in learners' retention of discrete facts (e.g., foreign vocabulary words) without necessarily demonstrating a better understanding of the subject matter through testing (Daniel & Poole, 2009). However, a growing body of research has shown that testing can serve as a versatile learning tool by enhancing the long-term retention of nontested information that is conceptually related to previously retrieved information (Chan, 2009; Chan, McDermott, & Roediger, 2006), by stimulating the subsequent learning of new information (Izawa, 1970; Karpicke, 2009; Szpunar, McDermott, & Roediger, 2008; Tulving & Watkins, 1974) and by permitting better transfer to new questions (Butler, 2010; Johnson & Mayer, 2009; Rohrer, Taylor, & Sholar, 2010). In the present research, we further examine the potential benefits of testing by asking whether testing can improve individuals' learning and retention of the conceptual organization of study materials, relative to studying the materials alone—a question not yet addressed in the literature.

Psychologists have long grappled with questions of how the processes involved in mentally organizing information influence learning and retention (e.g., Ausubel, 1963; Bartlett, 1932; Katona, 1940). One theoretical assumption that has guided much of the cognitive research examining organization and learning was Miller's (1956) conception of recoding, or *chunking*, in which he argued that the key to learning and retaining large quantities of information was to mentally repackage, or *chunk*, the study materials into smaller units. Evidence for chunking has come primarily from studies using serial recall and free recall paradigms in which subjects often study and attempt to recall verbal materials such as lists of words over multiple alternating study and test trials (e.g., Bower & Springston, 1970; Tulving, 1962), but it has also come from other techniques (e.g., Mandler, 1967).

In support of the chunking hypothesis, researchers have pointed to the finding that when people study lists of words coming from different conceptual categories in a randomized order, they tend to recall them in an organized fashion by clustering conceptually related responses together (W. A. Bousfield, 1953; W. A. Bousfield, Cohen, & Whitmarsh, 1958). Furthermore, response clustering is often associated with greater retention (Mulligan, 2005; Puff, 1979). Similarly, Tulving (1962) found that when students learned

F. M. Zaromb, fzaromb@ets.org

a list of seemingly unrelated words, they recoded groups of items into higher order subjective units; furthermore, this organizing tendency, referred to as *subjective organization*, was predictive of free recall. Subjective organization is presumed to be reflected in the degree to which recall protocols become more consistent over multiple study and test trials, even though the sequence of item presentation changes from trial to trial. Mandler (1967) also showed powerful effects of organization on recall; after subjects sorted unrelated words into consistent groupings, they remembered them better than did subjects in other conditions exposed to the words the same number of times.

One question that was never addressed in this line of research is whether organizational phenomena such as category clustering and subjective organization are determined by processes that occur during study trials, during test trials, or both. In the present research, we investigated the effects of repeated testing on organization by manipulating the number of study trials and test trials in learning a list of categorized words. We equated the conditions of studying and testing by allotting the same amount of time for study and test trials and by equating the total number of study and test trials in each learning condition. Of interest was whether varying the number of times that subjects studied or attempted to recall lists of categorized words would affect the level of recall and how recall was organized in both initial and delayed tests of free recall. Cued recall tests were also included.

We focused on several different measures to examine recall performance and organization. Total recall was measured by the proportion of all words recalled from each list. Recall of the categorized lists was also decomposed into two components that bear a multiplicative relationship to total recall: category recall (Rc) and recall of items within categories (Rw/c; Tulving & Pearlstone, 1966). Rc is defined as the number of times at least one member of a taxonomic category represented in the original study list is recollected, and Rw/c is the average number of items recalled from each of the list categories represented in a subject's output protocol (Cohen, 1963). The measures index how many categories can be recalled and the completeness of the recall from the categories once accessed.

The organization of recall was measured using the adjusted ratio of clustering (ARC; Roenker, Thompson, & Brown, 1971). ARC assesses the degree to which subjects' recall patterns correspond to the conceptual structure of the study materials and is also considered a relatively pure measure of organization, because it controls for differences in recall level across subjects or learning conditions. ARC quantifies the extent to which subjects tend to cluster responses according to taxonomic categories (or other predefined types of categories). ARC scores range in value from −1.0 to 1.0, where 0 indicates that the amount of clustering reflected in subjects' response patterns is no greater than that expected by chance alone, and 1.0 indicates perfect clustering. By contrast, negative scores may reflect atypical patterns of recall organization not captured by traditional category clustering measures (for reviews of ARC and other clustering measures, see Kahana, Howard, & Polyn, 2008; Murphy, 1979; Murphy & Puff, 1982; Pellegrino & Hubert, 1982).

Another form of organization that may be directly influenced by retrieval practice is subjective organization (e.g., Mulligan, 2002). Even with the use of categorized lists, subjects may tend to adopt idiosyncratic forms of conceptual organization to chunk list items into higher order subjective units, or they may adopt uniform organization within category recall. The measure of subjective organization that we used is bidirectional intertrial repetition (A. K. Bousfield & W. A. Bousfield, 1966; W. A. Bousfield, Puff, & Cowan, 1964), also called *pair frequency* (PF; Sternberg & Tulving, 1977). PF represents the number of pairs of items recalled on adjacent test trials in adjacent output positions in either forward or reverse order. Moreover, PF takes into account the baseline level of subjective organization that might be expected by chance alone in a given recall protocol. The measure can go from 0 (chance organization) to much higher levels (depending on the number of items and pairs recalled).

Of course, there are other measures of organization, and debates surrounding the issue of which is the best measure have not been resolved (Murphy, 1979). The measures that we employed are commonly accepted in the literature and, when used in combination, provide a comprehensive picture of how testing affects the learning and utilization of organizational information to aid episodic retrieval, relative to studying alone.

EXPERIMENT 1

The purpose of Experiment 1 was to examine the effects of repeated studying and testing on the learning and retention of lists of words representing different taxonomic categories. We varied the number of times that subjects studied or attempted to recall the lists of categorized words and measured how this manipulation affected memory performance, as measured by total word recall, Rc, and words per category recall (Rw/c), and how it affected organization of recall, as measured by response output organization (ARC, PF). Specifically, subjects were asked to study or attempt to recall three lists of 50 words sampled from 10 categories (with five instances per category). There were three conditions: The subjects studied one list eight consecutive times without taking any tests, the subjects studied a second list six times and were tested twice, or the subjects studied a third new list over four alternating study and test trials. Two days later, the subjects were asked to recall as many words as they could remember from the three lists.

Method

Subjects. Thirty-six Washington University undergraduates participated for either payment or course credit.

Design. We manipulated three learning conditions in a within-subjects design. In one learning condition (study only), the subjects completed eight consecutive study trials and no test trials (SS SS SS SS, where S represents an individual study trial). In a second condition (the two-test condition), the subjects completed two test trials and six study trials, according to the following se-

Table 1
Mean Proportion of Words Recalled, Number of Categories Recalled (Rc), Number of Words per Category Recalled (Rw/c), and Adjusted Ratio of Clustering (ARC) Scores for the Two-Test and Four-Test Conditions in Initial Tests of Free Recall in Experiment 1

		Initial Test							
		Test 1		Test 2		Test 3		Test 4	
Measure	Condition	*M*	CI	*M*	CI	*M*	CI	*M*	CI
Recall	Two test	.31	.04			.52	.06		
	Four test	.33	.04	.51	.04	.56	.06	.63	.07
Rc	Two test	6.58	0.68			8.25	0.46		
	Four test	6.61	0.72	8.36	0.38	8.78	0.38	8.83	0.46
Rw/c	Two test	2.34	0.20			3.13	0.24		
	Four test	2.45	0.18	3.02	0.20	3.14	0.26	3.48	0.30
ARC	Two test	.66	.08			.73	.08		
	Four test	.68	.08	.78	.04	.79	.06	.82	.05

Note—CI, 95% confidence interval.

quence: ST SS ST SS, where T represents an individual test trial. In the third condition (the four-test condition), the subjects completed four alternating study and test trials: ST ST ST ST—the standard condition in free recall learning (e.g., Tulving, 1962). Words were presented in a different randomized order on each study trial, but each learning sequence (SS SS SS SS, ST SS ST SS, or ST ST ST ST) occurred in blocks that were counterbalanced with list presentation such that each list was matched to every learning sequence an equal number of times across subjects.

Materials. One hundred fifty words were sampled from 30 categories (5 words per category) in the expanded and updated version of the Battig and Montague word norms (Van Overschelde, Rawson, & Dunlosky, 2004) to create three 50-word study lists. The 50 words in each list included five medium frequency nouns belonging to each of 10 taxonomic categories.

Procedure. The subjects participated in two sessions scheduled 2 days apart. In the first session, the subjects were informed that they would be asked to study and recall several lists of words presented by a computer. The session consisted of 24 trials in which the subjects studied or were tested on their ability to recall each of the three lists a total of eight times. The number of study and test trials varied for each list depending on the learning condition (study-only, two-test, or four-test condition). During the study trials, the computer displayed each word one at a time for 2 sec, followed by a 400-msec interstimulus interval. Each list was presented in a different color and location on the computer screen. Words in the first list appeared in green text in the upper left-hand quadrant of the screen. Words in the second list appeared in yellow text in the upper right-hand quadrant. Finally, words in the third list appeared in red text in the lower right-hand quadrant. (The reason for the different presentation formats will become clear later.) The subjects were asked to read the words aloud during list presentation. The total study time was 2 min per trial.

Across the three learning conditions, each study trial was followed by another study trial with the same list, a study trial with a new list, or a test trial. During the test trials, the subjects were given 2 min to recall as many words out loud as they could remember from the most recently studied list in any order in which the words came to mind. The computer recorded all verbal responses with the aid of a microphone. In addition, the subjects solved arithmetic problems for 15 sec between trials. PyEPL software (Geller, Schleifer, Sederberg, Jacobs, & Kahana, 2007) was used for stimulus presentation and recording the subjects' verbal and keyboard responses. This first session lasted 1 h.

Following a 2-day retention interval, the subjects returned to the lab and were given four consecutive tests. In the first three tests, the subjects had 5 min to recall words from each of the three lists, separately, in any order in which the words came to mind. To aid list discrimination during recall, the experimenter tested lists in the order in which they had been presented and reminded the subjects of the text colors and positioning on the screen of the words from the separate lists. For the final test, the subjects were given 10 min to recall words from all three lists; however, in contrast to the previous trials, the experimenter provided a list of all of the category names. The second session lasted 30 min.

Results

All results, unless otherwise stated, were significant at the .05 level. For all sets of individual comparisons, we controlled the Type I error rate using the false discovery rate procedure (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001).

Recall of words. The first row of Table 1 shows that during the initial learning phase, the mean proportion of words recalled increased with repeated testing in both the two-test (.31 to .52) [$t(35) = 9.46$, $SEM = 0.02$, $d = 1.39$] and the four-test (.33 to .63) [$t(35) = 8.05$, $SEM = 0.04$, $d = 1.66$] conditions. The small differences in test performance between the four-test and two-test conditions on their corresponding tests were not significant (.33 vs. .31) ($t < 1$) and (.56 vs. .52) [$t(35) = 1.46$, $SEM = 0.02$, $d = 0.24$, n.s.]. As Tulving (1962) showed, test trials seem to fully substitute for study trials in free recall (see also Karpicke & Roediger, 2007).

Delayed recall performance can be examined in two alternative ways: as a function of either study trials or test trials. Under the usual assumption that learning occurs primarily during study episodes, one might expect recall to increase on a delayed test as the number of study trials increased from four to six to eight. However, the top panel of Figure 1 shows that recall 2 days after learning actually declined in performance as a function of the number of study trials. The decline occurred in both free recall and cued recall. Of course, the number of test trials covaried with the number of study trials, and replotting the data in the top panel of Figure 1 as a function of the number of test trials (bottom panel of Figure 1) shows that testing exerted a powerful influence in both free and cued recall. This outcome represents the standard testing effect: Retrieval practice during tests (even when not all items are recalled; see Table 1) often has a much more powerful influence on later retention than does repeated study (of 100% of the material).

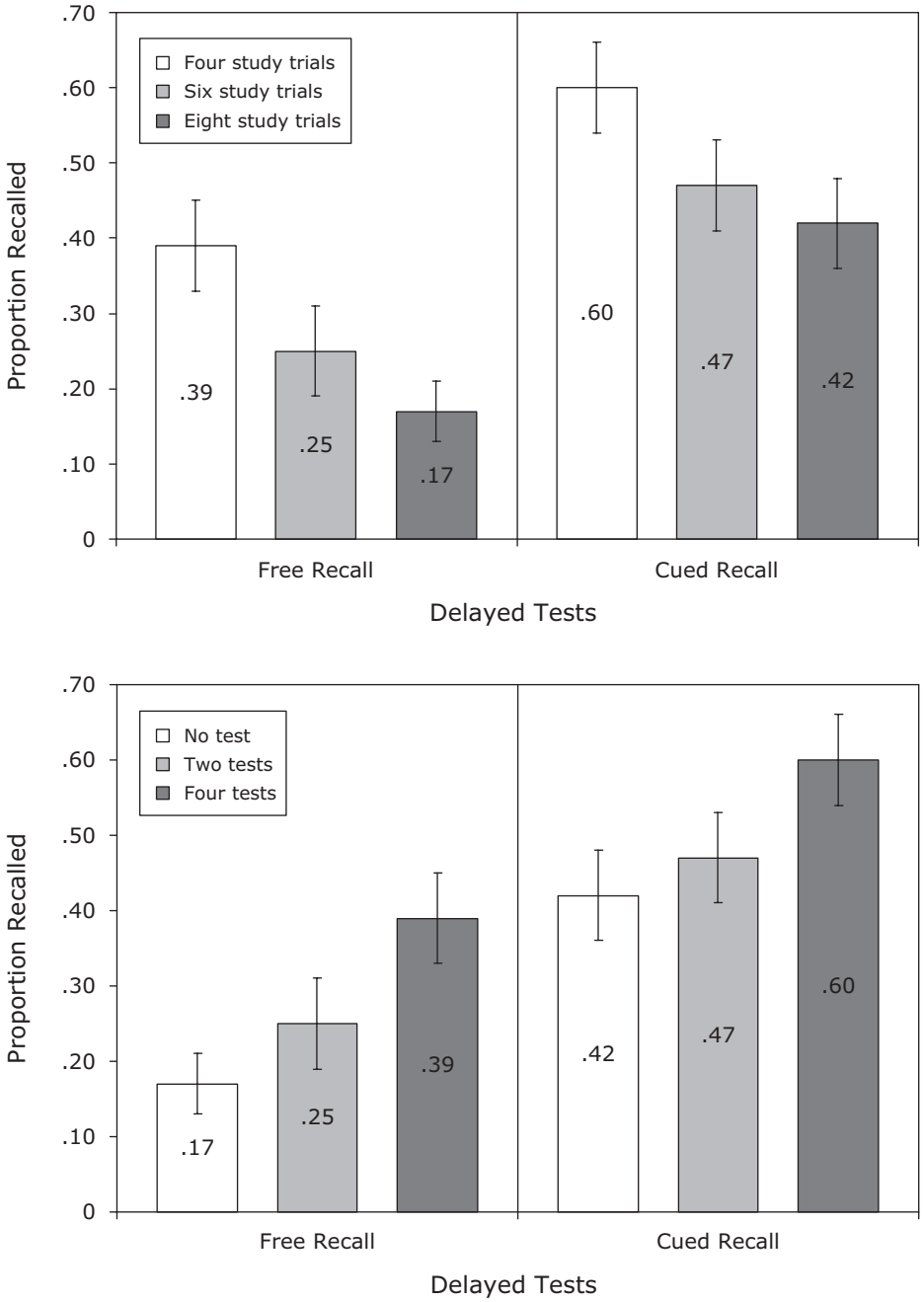

Figure 1. Mean proportion of words recalled in delayed tests of free and cued recall as a function of the number of study trials (top panel) and test trials (bottom panel) given in the initial learning phase in Experiment 1. Error bars represent 95% confidence intervals.

We conducted an ANOVA on the data from delayed free recall, which confirmed a significant effect of learning condition [$F(2,70) = 26.60$, $MS_e = 0.02$, $\eta_p^2 = .43$]. Individual pairwise comparisons revealed that, relative to the study-only condition, recall was superior in the two-test (.25 vs. .17) [$t(35) = 3.17$, $SEM = 0.03$, $d = 0.50$] and four-test (.39 vs. .17) [$t(35) = 7.13$, $SEM = 0.03$, $d = 1.32$] conditions, and taking four tests enhanced recall to a greater extent than did taking only two tests (.39 vs. .25) [$t(35) = 4.03$, $SEM = 0.04$, $d = 0.76$]. In general, the pattern of results for cued recall was the same as that for free recall, and similar patterns of statistical significance obtained for these and the subsequent analyses in both Experiments 1 and 2 (see Zaromb, 2010, for these analyses). Of course, cued recall followed free recall, so the parallel trends may be carryover effects from free recall. Nevertheless, long-term free recall was superior in the repeated testing conditions, relative to the repeated studying condition, and was further enhanced by increasing the number of test trials during the initial learning phase.

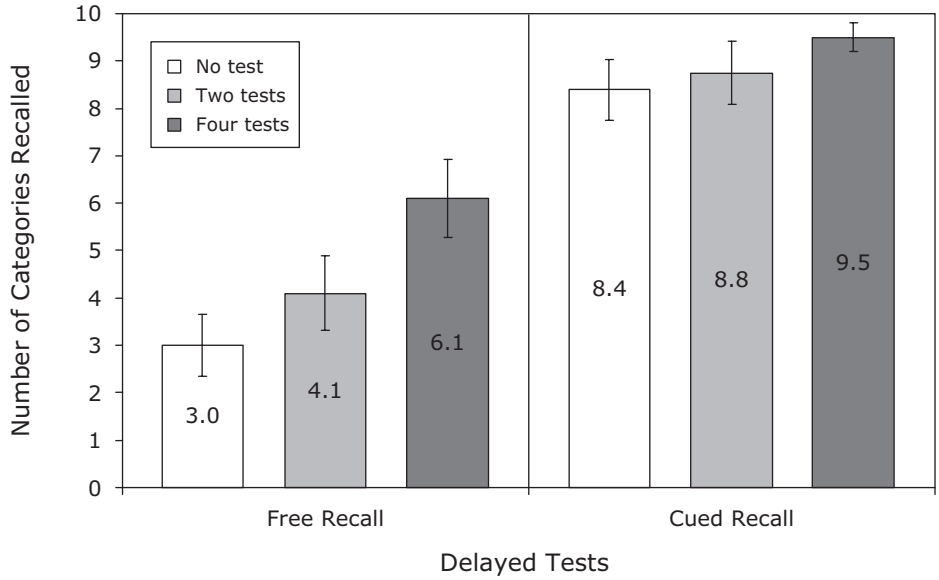

Figure 2. Mean number of categories recalled as a function of the number of tests given during the learning phase in delayed tests of free and cued recall in Experiment 1. Error bars represent 95% confidence intervals.

Recall of categories. The second row of Table 1 shows the mean number of categories in which at least one instance was recalled during initial tests of free recall. Rc increased during the learning phase with repeated testing in both the two-test (6.58 to 8.25) [$t(35) = 5.69$, $SEM = 0.02$, $d = 0.95$] and the four-test (6.61 to 8.83) [$t(35) = 6.16$, $SEM = 0.36$, $d = 1.22$] conditions. The differences in Rc between the four-test and two-test conditions on corresponding tests were not significant (6.61 vs. 6.58 and 8.78 vs. 8.25, respectively) [$t < 1$ and $t(35) = 1.92$, $SEM = 0.27$, $d = 0.42$, n.s., respectively].

Figure 2 shows that Rc increased as a function of the number of test trials given during the learning phase in delayed tests of free and cued recall. An ANOVA confirmed a significant effect of learning condition in free recall [$F(2,70) = 28.77$, $MS_e = 3.11$, $\eta_p^2 = .45$]. Individual comparisons revealed that, relative to that in the study-only condition, Rc in free recall was superior in the two-test (4.08 vs. 2.97) [$t(35) = 2.81$, $SEM = 0.40$, $d = 0.53$] and four-test (6.08 vs. 2.97) [$t(35) = 8.06$, $SEM = 0.39$, $d = 1.47$] conditions, and taking four tests improved Rc to a greater extent than did taking only two tests (6.08 vs. 4.08) [$t(35) = 4.34$, $SEM = 0.46$, $d = 0.86$]. In summary, the repeated testing conditions improved Rc, relative to the repeated study condition, in delayed free recall, and Rc was further enhanced by increasing the number of test trials during the initial learning phase.

Recall of items within categories. The third row of Table 1 shows the mean number of instances recalled within each taxonomic category (Rw/c) accessed during initial tests of free recall. During the learning phase, Rw/c increased across tests in both the two-test (2.34 to 3.13) [$t(35) = 5.74$, $SEM = 0.14$, $d = 1.20$] and four-test (2.45 to 3.48) [$t(35) = 6.29$, $SEM = 0.16$, $d = 1.37$] conditions. However, the differences in Rw/c between the four-test and two-test conditions on corresponding tests were not significant (2.45 vs. 2.34 and 3.14 vs. 3.13, respectively) (ts < 1).

Figure 3 shows that on the 2-day delayed tests of free and cued recall, Rw/c increased as a function of the number of test trials given during the learning phase. There was a significant effect of learning condition in Rw/c recall [$F(2,58) = 6.88$, $MS_e = 0.39$, $\eta_p^2 = .19$]. Individual comparisons revealed that, relative to that in the study-only condition, Rw/c was enhanced in the four-test condition (3.19 vs. 2.65) [$t(35) = 3.60$, $SEM = 0.17$, $d = 0.72$] and in the two-test condition, although the latter difference did not reach the conventional level of statistical significance (3.02 vs. 2.65) [$t(35) = 2.46$, $p = .06$, $SEM = 0.15$, $d = 0.41$]. Rw/c did not significantly differ between the four-test and two-test conditions (3.19 vs. 3.02) [$t(35) = 1.89$, $SEM = 0.14$, $d = 0.33$, n.s.]. In summary, repeated testing improved Rw/c in delayed free recall, as compared with repeated studying alone.

Category clustering. Category clustering was measured by computing ARC (Roenker et al., 1971) scores for each subject's output protocol in initial and delayed tests of free recall. There was a high degree of category clustering during both the initial learning phase and the final free recall phase, with ARC scores ranging from .66 to .84. As is shown in the bottom row of Table 1, for tests that occurred during the learning phase, ARC scores increased in the four-test condition (.68 to .82) [$t(35) = 3.08$, $SEM = 0.04$, $d = 0.71$] and in the two-test condition (.66 to .73), although the latter improvement was not statistically significant [$t(35) = 1.34$, $SEM = 0.05$, $d = 0.27$, n.s.], probably due to high variability among the subjects. The differences in ARC scores between the four-test and two-test conditions on corresponding tests that occurred during the first (.68 vs. .66) and third (.79

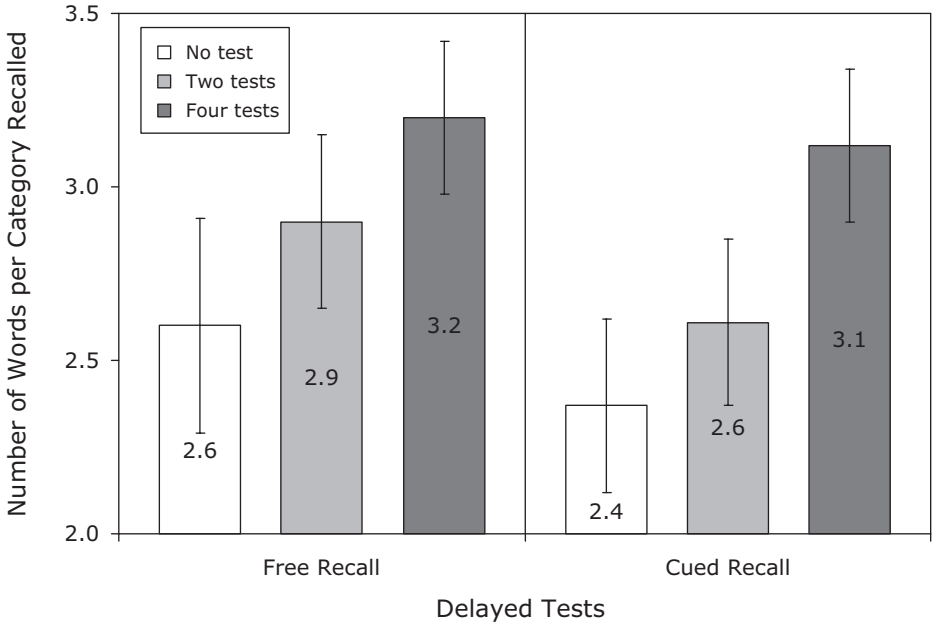

Figure 3. Mean number of words per category recalled as a function of the number of tests given during the learning phases in delayed tests of free and cued recall in Experiment 1. Error bars represent 95% confidence intervals.

vs. .73) trial blocks were not significant either [$t < 1$ and $t(35) = 1.58$, $SEM = 0.04$, $d = 0.27$, n.s., respectively]. Moreover, whereas ARC scores and recall performance were uncorrelated on the initial tests (Pearson's $r = .00$ in both the two-test and the four-test conditions), they were correlated on the final tests ($r = .56$ and .55 for the two-test and four-test conditions, respectively).

The ARC scores remained high 2 days later across all three learning conditions, with the greatest degree of category clustering in the two-test condition ($M = .84$, $SD = .21$), followed by the four-test condition ($M = .80$, $SD = .21$), and the poorest in the study-only condition ($M = .72$, $SD = .31$). This outcome is a surprise, because previous work has shown that repeated testing can enhance clustering after long delays, relative to taking a single, time-matched test (Mulligan, 2005). Therefore, one would expect the greatest clustering with more tests, if testing enhances clustering. An ANOVA revealed, however, that the differences among the learning conditions were not statistically significant [$F(2,48) = 2.34$, $MS_e = 0.05$, $\eta_p^2 = .09$, n.s.]. Moreover, ARC scores were also uncorrelated with recall performance across all three conditions (all r values $< .21$, n.s.). Although repeated testing in the four-test condition demonstrated significant improvements in output organization during initial learning, repeated testing and the study-only conditions produced similarly high degrees of output organization in delayed recall, and organization was uncorrelated with recall. However, this finding of no difference among conditions may be due to a ceiling effect in clustering for the two-test and four-test conditions (.84 and .80, respectively). In fact, over 28% of the subjects in each of the repeated studying and testing conditions demonstrated perfect recall organization (ARC scores of 1.00), which indicates that the present learning conditions allowed many of the subjects to master the conceptual organization of the study materials.

Subjective organization. Because categorical organization was near ceiling, we examined subjective organization to try to discern differences among conditions. Subjective organization was measured using PF (Sternberg & Tulving, 1977). Again, PF represents the number of pairs of items commonly recalled on adjacent test trials in adjacent output positions in either forward or reverse order. Figure 4 shows PF scores for initial and delayed free recall trials in the two-test and four-test conditions. Because a minimum of two recall trials are required to compute PF, it was not possible to measure subjective organization in the study-only condition.

During the initial learning phase, PF scores significantly increased across recall trials in the four-test condition (2.00 to 5.67) [$t(35) = 4.05$, $SEM = 0.57$, $d = 0.99$]. Although PF scores did not significantly differ among initial pairs of tests in the four-test and two-test conditions (2.00 vs. 1.81) ($t < 1$), 2 days later the PF scores measured between the final test trial during the learning phase and the delayed test of free recall were significantly higher in the four-test condition (4.31 vs. 2.20) [$t(33) = 3.67$, $SEM = 0.63$, $d = 0.72$]. In addition, the final PF scores were highly correlated with delayed recall performance ($r = .68$ and .83 in the two-test and four-test conditions, respectively). Therefore, increasing the number of tests led to significant increases in subjective organization across initial and delayed recall tests. The PF measure captured a form of organization significantly correlated with delayed recall that the ARC measure did not, despite the fact that we used categorized lists. This outcome supports the

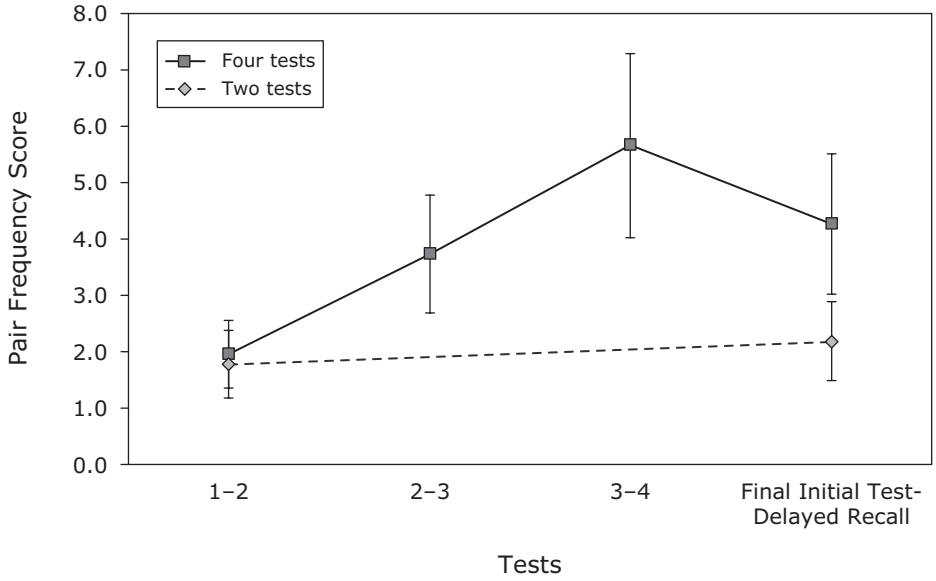

Figure 4. Pair frequency scores as a function of the number of tests given during the learning phase in initial and delayed tests of free recall in Experiment 1. Error bars represent 95% confidence intervals.

hypothesis that even though category clustering was near ceiling for subjects in all three conditions, differences in later recall were correlated with consistent responding in recall of items within and across categories, measured by PF. Enhanced organization may be responsible for the testing effect in free recall.

Intrusions. We further examined recall accuracy by measuring the proportion of all words recalled during final free recall that were words from other study lists (interlist intrusions) or words not presented during the course of the experiment (extralist intrusions). Repeated testing reduced both inter- and extralist intrusions, as compared with the repeated studying condition. The highest proportion of interlist intrusions was committed in the study-only condition ($M = .34$, $SD = .25$), followed by the two-test ($M = .23$, $SD = .22$) and four-test ($M = .16$, $SD = .15$) conditions. Although the subjects recalled fewer extralist intrusions, the pattern was similar. The highest proportion of extralist intrusions occurred in the study-only condition ($M = .12$, $SD = .12$), followed by the two-test ($M = .08$, $SD = .08$) and four-test ($M = .08$, $SD = .08$) conditions.

We conducted a 2 (intrusion type: interlist vs. extralist) $\times$ 3 (learning condition: study only vs. two test vs. four test) ANOVA, which revealed a significant effect of intrusion type, with a higher overall rate of interlist intrusions (.24 vs. .08) [$F(1,32) = 48.31$, $MS_e = 0.03$, $\eta_p^2 = .60$]. There was a significant effect of learning condition [$F(2,64) = 8.99$, $MS_e = 0.25$, $\eta_p^2 = .22$] and an interaction between the two factors [$F(2,64) = 3.49$, $MS_e = 0.02$, $\eta_p^2 = .10$]. These effects were due to a higher proportion of interlist intrusions committed in the study-only than in the four-test condition (.34 vs. .16) [$t(34) = 4.48$, $SEM = 0.04$, $d = 0.87$]. However, neither the differences between the study-only and two-test conditions (.34 vs. .23) [$t(32) = 1.82$, $SEM = 0.06$, $d = 0.47$, n.s.] nor those between the two-test and four-test conditions (.23 vs. .16) [$t(33) = 1.64$, $SEM = 0.04$, $d = 0.37$, n.s.] were significant.

Similarly, there was a higher proportion of extralist intrusions in the study-only condition than in the four-test condition (.12 vs. .06) [$t(34) = 3.54$, $SEM = 0.02$, $d = 0.63$]. Neither the differences between the study-only and two-test conditions (.12 vs. .08) [$t(32) = 2.05$, $SEM = 0.02$, $d = 0.38$, n.s.] nor those between the two-test and four-test conditions ($t < 1$) were significant. In summary, repeated testing reduced false recall of both inter- and extralist intrusions, relative to repeated studying alone.

Discussion

This experiment confirmed a powerful effect of repeated testing (relative to repeated studying) on delayed retention tests. Studying the list six times and taking two tests produced greater recall 2 days later than did the condition in which the subjects studied the list eight times. Furthermore, studying the list only four times, while taking four tests, produced better retention than in either of the other two conditions. Keep in mind that if sheer exposure were the primary factor determining performance, the repeated study condition should have greatly outperformed the other two. When the subjects were given tests in the initial learning phase, they only recalled (on average) about 50% of the items, whereas the subjects in the repeated study condition were, of course, reexposed to 100% of the items on each study trial. In addition, repeated testing improved overall accuracy by minimizing false recall of both inter and extralist intrusions, relative to repeated studying alone. These results both replicate and extend previous findings that testing reduces the commission of prior-list

intrusions in free recall (Szpunar et al., 2008). Taken together, these findings provide further striking evidence for the power of testing (Roediger & Karpicke, 2006a).

One purpose of this experiment was to determine what components of recall were improved by testing, relative to studying alone—access to higher order units, access to items within units, or both. The last option was confirmed, because testing benefited both measures of category access (Rc) and recall of items within each accessed category (Rw/c) in delayed recall. These results are surprising, because many prior studies have shown that these two factors contribute independently to recall. That is, variables that influence Rc usually have no influence on Rw/c, and vice versa (e.g., Burns & Brown, 2000; Cohen, 1963, 1966; Hunt & Seta, 1984; Tulving & Pearlstone, 1966).

In addition, we asked whether testing enhances recall organization, relative to studying alone. In terms of category clustering, the answer appears to be "no," because the amount of category clustering in final free recall was similarly high across all learning conditions. Although repeated studying led to poor retention, relative to the conditions that included testing, the subjects clustered recall in terms of categories even in this condition. However, when we used the more subtle pair frequency measure of subjective organization, we found significant differences among conditions, supporting the claim that testing enhances organization.

Subjective organization (PF) increased as a function of the number of recall tests performed during the learning phase, with response patterns more consistent with four than with two tests (in both initial and delayed recall). Of course, consistent with prior research (e.g., Klein, Loftus, Kihlstrom, & Aseron, 1989; Mulligan, 2001; Mulligan & Duke, 2002), category clustering also increased over test trials during the initial learning phase. However, the main difference between the measures was in delayed recall. In that case, the PF scores were greater when the subjects had four tests than when they had two tests during initial learning (unlike the outcome with category clustering as indexed by ARC scores). In addition, PF measures were also correlated with the subjects' levels of recall on the final free recall test, unlike the clustering measures. Therefore, the benefits of testing in free recall may be due, at least in part, to processes that contribute to subjective organization.

Nevertheless, aside from the possible ceiling effect, it is still unclear why testing had no effect on category clustering, since some prior studies have shown that testing does have a positive effect on clustering in delayed free recall (Masson & McDaniel, 1981; Mulligan, 2005). For instance, Masson and McDaniel (in their Experiment 1) presented subjects with a list of words representing several taxonomic categories and gave either intentional or incidental learning instructions and different encoding tasks for the study of individual words (the subjects wrote down a semantic or a phonological associate of each list item). Half of the subjects were given a free recall test immediately following the initial study period, and all of the subjects were given delayed recall and recognition tests a day later. They found that the subjects who were initially tested on the word list produced higher ARC scores in delayed recall than the subjects who did not receive a test during the first session.

Several differences exist between the prior and present research that may explain the divergent findings. First, Experiment 1 provided the subjects with considerably more opportunities to study the taxonomic organization of the lists (between four and eight study trials, as compared with only one study trial in Experiment 1 of Masson & McDaniel [1981]). As a result, ARC scores in delayed recall were much higher in the present experiment, in which they ranged from .72 to .84, than those reported by Masson and McDaniel, in which they ranged from .11 to .47. Second, whereas Masson and McDaniel used a between-subjects experimental design, in the present experiment we used a within-subjects design, which provided the subjects with a total of 24 study and test trials to learn the taxonomic organization of three separate lists. Perhaps decreasing the number of learning trials and using a between-subjects design would permit differentiating the effects of studying and testing on category clustering.[1]

Masson and McDaniel (1981, Experiment 1) also did not equate the number of study and test trials across the learning conditions. Perhaps the organization scores were higher for the prior testing condition because the subjects had an additional opportunity to learn the material and because a second study trial during the first session would have been just as effective as the recall test in promoting additional processing of organizational information among list items. Finally, Masson and McDaniel used encoding tasks that may have promoted greater processing of semantic and/or phonological features unique to each word (item-specific processing) while minimizing the processing of interitem relational information. Output organization might have been greater had the subjects been given the opportunity to study the list items as they saw fit under standard intentional learning conditions, in which case they might have been more likely to process interitem semantic relations. We examined these possible explanations of differences between our results and those of Masson and McDaniel in Experiment 2 by using an experimental design similar to that of Masson and McDaniel (in their Experiment 1), but with some changes to address the issues noted above.

EXPERIMENT 2

The purpose of Experiment 2 was to further examine the effects of testing on the learning and retention of lists of words representing different taxonomic categories. Of interest was whether the retrieval processes that occur during a recall test stimulate organizational processing to a greater extent than does a study trial of equal duration. Using an experimental design adapted from Masson and McDaniel (1981), we compared delayed recall performance, measured by total word recall, category recall (Rc), and words per category recall (Rw/c), and organization, measured by clustering (ARC), for subjects who

received one study trial followed by an immediate recall test with those for groups that received one or two study trials alone. All groups were given delayed tests of free and category cued recall 24 h later.

In one condition, a group of subjects studied several lists of words for one study trial, with instructions to rate the pleasantness of each word. A second group studied each list once, with intentional learning instructions to learn each word as well as possible during list presentation. A third group rated the pleasantness of each word during an initial study trial and then studied each list a second time under intentional learning instructions. Finally, a fourth group initially studied each list of words with instructions to make pleasantness judgments and then attempted to recall each list immediately following list presentation.

The logic underlying these comparisons is as follows. The comparison of the pleasantness-rating study phase by itself and the same kind of study phase plus an initial test conceptually replicates the design of Masson and McDaniel (1981, Experiment 1). The condition with two study conditions (pleasantness rating and intentional learning) equates exposure to that in the study + test condition. The addition of the single intentional study control condition made it possible to ask what effect studying under intentional learning has on later performance and permits comparison with the pleasantness-rating single-study condition. A day later, the subjects in all four conditions took final tests of free and category cued recall.

Method

Subjects. Sixty-four Washington University undergraduates participated for either payment or course credit.

Design. There were four learning conditions distributed among the subjects. In the S_p condition, 16 subjects studied three lists of words only once with instructions to rate the pleasantness of each list item on a 5-point scale. In the S_i condition, 16 subjects studied all three lists of words only once, with intentional learning instructions to learn each of the list items as well as possible during list presentation. In the S_pS_i condition, another group of 16 subjects rated the pleasantness of each list item during an initial study trial and then studied the list a second time, with standard intentional learning instructions. Finally, in the testing (S_pT) condition, 16 subjects first studied the list of words with instructions to make pleasantness judgments for each item and then attempted to recall the list immediately afterward. Words were presented in a different randomized order on each study trial in the condition that involved two study trials. The critical tests took place a day later, when the subjects in all four conditions attempted to recall the word lists using tests of free and category cued recall.

Materials. Ninety words sampled from 18 categories (5 words per category) in the expanded and updated version of the Battig and Montague word norms (Van Overschelde et al., 2004) were used to create three 30-word study lists. The 30 words in each list included five medium frequency nouns belonging to each of six taxonomic categories.

Procedure. The subjects participated in two sessions scheduled 1 day apart. In the first session, the subjects were informed that they would study several lists of words presented by a computer in preparation for a memory test the next day. During the study trials, the computer displayed each word in the center of the monitor display one at a time for 4.5 sec, followed by a 500-msec interstimulus interval. The words were presented in randomized order on each study trial. For the S_p study trials, the subjects were informed that they had 5 sec during the presentation of each word to type a number between 1 and 5 indicating their pleasantness judgment for the current item. For the S_i study trials, the subjects were instructed only to learn each word as well as possible as it was presented. The total time for each study trial was 2.5 min.

During the test trial in the S_pT condition, the subjects were given 2.5 min to write down on a blank sheet of paper as many words as they could remember from the most recently studied list in any order in which the words come to mind. In order to keep the spacing between each of the three study lists constant across the four learning conditions, the subjects in the S_p and S_i conditions played Tetris for an additional 2.5 min in between study trials. E-Prime experimental software (Psychology Software Tools, Sharpsburg, PA) was used for stimulus presentation and recording the subjects' keyboard responses. The first session lasted about 30 min.

Following a 1-day retention interval, the subjects were given tests of final free and cued recall. During the free recall test, the subjects had 10 min to write down on a blank sheet of paper as many words as they could remember from all three lists in any order in which the words came to mind. Finally, the subjects had 10 min to recall words from all three lists; however, in contrast to the previous test, the subjects were also provided with a list of all of the category names to aid recall of the words. The second session lasted 20 min.

Results

We report analyses only for the delayed tests of free and cued recall, because unlike in Experiment 1, only one learning condition (S_pT) included tests during the initial learning phase, and it was only possible to compare recall performance and organization across all conditions in the delayed tests. On the initial test trial, the subjects in the SpT condition recalled, on average, 68% ($SD = 0.12$) of the words from 5.48 ($SD = 0.36$) categories (Rc) and 3.71 ($SD = 0.56$) items per category (Rw/c) of each 30-item list. Recall was also highly organized, as was indicated by a mean ARC score of .79 ($SD = .12$).

Recall of words. The top row of Table 2 shows that testing during the initial learning phase improved recall performance in delayed tests of free and cued recall. There was a significant effect of learning condition in free recall [$F(3,60) = 22.19$, $MS_e = 0.01$, $\eta_p^2 = .53$], which was due to enhanced recall in the prior testing condition (S_pT), relative to the S_p (.45 vs. .19) [$t(30) = 6.48$, $SEM = 0.04$, $d = 2.35$], S_i (.45 vs. .18) [$t(30) = 7.84$, $SEM = 0.03$, $d = 2.84$], and S_pS_i (.45 vs. .21) [$t(30) = 5.99$, $SEM = 0.04$, $d = 2.17$] conditions. No other comparisons among the study-only conditions were statistically significant. Thus, testing improved long-term free recall, relative to studying alone, and neither varying the encoding instructions (pleasantness ratings vs. standard intentional learning) nor varying the number of study opportunities (one vs. two study trials) affected delayed recall performance.

Recall of categories. The second row of Table 2 shows that testing during the initial learning phase improved Rc in delayed tests of free and cued recall. There was a significant effect of learning condition on free recall [$F(3,60) = 11.49$, $MS_e = 7.33$, $\eta_p^2 = .37$], which was due to enhanced Rc in the prior testing condition (S_pT), relative to the S_p (12.56 vs. 8.31) [$t(30) = 4.64$, $SEM = 0.92$, $d = 1.64$], S_i (12.56 vs. 7.56) [$t(30) = 6.01$, $SEM = 0.83$, $d = 2.13$], and S_pS_i (12.56 vs. 8.19) [$t(30) = 5.80$, $SEM = 0.75$, $d =$

Table 2
Mean Proportion of Words Recalled, Number of Categories Recalled (Rc), Number of Words per Category Recalled (Rw/c), and Adjusted Ratio of Clustering (ARC) Scores As a Function of Initial Learning Condition in Delayed Tests of Free and Cued Recall in Experiment 2

	Free Recall								Cued Recall							
	S_p		S_i		S_pS_i		S_pT		S_p		S_i		S_pS_i		S_pT	
Measure	*M*	CI	*M*	CI	*M*	CI	*M*	CI	*M*	CI	*M*	CI	*M*	CI	*M*	CI
Recall	.19	.06	.18	.04	.21	.06	.45	.06	.34	.05	.29	.06	.37	.06	.61	.06
Rc	8.31	1.68	7.56	1.50	8.19	1.32	12.56	0.74	14.69	1.23	13.06	1.70	15.69	1.09	17.25	0.67
Rw/c	1.99	0.23	2.04	0.25	2.16	0.35	3.17	0.28	2.07	0.22	1.93	0.18	2.09	0.26	3.17	0.27
ARC	.60	.20	.48	.17	.60	.17	.85	.04								

Note—CI, 95% confidence intervals; S_p, study with pleasantness ratings; S_i, study with intentional learning instructions; S_pS_i, repeated study with pleasantness ratings on the first trial and intentional learning instructions on the second trial; S_pT, study with pleasantness ratings followed by a recall test.

2.05] conditions. No other comparisons were statistically significant. In summary, testing during the initial learning phase improved Rc, relative to studying alone, and neither varying the encoding instructions nor varying the number of study trials affected category recall.

Recall of items within categories. As is shown in the third row of Table 2, testing during the initial learning phase improved Rw/c in delayed tests of free and cued recall. There was a significant effect of learning condition on free recall [$F(3,60) = 15.74$, $MS_e = 0.32$, $\eta_p^2 = .44$], which was due to enhanced Rw/c in the prior testing condition (S_pT), relative to the S_p (3.17 vs. 1.99) [$t(30) = 6.53$, $SEM = 0.18$, $d = 2.30$], S_i (3.17 vs. 2.04) [$t(30) = 6.04$, $SEM = 0.19$, $d = 2.13$], and S_pS_i (3.17 vs. 2.09) [$t(30) = 4.49$, $SEM = 0.22$, $d = 1.59$] conditions. No other comparisons among the study-only conditions were significant. In summary, long-term free recall of words within categories was superior in the prior testing condition, relative to that in the study-only conditions, and neither varying the encoding instructions nor varying the number of study trials affected Rw/c.

Category clustering. As is shown in the bottom row of Table 2, testing during the initial learning phase improved category clustering in delayed free recall. An ANOVA confirmed a significant effect of learning condition on category clustering [$F(3,58) = 3.93$, $MS_e = 0.10$, $\eta_p^2 = .16$], which was due to enhanced ARC scores in the prior testing condition (S_pT), relative to the S_p (.85 vs. .60) [$t(30) = 2.50$, $SEM = 0.10$, $d = 0.87$], S_i (.85 vs. .48) [$t(29) = 4.41$, $SEM = 0.08$, $d = 1.59$], and S_pS_i (.85 vs. .61) [$t(29) = 2.78$, $SEM = 0.09$, $d = 0.97$] conditions. No other comparisons among the study-only conditions were significant. In addition, ARC scores were positively correlated with delayed recall ($r = .51$). In contrast to the results of Experiment 1, testing improved the organization of recall, and organization was correlated with the number of words recalled. Furthermore, neither varying the encoding instructions nor varying the number of study trials affected output organization.

Intrusions. We further examined recall accuracy by measuring the proportion of all words recalled in delayed free recall that were words not presented during the course of the experiment (extralist intrusions). Testing during the learning phase reduced false recall on the delayed test. The highest proportion of extralist intrusions was committed in the Si condition ($M = .36$, $SD = .25$), followed by the S_p ($M = .23$, $SD = .21$) and S_pS_i ($M = .21$, $SD = .22$) conditions. The lowest proportion of extralist intrusions occurred in the prior testing (S_pT) condition ($M = .06$, $SD = .07$).

Critically, there was a significant effect of learning condition in recall of intrusions [$F(3,60) = 6.04$, $MS_e = 0.04$, $\eta_p^2 = .23$], which was due to a lower proportion of extralist intrusions committed in the prior testing condition (S_pT), relative to the S_p (.06 vs. .23) [$t(30) = 3.07$, $SEM = 0.06$, $d = 1.09$], S_i (.06 vs. .36) [$t(30) = 4.64$, $SEM = 0.07$, $d = 1.63$], and S_pS_i (.06 vs. .21) [$t(30) = 2.71$, $SEM = 0.06$, $d = 0.92$] conditions. No other comparison among the study-only conditions was significant. As in Experiment 1, testing during the initial learning phase reduced false recall, relative to studying alone, following a long delay.

Discussion

Experiment 2 confirmed several positive effects of testing on long-term retention and organization, relative to studying alone. Consistent with the results of Experiment 1 and prior research, studying a list and taking an immediate recall test produced greater veridical recall and reduced false recall a day later, relative to conditions in which the subjects studied a list only one or two times (Masson & McDaniel, 1981; Szpunar et al., 2008). Somewhat surprisingly, neither varying the conditions of encoding nor increasing the number of study trials affected recall after 24 h. Although it is reasonable to expect that repeatedly studying information should improve recall, relative to a single study opportunity, repetition does not always boost retention (e.g., Callender & McDaniel, 2009), especially after long delays (Karpicke & Roediger, 2008).

The main findings of Experiment 2 are that testing benefited measures of category access (Rc), recall of items within each accessed category (Rw/c), and organization of recall (ARC), relative to learning conditions of studying alone. Our results confirm that testing can improve organization of recall—or category clustering—in delayed free recall, relative to restudying material (Masson & McDaniel, 1981). That organization was positively correlated with delayed recall further suggests that the testing effect

in free recall may be due in part to enhanced organizational processing.

GENERAL DISCUSSION

Two experiments confirmed the positive effects of testing in enhancing long-term retention, relative to restudying lists of categorized words, and showed that testing enhances organization during recall. In Experiment 1, total recall of words, category access (Rc), recall of words within categories (Rw/c) and one measure of organization (PF) all increased as the number of tests increased from none to two to four (while holding total exposure constant). In Experiment 2, studying a list and taking an immediate recall test produced greater recall a day later than did conditions in which the subjects studied the list alone, and testing once again improved Rc, Rw/c, and organization, as measured by category clustering (ARC). Furthermore, testing improved memory accuracy by reducing the false recall of interlist (Experiment 1) and extralist (Experiments 1 and 2) intrusions. Taken together, these findings provide further striking evidence for the power of testing (Roediger & Karpicke, 2006a) and help to provide understanding of why testing effects occur, at least in free recall.

The main purpose of these experiments was to investigate whether the benefits of testing extended to individuals' learning of conceptual organization, relative to studying alone—a question that had not yet been addressed in the literature. First, we asked what components of recall were improved by testing, relative to studying alone—access to higher order units, access to items within units, or both. In both experiments, the last option was confirmed, because testing benefited measures of both category access (Rc) and recall of items within each accessed category (Rw/c) in delayed tests of free and cued recall.

If individuals learn categorized word lists by chunking items into category-based units, once they can access the units during retrieval, their contents (the individual items) will be accessed as well, to some degree. In their classic work supporting the distinction between item availability and accessibility, Tulving and Pearlstone (1966) showed that Rc and Rw/c were largely independent of each other, because variables that affected Rc (such as category cuing and list length) had little influence on Rw/c. Hunt and Seta (1984) argued that Rc and Rw/c measure the extent to which relational and item-specific information, respectively, is used to guide episodic retrieval. Although Rc measures the extent to which individuals can retrieve higher order units or chunks, Rw/c reflects the degree to which individuals can retrieve category members.

Indeed, experimental conditions designed to promote organizational processing (e.g., instructing subjects to organize study items, providing category names during study) have been found to selectively increase Rc, and those designed to enhance item-specific processing (e.g., generating study items) have been shown to increase Rw/c (e.g., Cohen, 1963, 1966; McDaniel, Waddill, & Einstein, 1988; Schmidt & Cherry, 1989). To the extent to which these measures assess the extent to which relational (Rc) and item-specific (Rw/c) information is used to guide episodic retrieval (e.g., Hunt & Seta, 1984), our findings show that testing may promote both relational and item-specific processing, relative to studying alone.

Karpicke and Zaromb (2010) recently found that testing enhances memory for previously read list items, relative to passively rereading or actively generating the items, on final tests of recall and item recognition. They also showed that these effects are robust in both within- and between-subjects experimental designs (unlike the generation effect). They argued that testing may enhance item-specific processing that constrains retrieval to the set of list items to be remembered on a later test. Note that when the subjects in our experiments falsely recalled extralist intrusions, over 80% of these intrusions were other category exemplars, which suggests that testing may reduce false recall by constraining retrieval to the target category exemplars. Gallo and Roediger (2002) showed a similar effect in that recall testing of previously studied associate (DRM) lists reduced later false recognition. They argued that testing enhanced the recollective distinctiveness of list items, which, in turn, reduced false recognition on a later test (see also Brewer, Marsh, Meeks, Clark-Foos, & Hicks, 2010). Taken together, one might argue that it is the combination of these two types of processing—relational and item specific—that produces superior retention and underlies the positive effects of testing on long-term retention (Hunt, 2006; Matthews, Smith, Hunt, & Pivetta, 1999).

One criticism of interpreting Rc and Rw/c as measures of organizational and item-specific processing is that they do not adjust for differences in recall performance across individuals or learning conditions (Burns & Brown, 2000; Murphy, 1979). For instance, Burns and Brown argued for the use of the adjusted category access ratio (ACA) and adjusted items per category recalled ratio (AIPC) in conjunction with Rc and Rw/c, because these measures do correct for recall-level differences (see Burns & Brown, 2000, for details). ACA and AIPC scores of 0 indicate chance-level Rc and Rw/c scores, respectively, and scores above 0 indicate that Rc and Rw/c scores are greater than that expected by chance alone.

We applied Burns and Brown's (2000) measures to our data and obtained curious outcomes. In both experiments, access of categories (ACA, the corrected version of Rc) was consistently well below chance in final recall in both the nontested and the tested conditions. Furthermore, corrected access of items within categories (AIPC, the corrected version of Rw/c) was near chance levels in the nontested conditions and above chance in the tested condition, but only in Experiment 2.

These findings raise questions, one of which is the interpretation of *below-chance* clustering of categorized lists (but see Burns & Brown [2000] for a suggestion). This outcome gives one pause about the assumptions being used in the measure. If subjects obviously use organized recall (as was indicated both by near-ceiling category clustering and by above-chance PF scores in Ex-

periment 1), perhaps the estimate of chance is too high in these measures (hence, leading the data to appear to be below chance). Our preferred use of Rc and Rw/c measures is the same as that of Tulving and Pearlstone (1966) and many others—as descriptive measures: Total recall of categorized lists can be decomposed into two components that bear a multiplicative relationship (i.e., recall of words or Rw = Rc × Rw/c). The Rc and Rw/c measures are, by definition, components of overall recall and do not need to be corrected for descriptive purposes. On the other hand, future research may indeed show that Hunt and Seta's (1984) interpretation of Rc and Rw/c as reflecting relational and item-specific processing may be in need of reexamination, as Burns and Brown claimed.

A second question that we asked was whether testing improves recall organization, as measured by category clustering (ARC) and subjective organization (PF). In Experiment 1, repeated studying led to poor retention, relative to the conditions that included testing, but it was nearly as effective in producing highly organized recall. That a significant proportion of the subjects achieved perfect clustering suggests that ARC was near ceiling and therefore not sufficiently sensitive to potential underlying differences in organization between repeated studying and testing. In agreement with this hypothesis, we found that the PF measure of subjective organization was useful and revealed differences among conditions, even though we used categorized lists. Subjective organization (PF) increased as a function of the number of recall tests performed during the learning phase, with response patterns more consistent with four than with two tests (both in initial and delayed recall). In addition, PF measures were also correlated with the subjects' levels of recall on the final free recall test.

In Experiment 2, we used conditions that eliminated the ceiling effect on ARC scores that existed in Experiment 1 by using fewer study and test trials (and a between-subjects design, to minimize interlist interference). We obtained a testing effect with only a single test (replicating Masson & McDaniel, 1981). More importantly, testing did produce greater category clustering, and organization was correlated with recall. This finding provides additional evidence that organizational processes may contribute to the positive effects of testing on long-term retention.

In summary, the main findings from our experiments are that testing enhances three different measures of categorized list recall: access to higher order units (Rc), access to their contents (Rw/c), and recall organization of the lists (ARC and/or PF). We conclude that testing stimulates the development of both categorized knowledge (assessed by ARC) and personal idiosyncratic organization (measured by PF). Put another way, testing appears to permit subjects to develop retrieval plans (Slamecka, 1968) on the basis of both their categorical knowledge and recollection of previous recall attempts. These complementary retrieval schemas that arise through testing may be responsible for the testing effect obtained in delayed free recall. This conjecture about subjective organization and category clustering is new, however, and awaits further assessment.

AUTHOR NOTE

Support for this research was provided by a Collaborative Activity grant from the James S. McDonnell Foundation (220020041). Data from Experiment 2 were included as part of the first author's doctoral thesis under the direction of the second author. Thanks to Larry Jacoby, Mark McDaniel, and Kathleen McDermott for serving on the dissertation committee and for helpful suggestions in designing Experiment 2 and to Larissa D'Abreu for assistance with data collection and scoring. Correspondence concerning this article should be addressed to F. M. Zaromb, Center for Foundational and Validity Research, Educational Testing Service, 660 Rosedale Road, MS 16R, Princeton, NJ 08541 (e-mail: fzaromb@ets.org).

REFERENCES

AUSUBEL, D. P. (1963). *The psychology of meaningful verbal learning*. New York: Grune & Stratton.

BARTLETT, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

BENJAMINI, Y., & HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, **57**, 289-300.

BENJAMINI, Y., & YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165-1188.

BOUSFIELD, A. K., & BOUSFIELD, W. A. (1966). Measurement of clustering and sequential constancies in repeated free recall. *Psychological Reports*, **19**, 935-942.

BOUSFIELD, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology*, **49**, 229-240.

BOUSFIELD, W. A., COHEN, B. H., & WHITMARSH, G. A. (1958). Associative clustering in the recall of words of different taxonomic frequencies of occurrence. *Psychological Reports*, **4**, 39-44.

BOUSFIELD, W. A., PUFF, C. R., & COWAN, T. M. (1964). The development of constancies in sequential organization during repeated free recall. *Journal of Verbal Learning & Verbal Behavior*, **3**, 489-495.

BOWER, G. H., & SPRINGSTON, F. (1970). Pauses as recoding points in letter series. *Journal of Experimental Psychology*, **83**, 421-430.

BREWER, G. A., MARSH, R. L., MEEKS, J. T., CLARK-FOOS, A., & HICKS, J. L. (2010). The effects of free recall testing on subsequent source memory. *Memory*, **18**, 385-393.

BURNS, D. J., & BROWN, C. A. (2000). The category access measure of relational processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 1057-1062.

BUTLER, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **36**, 1118-1133.

BUTLER, A. C., & ROEDIGER, H. L., III (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, **19**, 514-527.

CALLENDER, A. A., & MCDANIEL, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology*, **34**, 30-41.

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

CHAN, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory & Language*, **61**, 153-170.

CHAN, J. C. K., MCDERMOTT, K. B., & ROEDIGER, H. L., III (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, **135**, 553-571.

COHEN, B. H. (1963). Recall of categorized word lists. *Journal of Experimental Psychology*, **66**, 227-234.

COHEN, B. H. (1966). Some or none characteristics of coding behavior. *Journal of Verbal Learning & Verbal Behavior*, **5**, 182-187.

DANIEL, D. B., & POOLE, D. A. (2009). Learning for life: An ecological approach to pedagogical research. *Perspectives on Psychological Science*, **4**, 91-96.

GALLO, D. A., & ROEDIGER, H. L., III (2002). Variability among word lists in eliciting memory illusions: Evidence for associative activation and monitoring. *Journal of Memory & Language*, **47**, 469-497.

GATES, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, **6**, 1-104.

GELLER, A. S., SCHLEIFER, I. K., SEDERBERG, P. B., JACOBS, J., & KAHANA, M. J. (2007). PyEPL: A cross-platform experiment-programming library. *Behavior Research Methods*, **39**, 950-958.

HUNT, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 3-25). New York: Oxford University Press.

HUNT, R. R., & MCDANIEL, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory & Language*, **32**, 421-445.

HUNT, R. R., & SETA, C. E. (1984). Category size effects in recall: The roles of relational and individual item information. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 454-464.

IZAWA, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology*, **83**, 340-344.

JOHNSON, C. I., & MAYER, R. E. (2009). The testing effect with multimedia learning. *Journal of Educational Psychology*, **101**, 621-629.

KAHANA, M. J., HOWARD, M. W., & POLYN, S. M. (2008). Associative processes in episodic memory. In J. Byrne (Series Ed.) & H. L. Roediger III (Vol. Ed.), *Learning and memory: A comprehensive reference. Vol. 2: Cognitive psychology of memory* (pp. 467-490). Amsterdam: Elsevier.

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. *European Journal of Cognitive Psychology*, **19**, 528-558.

KARPICKE, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, **138**, 469-486.

KARPICKE, J. D., & ROEDIGER, H. L., III (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory & Language*, **57**, 151-162.

KARPICKE, J. D., & ROEDIGER, H. L., III (2008). The critical importance of retrieval for learning. *Science*, **319**, 966-968.

KARPICKE, J. D., & ZAROMB, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory & Language*, **62**, 227-239.

KATONA, G. (1940). *Organizing and memorizing*. New York: Columbia University Press.

KLEIN, S. B., LOFTUS, J., KIHLSTROM, J. F., & ASERON, R. (1989). Effects of item-specific and relational information on hypermnesic recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 1192-1197.

MANDLER, G. (1967). Organization and memory. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 1, pp. 327-372). New York: Academic Press.

MASSON, M. E. J., & MCDANIEL, M. A. (1981). The roles of organizational processes in long-term retention. *Journal of Experimental Psychology: Human Learning & Memory*, **7**, 100-110.

MATTHEWS, T. D., SMITH, R. E., HUNT, R. R., & PIVETTA, C. E. (1999). Role of distinctive processing during retrieval. *Psychological Reports*, **84**, 904-916.

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., MOORE, B. A., & WHITEMAN, H. L. (1998). Dynamic changes in hypermnesia across early and late tests: A relational/item-specific account. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 173-185.

MCDANIEL, M. A., ROEDIGER, H. L., III, & MCDERMOTT, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, **14**, 200-206.

MCDANIEL, M. A., WADDILL, P. J., & EINSTEIN, G. O. (1988). A contextual account of the generation effect: A three-factor theory. *Journal of Memory & Language*, **27**, 521-536.

MILLER, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, **63**, 81-97.

MULLIGAN, N. W. (2001). Generation and hypermnesia. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 436-450.

MULLIGAN, N. W. (2002). The emergence of item-specific encoding effects in between-subjects designs: Perceptual interference and multiple recall tests. *Psychonomic Bulletin & Review*, **9**, 375-382.

MULLIGAN, N. W. (2005). Total retrieval time and hypermnesia: Investigating the benefits of multiple recall tests. *Psychological Research*, **69**, 272-284.

MULLIGAN, N. W., & DUKE, M. D. (2002). Positive and negative general effects, hypermnesia, and total recall time. *Memory & Cognition*, **30**, 1044-1053.

MURPHY, M. D. (1979). Measurement of category clustering in free recall. In C. R. Puff (Ed.), *Memory organization and structure* (pp. 51-83). New York: Academic Press.

MURPHY, M. D., & PUFF, C. R. (1982). Free recall: Basic methodology and analyses. In C. R. Puff (Ed.), *Handbook of research methods in human memory and cognition* (pp. 99-128). New York: Academic Press.

PASHLER, H., ROHRER, D., CEPEDA, N. J., & CARPENTER, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, **14**, 187-193.

PELLEGRINO, J. W., & HUBERT, J. L. (1982). The analysis of organization and structure in free recall. In C. R. Puff (Ed.), *Handbook of research methods in human memory and cognition* (pp. 129-172). New York: Academic Press.

PUFF, C. R. (1979). Memory organization research and theory: The state of the art. In C. R. Puff (Ed.), *Memory organization and structure* (pp. 3-17). New York: Academic Press.

ROEDIGER, H. L., III, & KARPICKE, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, **1**, 181-210.

ROEDIGER, H. L., III, & KARPICKE, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, **17**, 249-255.

ROENKER, D. L., THOMPSON, C. P., & BROWN, S. C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin*, **76**, 45-48.

ROHRER, D., TAYLOR, K., & SHOLAR, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **36**, 233-239.

SCHMIDT, S. R., & CHERRY, K. (1989). The negative generation effect: Delineation of a phenomenon. *Memory & Cognition*, **17**, 359-369.

SLAMECKA, N. J. (1968). An examination of trace storage in free recall. *Journal of Experimental Psychology*, **76**, 504-513.

SPITZER, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, **30**, 641-656.

STERNBERG, R. J., & TULVING, E. (1977). The measurement of subjective organization in free recall. *Psychological Bulletin*, **84**, 539-556.

SZPUNAR, K. K., MCDERMOTT, K. B., & ROEDIGER, H. L., III (2008). Testing during study insulates against the build-up of proactive interference. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **34**, 1392-1399.

TSE, C.-S., BALOTA, D. A., & ROEDIGER, H. L., III (in press). The benefits and costs of repeated testing on the learning of face–name pairs in healthy older adults. *Psychology & Aging*.

TULVING, E. (1962). Subjective organization in free recall of unrelated words. *Psychological Review*, **69**, 344-354.

TULVING, E., & PEARLSTONE, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning & Verbal Behavior*, **5**, 381-391.

TULVING, E., & WATKINS, M. J. (1974). On negative transfer: Effects of testing one list on the recall of another. *Journal of Verbal Learning & Verbal Behavior*, **13**, 181-193.

VAN OVERSCHELDE, J. P., RAWSON, K. A., & DUNLOSKY, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1963) norms. *Journal of Memory & Language*, **50**, 289-335.

ZAROMB, F. M. (2010). *Organizational processes contribute to the test-*

ing effect in free recall. Unpublished doctoral dissertation, Washington University, St. Louis.

NOTE

1. Although the focus of the present research was on whether testing enhances conceptual organization more than does studying alone, it is also worth asking how one can reconcile our finding that four tests during the learning phase did not produce greater category clustering than taking only two tests with that of Mulligan (2005, Experiment 2) who showed that taking four successive recall tests of 5-min duration each produced greater clustering 2 days later than did taking a single 20-min recall test. Moreover, Mulligan reported similar item gains but fewer item losses between the initial learning phase and the final recall test in the multiple- than in the single-test condition. These and related findings have been taken as support for the view that repeated testing promotes the development of increasingly stable retrieval strategies (e.g., Hunt & McDaniel, 1993; McDaniel, Moore, & Whiteman, 1998). By contrast, we found similar item gains (3.9 vs. 3.5) ($t < 1$) and item losses (17.1 vs. 15.5) ($t = 1$) between the last test during the study session and the delayed test in the two-test and four-test conditions, respectively. We speculate that our inclusion of additional study trials between tests and, possibly, the use of a within-subjects experimental design may have minimized differences among the repeated testing conditions.

(Manuscript received September 6, 2009;
revision accepted for publication April 26, 2010.)