Check for updates

Research Article

Psychological Science
2016, Vol. 27(2) 223–230
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797615617778
pss.sagepub.com

The Critical Importance of Retrieval—and Spacing—for Learning

Nicholas C. Soderstrom, Tyson K. Kerr, and Robert A. Bjork
Department of Psychology, University of California, Los Angeles

Abstract

We examined the impact of repeated testing and repeated studying on long-term learning. In Experiment 1, we replicated Karpicke and Roediger's (2008) influential results showing that once information can be recalled, repeated testing on that information enhances learning, whereas restudying that information does not. We then examined whether the apparent ineffectiveness of restudying might be attributable to the spacing differences between items that were inherent in the between-subjects design employed by Karpicke and Roediger. When we controlled for these spacing differences by manipulating the various learning conditions within subjects in Experiment 2, we found that both repeated testing and restudying improved learning, and that learners' awareness of the relative mnemonic benefits of these strategies was enhanced. These findings contribute to understanding how two important factors in learning—test-induced retrieval processes and spacing—can interact, and they illustrate that such interactions can play out differently in between-subjects and within-subjects experimental designs.

Keywords

learning, memory, testing, retrieval practice, spacing

Received 6/15/15; Revision accepted 10/26/15

Memory tests do not merely assess memory. The retrieval practice promoted by testing acts as a "memory modifier" (Bjork, 1975) by rendering successfully retrieved information more recallable in the future than if that same information had not been tested or had been merely restudied (i.e., *the testing effect*). Testing can also potentiate, or enhance, the effectiveness of subsequent study sessions (for a review of the direct and indirect benefits of testing, see Roediger, Putnam, & Smith, 2011).

In a highly cited article, Karpicke and Roediger (2008) reported a particularly dramatic demonstration of the benefits of retrieval practice. In their experiment, subjects studied Swahili-English vocabulary pairs (e.g., *elimu-science*) according to several different learning schedules that varied in the amount of repeated studying and repeated testing during the acquisition phase. They used four between-subjects conditions, each with four study (S)–test (T) cycles: ST, in which all pairs were studied and then tested in each cycle; S_NT, in which pairs recalled on a test were dropped from subsequent study periods but retained in subsequent test periods ("N" refers to "nonrecalled" items, and thus S_N indicates that only nonrecalled items were restudied); ST_N, in which pairs recalled on a test were dropped from subsequent test periods but retained in subsequent study periods; and S_NT_N, in which pairs recalled on a test were dropped from both subsequent study and subsequent test periods. At the end of the acquisition phase, subjects were asked to predict how many pairs they would remember on a test in 1 week. They then returned after 1 week for a final retention test that included all of the pairs.

The results of Karpicke and Roediger's (2008) experiment were clear and striking. Although cumulative learning at the end of the acquisition phase was virtually identical across the four conditions—in the sense that each subject, regardless of condition, had recalled almost every pair at least once successfully—and although subjects in the four

Corresponding Author:
Nicholas C. Soderstrom, Department of Psychology, 1285 Franz Hall, Box 951563, University of California, Los Angeles, Los Angeles, CA 90095-1563
E-mail: nsoderstrom@psych.ucla.edu

conditions provided similar memory predictions, the authors found that repeated testing, but not repeated studying, had large benefits for long-term learning. Specifically, in both of the conditions in which the pairs were always tested (ST and S_NT), subjects recalled approximately 80% of the items on the final test. In the conditions in which retrieved pairs were dropped from subsequent testing (ST_N and S_NT_N), subjects recalled approximately 36% and 33% of the items, respectively. Overall, then, repeated testing appeared to act as a powerful learning tool, whereas repeated studying seemed to produce no additional benefits for learning.

In thinking about Karpicke and Roediger's (2008) dramatic findings, though, it occurred to us that differences arising from their between-subjects experimental design might have contributed to their findings. Although manipulating the learning conditions between subjects makes sense from a practical standpoint, given that learners are likely to adopt a consistent policy in regulating their own studying, we wondered whether differences in the spacing between successive study periods in the ST, S_NT, ST_N, and S_NT_N conditions might have contributed to the apparent ineffectiveness of the study trials across the conditions. As illustrated in Figure 1, in the two conditions that produced the best long-term learning—the ST and S_NT conditions—after the first study period all three succeeding study periods were always preceded by 40 test trials; in contrast, the amount of testing that preceded those study periods in the ST_N and S_NT_N conditions declined markedly over time. This difference is important because a large literature on the *spacing effect* suggests that spacing repeated study opportunities, compared with massing them, enhances learning (for a review, see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006).

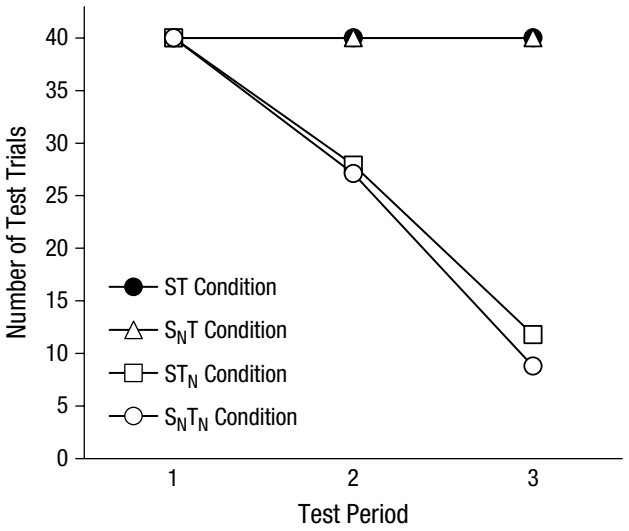

Fig. 1. The average number of test trials during the three test periods that preceded study periods in the four conditions of Karpicke and Roediger's (2008) experiment. For an explanation of the condition labels, see the text.

The sharp decline in testing over time in the ST_N condition might also explain why this condition did not produce *test-potentiated learning*, that is, enhancement of the effectiveness of studying after testing (e.g., see Izawa, 1966; Soderstrom & Bjork, 2014). Test-potentiated learning is positively related to the amount of testing that occurs prior to studying (e.g., Arnold & McDermott, 2013), and thus the amount of testing that occurred in the ST_N condition may not have been sufficient to convey such benefits.

The Present Study

To test whether the apparent ineffectiveness of restudying in Karpicke and Roediger's (2008) experiment might be attributable to the spacing differences between items in their between-subjects design, we carried out two experiments, the first a direct replication of Karpicke and Roediger's experiment and the second a version of the experiment in which the four learning conditions were manipulated within subjects (i.e., items designated ST, S_NT, ST_N, and S_NT_N were intermixed and experienced by each subject). With this within-subjects design, we were better able to control for the amount of testing that occurred before the study periods (and, therefore, the spacing of study opportunities). We reasoned that if such spacing is important for the benefits of repeated studying in the current paradigm, repeated studying would have little, if any, benefit in our experiment with the between-subjects design (as Karpicke & Roediger found) but would confer significant benefits in our experiment with the within-subjects design.

Method

Sample sizes for the experiments were determined on the basis of prior work examining test-enhanced learning (e.g., Karpicke & Roediger, 2008; Soderstrom & Bjork, 2014). Sixty-four undergraduates (mean age = 20.53 years; 48 female, 16 male) at the University of California, Los Angeles (UCLA), participated in Experiment 1 for partial course credit. During the acquisition phrase, they attempted to learn 40 Swahili-English word pairs (e.g., *elimu-science*; taken from Karpicke & Roediger, 2008) across a total of four study-test cycles (i.e., eight alternating study and test periods). During the study periods, the Swahili words and their English translations were presented one at a time for 5 s each, and subjects were asked to study each pair with the goal of subsequently being able to recall the English word when presented with the Swahili word. During the test periods, the Swahili words were presented one at a time for 8 s each, and subjects attempted to type in the English translations within that time.

A between-subjects design was used. A separate group of subjects (16 in each group) was assigned to each of the four learning conditions of Karpicke and Roediger's (2008) experiment: ST, S_NT, ST_N, and S_NT_N. As described earlier, these conditions differed with respect to how correctly recalled items were treated in subsequent study and test periods. Subjects in all of the conditions performed a 30-s distractor task (solving simple multiplication problems) after each study period. The orders in which items were presented during study and test periods were randomized, and no feedback was given during test periods.

After the final test period during the acquisition phase, subjects in all of the conditions were asked to make an aggregate judgment of their learning; specifically, they predicted how many of the 40 pairs they would recall on a test that would be administered in 1 week. Subjects then returned after 1 week for a final retention test. During this test, the Swahili words were presented one at a time for 15 s each, and subjects attempted to type their English translations within that time. Immediately after the retention test, subjects made a final metacognitive judgment regarding the perceived effectiveness of the learning procedures used in the experiment. Specifically, the four learning procedures—ST, S_NT, ST_N, and S_NT_N—were explained to subjects in detail, and they were asked to rank-order how effective the procedures would be for their own learning. The rankings were made on a scale from 1, *least effective*, to 4, *most effective*. After the procedures were ranked, subjects were thanked for their participation.

In Experiment 2, 36 undergraduates (mean age = 20.08 years; 30 female, 6 male) at UCLA participated for partial course credit. Given that Experiments 1 and 2 were completed in series, they were conducted at slightly different times during the academic quarter, although by the same research assistants. The design, materials, and procedure of Experiment 2 were similar to those of Experiment 1 with the exception that learning condition was manipulated within subjects. The 40 word pairs were equally divided into ST, S_NT, ST_N, and S_NT_N items (10 of each), and each type of item was experienced by each subject. Thus, the initial study-test cycle included all of the pairs, and subsequent study and test periods included some of these pairs, according to the condition to which they had been assigned. The specific pairs that were designated to be ST, S_NT, ST_N, and S_NT_N items were randomly chosen for each subject and retained their status throughout the experiment.

Results

Table 1 shows the mean number of trials during each study and test period of the acquisition phase, as well as the total number of trials, for each condition in Experiment 1. The ST condition, in which all the pairs were presented during each study and test period, contained the most trials (320.00), and the S_NT_N condition, in which successfully retrieved items were dropped from subsequent study and test periods, contained the fewest trials (194.00). The total number of trials was similar between the S_NT (263.94) and ST_N (265.94) conditions, but, of course, there were more test trials in the S_NT condition and more study trials in the ST_N condition by virtue of the differing dropout procedures in these two conditions. Thus, it is clear that the spacing intervals between study

Table 1. Mean Number of Trials During the Acquisition Phase for Each Learning Condition in Experiment 1

	Study-test cycle								
	1		2		3		4		
Condition	Study	Test	Study	Test	Study	Test	Study	Test	Total number of trials
ST	40.00	40.00 (40.00)	40.00 (40.00)	40.00 (40.00)	40.00 (40.00)	40.00 (40.00)	40.00 (40.00)	40.00 (40.00)	320.00
S_NT	40.00	40.00 (40.00)	32.81 (36.40)	40.00 (36.40)	20.19 (30.10)	40.00 (30.10)	10.94 (25.47)	40.00 (25.47)	263.94
ST_N	40.00	40.00 (40.00)	40.00 (40.00)	34.56 (37.28)	40.00 (37.28)	21.50 (30.75)	40.00 (30.75)	9.88 (24.94)	265.94
S_NT_N	40.00	40.00 (40.00)	32.56 (36.28)	32.56 (32.56)	17.19 (24.87)	17.19 (17.19)	7.25 (12.22)	7.25 (7.25)	194.00

Note: In the ST condition, all pairs were studied and then tested in each cycle; in the S_NT condition, pairs recalled on a test were dropped from subsequent study periods but retained in subsequent test periods; in the ST_N condition, pairs recalled on a test were dropped from subsequent test periods but retained in subsequent study periods; and in the S_NT_N condition, pairs recalled on a test were dropped from all subsequent study and test periods. The numbers in parentheses indicate the average number of trials intervening between a given item in the indicated period and the most recent prior exposure to that item.

periods differed significantly across the learning conditions as a result of the between-subjects nature of the experiment, as did the spacing in Karpicke and Roediger's (2008) experiment. In the ST and S_NT conditions—the conditions that yielded the highest recall—after the first study period all succeeding study periods were preceded by 40 test trials, whereas the amount of testing that preceded study sessions in the ST_N and S_NT_N conditions declined over successive cycles.

Table 2 shows the mean number of trials during each study and test period of the acquisition phase, as well as the total number of trials, for each type of item in Experiment 2. Given the within-subjects design of this experiment, each study and test period contained a mixture of the four types of items. As in Experiment 1, ST items were presented on the most trials (80.00), S_NT_N items were presented on the fewest trials (46.54), and the total number of trials was similar between S_NT (64.30) and ST_N (63.56) items. In contrast to Experiment 1, however, because the four item types were mixed within each study period, the amount of testing that occurred prior to study was similar for the four types of items. On average, 40.00, 35.61, and 27.89 test trials occurred immediately before the study periods in the second, third, and fourth cycles, respectively. Thus, although the amount of testing declined over time as a result of the various dropout procedures, the spacing that was produced by virtue of the test periods did not favor one type of item over another.

Figure 2 shows the cumulative learning curves for each of the four learning conditions in the acquisition phases of Experiments 1 and 2. A 4 (learning condition: ST, S_NT, ST_N, S_NT_N) × 4 (test period: 1, 2, 3, 4) mixed-model analysis of variance (ANOVA) on the data for Experiment 1 revealed that, as expected, recall improved during the course of the acquisition phase, $F(3, 180) = 727.01$, $p < .001$, $\eta_p^2 = .92$, and that there were no significant differences in the learning curves across the four conditions ($F < 1$). Furthermore, recall performance during the final (fourth) test period did not differ reliably across the conditions ($F < 1$); subjects in the four learning conditions finished the acquisition phase having recalled similar proportions (approximately .85) of items at least once. Given that acquisition performance was similar across the conditions, it is unsurprising that subjects' aggregate judgments of their learning did not differ significantly across the conditions ($F < 1$). The mean number of items subjects predicted they would recall was 15.31 in the ST condition, 16.87 in the S_NT condition, 13.31 in the ST_N condition, and 14.00 in the S_NT_N condition.

Results were similar for Experiment 2 (see Fig. 2). A 4 (item type: ST, S_NT, ST_N, S_NT_N) × 4 (test period: 1, 2, 3, 4) repeated measures ANOVA revealed that recall performance improved over the course of the acquisition phase, $F(3, 105) = 624.17$, $p < .001$, $\eta_p^2 = .95$, and that there were no significant differences in the learning curves for the different types of items ($F < 1$). Additionally, recall performance during the final (fourth) test period did not differ reliably across item types ($F < 1$); subjects in Experiment 2 finished the acquisition phase having recalled similar proportions of the four types of items (approximately .90) at least once. Also, aggregate judgments of learning were comparable to those in Experiment 1: Subjects in Experiment 2 predicted that they would recall, on average, 15.28 items, 95% confidence interval = [12.54, 18.01], in a week.

Figure 3 shows the proportion of items correctly recalled on the final retention test in the four learning conditions of the two experiments. In Experiment 1, a one-way ANOVA revealed that recall differed across the

Table 2. Mean Number of Trials During the Acquisition Phase for Each Type of Item in Experiment 2

	Study-test cycle								
	1		2		3		4		
Item type	Study	Test	Study	Test	Study	Test	Study	Test	Total number of trials
ST	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	80.00
S_NT	10.00	10.00	8.19	10.00	4.30	10.00	1.81	10.00	64.30
ST_N	10.00	10.00	10.00	7.92	10.00	4.03	10.00	1.61	63.56
S_NT_N	10.00	10.00	7.69	7.69	3.86	3.86	1.72	1.72	46.54
Total[a]	40.00	40.00 (40.00)	35.88 (37.94)	35.61 (35.75)	28.16 (31.89)	27.89 (28.03)	23.53 (25.71)	23.33 (23.43)	254.40

Note: ST items were studied and then tested in each cycle; S_NT items recalled on a test were dropped from subsequent study periods but retained in subsequent test periods; ST_N items recalled on a test were dropped from subsequent test periods but retained in subsequent study periods; and S_NT_N items recalled on a test were dropped from all subsequent study and test periods.
[a]The numbers in parentheses indicate the average number of trials intervening between a given item in the indicated period and the most recent prior exposure to that item. Note that these numbers were the same for the four item types by virtue of the within-subjects design.

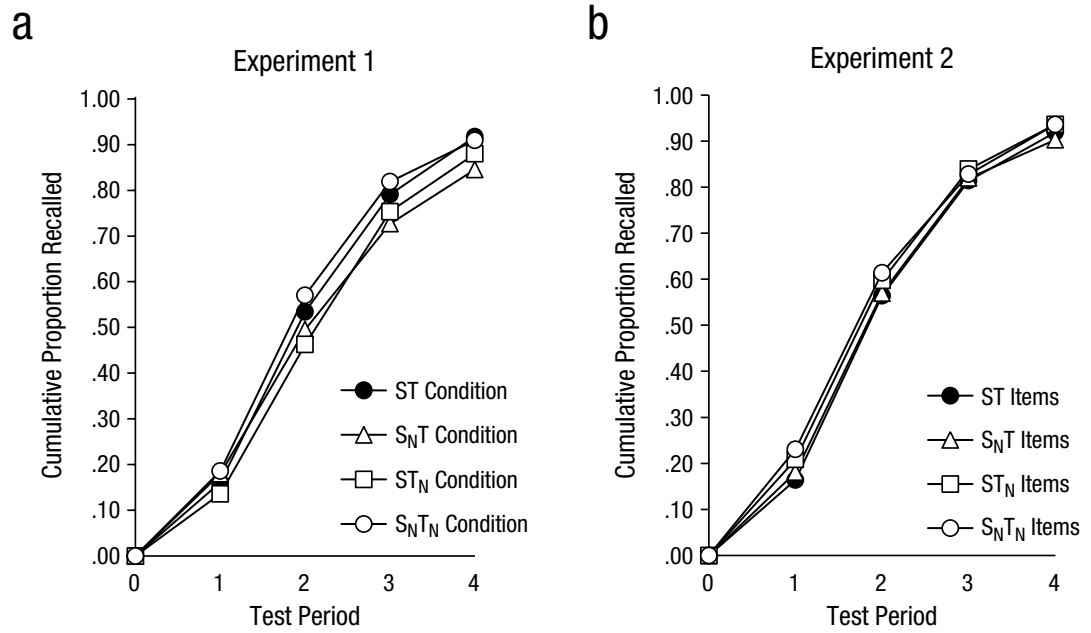

Fig. 2. Cumulative learning curves for the acquisition phases in Experiments 1 and 2. For an explanation of the condition (item-type) labels, see the text.

conditions, $F(3, 60) = 10.88$, $p < .001$, $\eta_p^2 = .35$. Tukey post hoc analyses showed that subjects in the ST and S_NT conditions recalled the highest proportion of items and that there was no significant difference in recall between these conditions ($p > .250$). Subjects in the ST condition recalled more items than subjects in both the ST_N ($p < .001$) and the S_NT_N ($p = .001$) conditions; likewise, subjects in the S_NT condition recalled more items than those

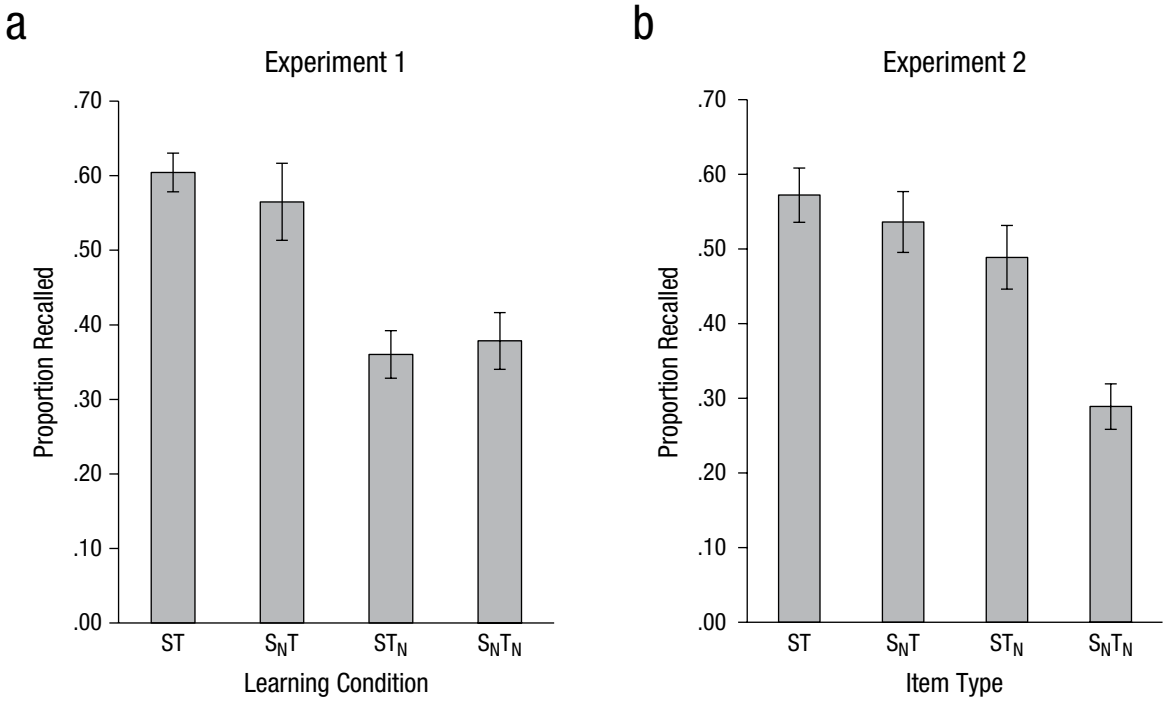

Fig. 3. Mean proportion of items recalled on the final retention test administered 1 week after the acquisition phase in Experiments 1 and 2. Results are shown separately for each learning condition (item type); for an explanation of the labels, see the text. Error bars represent ±1 *SEM*.

in both the ST_N ($p = .002$) and the S_NT_N ($p = .005$) conditions. No significant difference in recall was found between the ST_N and S_NT_N conditions ($p > .250$). Thus, repeated testing, but not repeated studying, appeared to be the critical ingredient for learning in Experiment 1; the pattern found by Karpicke and Roediger (2008) was replicated.

A one-way ANOVA on final recall in Experiment 2 revealed a reliable effect of item type, $F(3, 105) = 30.59$, $p < .001$, $\eta_p^2 = .47$. Follow-up t tests indicated that recall did not differ significantly between ST and S_NT items ($p = .205$), nor did it differ significantly between S_NT and ST_N items ($p = .200$). ST items, however, were better recalled than ST_N items, $t(35) = 3.25$, $p = .003$, $d = 0.33$, which were better recalled than S_NT_N items, $t(35) = 5.76$, $p < .001$, $d = 0.89$. Thus, in Experiment 2, repeated testing was effective for long-term learning, but so was repeated studying, as evidenced by the fact that ST_N items were better recalled than S_NT_N items.

Finally, upon completion of the final test, subjects in both experiments were informed of the four different learning procedures and were asked to rank them according to how effective the procedures would be for their own learning. Figure 4 displays the average metacognitive rankings of the procedures provided by subjects. In Experiment 1, a mixed-model ANOVA showed that the rankings of the learning procedures differed, $F(3, 180) = 23.96$, $p < .001$, $\eta_p^2 = .29$, and that these rankings did not vary with the condition to which subjects were assigned ($F < 1$). The ST procedure was ranked as being the most effective for learning, and the S_NT_N procedure was ranked as being the least effective. The ST procedure was ranked higher than the ST_N procedure, $t(63) = 2.36$, $p = .022$, $d = 0.42$; the ST_N procedure was ranked higher than the S_NT procedure, $t(63) = 2.25$, $p = .028$, $d = 0.50$; and the S_NT procedure was ranked higher than the S_NT_N procedure, $t(63) = 4.98$, $p < .001$, $d = 0.90$. Thus, although repeated testing was better for actual learning than was repeated studying in Experiment 1 (i.e., final recall was better in the S_NT condition than in the ST_N condition), subjects endorsed repeated studying as being more effective than repeated testing (i.e., the ST_N procedure was ranked higher than the S_NT procedure).

In Experiment 2 (see Fig. 4), a one-way ANOVA showed that the rankings of the learning procedures differed, $F(3, 105) = 6.26$, $p = .001$, $\eta_p^2 = .15$. Follow-up t tests revealed that the rankings of the ST and S_NT procedures did not differ significantly ($p > .250$) and that both the ST and the S_NT procedures were ranked marginally higher than the ST_N procedure, $t(35) = 1.87$, $p = .069$, $d = 0.47$, and $t(35) = 1.52$, $p = .138$, $d = 0.42$, respectively. Finally, the ST_N procedure was ranked higher than the S_NT_N procedure, $t(35) = 2.16$, $p < .038$, $d = 0.59$. Taken together, then, subjects' rankings of the procedures in Experiment 2 aligned rather well with their pattern of actual recall.

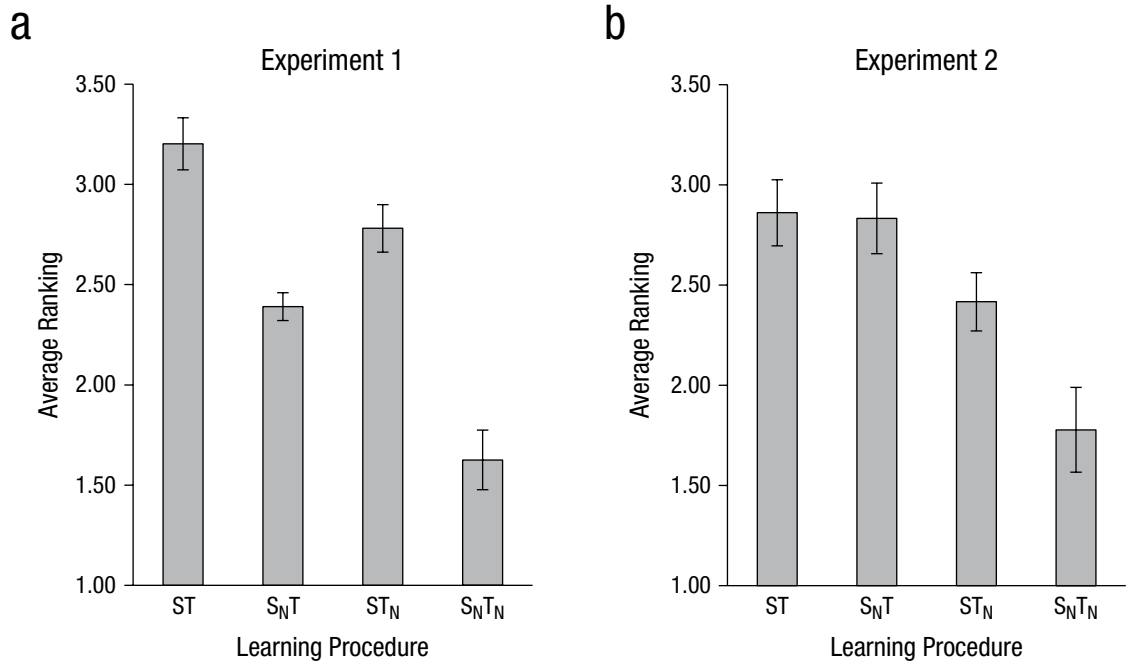

Fig. 4. Average metacognitive rankings of the four learning procedures in Experiments 1 and 2. The rankings were made on a scale from 1, *least effective*, to 4, *most effective*. For an explanation of the labels, see the text. Error bars represent ±1 *SEM*.

General Discussion

The findings from these experiments are important from several perspectives. They provide additional evidence that retrieval is critical for learning and that testing is often more powerful than restudying. At the same time, though, they demonstrate that restudying items that have been recalled earlier is far from useless and that the apparent uselessness of restudying such items in Karpicke and Roediger's (2008) experiment (and in our replication of that experiment) appears to be a product of the spacing of successive test and restudy trials when the ST, ST_N, S_NT, and S_NT_N conditions are implemented on a between-subjects basis. Decades ago, Poulton (1982; also see McDaniel & Bugg, 2008) reviewed evidence from both memory experiments and reaction time studies and concluded that researchers should utilize both within- and between-subjects designs when examining a given memory phenomenon in order to account for what he called "influential companions." He was actually most concerned about within-subjects designs and the possibility that experiencing a given condition can lead subjects to transfer some strategy or expectation appropriate to that condition to the materials in a different condition. However, the present findings illustrate that between-subjects designs can also create influential companions, even for items that are nominally in the same condition. In a related vein, recent work suggests that testing effects are larger in within-subjects designs compared with between-subjects designs (Mulligan & Peterson, 2015; but see Rowland, Littrell-Baez, Sensenig, & DeLosh, 2014). Clearly, the type of experimental design employed to investigate memory phenomena matters, and the current study adds to a growing list of studies that corroborate this fact.

We have demonstrated that restudy opportunities, if properly spaced, can enhance the learning of previously recalled information. Indeed, subjects in the ST_N condition learned more than subjects in the S_NT_N condition in the within-subjects experiment, but not in the between-subjects experiment. Why, then, was recall of ST items no better than recall of S_NT items in Experiment 2? After all, for ST items, subjects restudied previously recalled information in a spaced fashion in addition to being repeatedly tested on that information, whereas for S_NT items, subjects were repeatedly tested on previously recalled information but did not restudy that information. One possible explanation for this puzzle is that restudying previously recalled information, even if it is properly spaced, has negligible effects on learning if that same information is also repeatedly tested. That is, repeated studying might not have a significant impact on learning over and beyond what is already achieved by repeated testing. In Experiment 2, repeated studying boosted learning when testing was not repeated (ST_N vs. S_NT_N) but did not confer learning benefits when testing was also repeated (ST vs. S_NT). Of course, further research is warranted to examine this issue more thoroughly.

Another contribution of the present research concerns the clear differences in metacognitive judgments between Experiments 1 and 2. Upon completion of the final retention test, subjects in Experiment 1 ranked the ST condition as most effective and the S_NT_N condition as least effective; however, the ST_N condition was ranked as more effective than the S_NT condition, which indicates that subjects were prone to the metacognitive illusion that restudying is better for learning than is repeated testing. Perhaps reflecting the tendency for learners to endorse conditions of learning that are perceived to be relatively easy to execute, this finding is consistent with previous experimental (e.g., Karpicke, 2009; Roediger & Karpicke, 2006) and survey (e.g., McCabe, 2011) research, and accords with the findings demonstrating that what is effective for learning is often misaligned with what people *think* is effective for learning (for reviews, see Bjork, 1999; Bjork, Dunlosky, & Kornell, 2013).

In contrast to subjects' rankings of the learning procedures in Experiment 1, however, subjects' rankings in Experiment 2 matched quite well with their pattern of actual recall. Thus, equipping learners with the experiences associated with each type of learning procedure seemed to lead to an appreciation of the relative mnemonic benefits of retrieval practice and restudying, a finding that may be of interest to educators and researchers who seek to foster metacognitive sophistication in learners. Generally speaking, the differences in metacognitive judgments between our experiments comport with previous work showing that subjects, when predicting their future recall, were insensitive to retention interval when it was manipulated between subjects, but *were* sensitive to retention interval when it was varied within subjects (Koriat, Bjork, Sheffer, & Bar, 2004).

The present findings also illustrate the importance of distinguishing between learning—that is, the relatively permanent changes in knowledge that support long-term retention—and performance during acquisition—which reflects temporary fluctuations in knowledge. Indeed, overwhelming empirical evidence from both the verbal- and motor-learning domains supports the notion that learning and performance are dissociable (for a review, see Soderstrom & Bjork, 2015). In the current context, the cumulative recall patterns during acquisition were similar across the four learning conditions in both experiments (i.e., performance during acquisition was the same), yet retention differed substantially across the learning conditions after 1 week (i.e., learning differed). Thus, anyone interested in optimizing long-term retention should be cognizant of the fact that performance during acquisition can be an unreliable index of actual learning.

Concluding Comments

Given that many students report using self-testing during their own studying (Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007), it is important to identify ways to enhance the effectiveness of such a strategy. To this end, we first conducted a direct replication of Karpicke and Roediger's (2008) experiment showing that, once information can be recalled, repeated testing on that information enhances learning, whereas repeated studying of that information does not. However, the apparent ineffectiveness of restudying seemed to be attributable, at least in part, to the spacing differences between study sessions that were inherent in the between-subjects design used by Karpicke and Roediger. When we manipulated the learning conditions within subjects—and thus controlled for the amount of testing (and, therefore, spacing) that preceded study sessions—we found that both repeated testing and repeated studying improved learning, and that learners' awareness of the relative mnemonic benefits of retrieval practice and restudying was enhanced.

Author Contributions

N. C. Soderstrom and R. A. Bjork generated the research design from a general idea contributed by R. A. Bjork. T. K. Kerr programmed the experiments and helped with data collection. N. C. Soderstrom performed the data analyses, drafted the manuscript, and worked with R. A. Bjork on revising the manuscript. All authors approved the final version of the manuscript for submission.

Acknowledgments

We thank Gayan Seneviratna for his help with data collection.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

Grant 29192G from the James S. McDonnell Foundation supported this research.

References

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 940–945.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417–444.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*, 354–380.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*, 126–134.

Izawa, C. (1966). Reinforcement-test sequences in paired-associate learning. *Psychological Reports*, *18*, 879–919.

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. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, *133*, 643–656.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*, 219–224.

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*, 462–476.

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review*, *15*, 237–255.

Mulligan, N. W., & Peterson, D. J. (2015). Negative and positive testing effects in terms of item-specific and relational processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 859–871.

Poulton, E. C. (1982). Influential companions: Effects of one strategy on another in the within-subjects designs of cognitive psychology. *Psychological Bulletin*, *91*, 673–690.

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

Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mestre & B. Ross (Eds.), *Psychology of learning and motivation: Vol. 55. Cognition in education* (pp. 1–36). Oxford, England: Elsevier.

Rowland, C. A., Littrell-Baez, M. K., Sensenig, A. E., & DeLosh, E. L. (2014). Testing effects in mixed- versus pure-list designs. *Memory & Cognition*, *42*, 912–921.

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language*, *73*, 99–115.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*, 176–199.