Psychonomic Bulletin & Review
2007, 14 (3), 474-478

Testing beyond words: Using tests to enhance visuospatial map learning

Shana K. Carpenter and Harold Pashler
University of California, San Diego, La Jolla, California

Psychological research shows that learning can be powerfully enhanced through testing, but this finding has so far been confined to memory tasks requiring verbal responses. We explored whether testing can enhance learning of visuospatial information in maps. Fifty subjects each studied two maps, one through conventional study, and the other through computer-prompted tests. For the tests, the subjects were repeatedly presented with the same map with one feature deleted (e.g., a road or a river), and they tried to covertly recall the missing feature and its location. Subjects' map drawings after 30 min were significantly better for maps learned through tests in comparison with maps learned through the same amount of time devoted to conventional study. These results suggest that the testing effect is not limited to the types of memory that require discrete, verbal responses, and that utilizing covert retrievals may allow the effect to be extended to a variety of complex, nonverbal learning tasks.

Many studies have shown that a memory test is useful not only for assessing memory, but also for *improving* memory. Research going back several decades has shown that tests can strengthen memory more than extra opportunities to restudy the material. This was first noted in studies looking at recall of word lists (e.g., Allen, Mahler, & Estes, 1969; Lachman & Laughery, 1968). It has also been found in foreign-language vocabulary learning. For example, when people were given 5 sec to try to retrieve the English equivalent of an Eskimo word, and then both words were shown for an additional 5 sec, subsequent memory was strengthened more than when the English and Eskimo words were both provided for the entire 10-sec period (Carrier & Pashler, 1992).

This benefit of testing over restudying—which will be referred to here as the *testing effect*, but is also sometimes referred to as *retrieval practice*—has been further demonstrated with general knowledge facts (McDaniel & Fisher, 1991), face–name pairs (Carpenter & DeLosh, 2005; Landauer & Bjork, 1978), text passages (Nungester & Duchastel, 1982; Roediger & Karpicke, 2006a), paired-associate verbal items (Carpenter, Pashler, & Vul, 2006; Cull, 2000), and free recall of word lists (Carpenter & DeLosh, 2006; Kuo & Hirshman, 1996; Wheeler, Ewers, & Buonanno, 2003).

The potential utility of tests to improve students' learning in educational contexts has sparked a lot of enthusiasm among psychologists (e.g., Bjork, 1988; Dempster, 1989, 1996; Glover, 1989; McDaniel & Einstein, 2005; Roediger & Karpicke, 2006b; Roediger & Marsh, 2005; Wheeler & Roediger, 1992). This potential was also noted by a National Research Council–sponsored review of the practical results of five decades of learning and memory research (Druckman & Bjork, 1991, 1994).

Limitations of the Testing Effect

Efforts to harness the testing effect have so far been limited to memory tasks involving verbal responses. In both school and occupational contexts, however, people often learn information that is not well conveyed through words (e.g., Healy, King, & Sinclair, 1997; Wittman & Healy, 1995). Can testing also assist in learning visuospatially rich materials that do not require verbal responses? The present study explored this question with map learning.

An immediate obstacle confronts efforts to implement testing with maps. Consider, for example, trying to learn landscape features such as rivers, highways, and buildings. Although it is easy to collect and score brief verbal responses, it is not obvious how this could be done in a convenient way with map features. The method we used rests on two simple observations. First, there are some indications, on the basis of two studies conducted in our lab using verbal materials (Carpenter et al., 2006; Carpenter, Pashler, Wixted, & Vul, 2007), that the testing effect may not always require overt retrieval. Second, when a person is motivated to learn, he or she can score his or her own responses to determine which items require further study. Self-scoring is, after all, the basis for individuals' successful use of flash cards.

The method we used incorporates the potential for tests to benefit memory in both direct and indirect ways. Studies of the testing effect with verbal memory have shown, for example, that the act of retrieval per se may have some direct benefit on memory retention (e.g., Carpenter &

S. K. Carpenter, scarpenter@ucsd.edu

DeLosh, 2006; Carrier & Pashler, 1992). Or, the act of retrieval may reveal which items have been sufficiently learned and which have not, thus resulting in a more efficient use of subsequent study time (see, e.g., Izawa, 1992). The degree to which these direct versus indirect benefits make up the testing effect has not been clearly specified through past research, nor was it addressed in the present study. The main goal of the present study was to determine whether the testing effect could be obtained with a visuospatial task that does not require an overt verbal response.

Subjects learned two maps—one through a test, and the other through an additional study opportunity. After 30 min, they were asked to draw both maps as well as they could. To motivate subjects to learn the maps in both conditions, we advised them that a $10 bonus would be given to anyone whose performance on the final test fell within the top one third of all scores.

METHOD

Subjects and Materials

A total of 52 undergraduates from UCSD participated for course credit. Each subject learned two maps containing 12 features each, such as roads, rivers, and buildings (see Figure 1). One map was learned through testing, with feedback (the test/study condition), and the other map was learned through additional studying (the study condition). The subjects finished learning one of the maps through one method (e.g., study) before learning the other map through the other method (e.g., test/study). Four counterbalancing conditions were created: (1) Map A was presented for study, followed by Map B for test/study; (2) Map A was presented for test/study, followed by Map B for study; (3) Map B was presented for study, followed by Map A for test/study; and (4) Map B was presented for test/study, followed by Map A for study. Each subject was randomly assigned to one of these four counterbalancing conditions: 1 (n = 17), 2 (n = 11), 3 (n = 13), or 4 (n = 11).

Design and Procedure

The subjects first read instructions that explained the test/study and study procedures. They then practiced the test/study procedure using a display of colored shapes. After the practice phase, the subjects were told that they would learn two maps, one using the test/study procedure they had practiced, and the other using the study procedure. Before seeing the maps, the subjects were advised that they would be given a later, unspecified memory test over both maps, and that if they scored within the top one third of all subjects on this test, they would receive $10.

During test/study, the subjects were first given 20 sec to study the map with all 12 features. They were then advised that they were about to be tested over the features of that map. The subjects pressed the space bar to begin the test, and were shown an incomplete display of the map, in which 1 of the 12 features was missing. The subjects were told to figure out what was missing and to form a mental image of the missing feature in its location. When they had formed the image, they pressed the space bar to see the complete map again. The subjects scored their own responses as correct or incorrect by pressing a button to indicate whether they successfully remembered the missing feature and its location. As soon as they responded, a 1-sec blank screen appeared, and then the same map was shown again with a different feature missing. After all 12 features had been tested in random order, they were tested again in a new random order, and the test/study procedure continued in this fashion for 100 sec.

During study, the subjects were first given 20 sec to study the map with all 12 features. Then, they were given a screen of instructions telling them that they would be given another opportunity to study the complete map again for 100 additional seconds, and this began when they pressed the space bar. The total duration for the test/study and the study was thus held constant at 120 sec.

After engaging in an unrelated visual attention task for 30 min, the subjects were given a blank sheet of paper and instructed to draw both maps as well as they could. One subject (from Condition 2) drew only one of the maps, and another subject (from Condition 4) drew the same map twice. Data from these 2 subjects were discarded, and all of the analyses were based on the data from the 50 remaining subjects.

Scoring the Map Drawings

The map drawings were scored by an experimenter who was blind to the conditions to which the maps were assigned. One point was

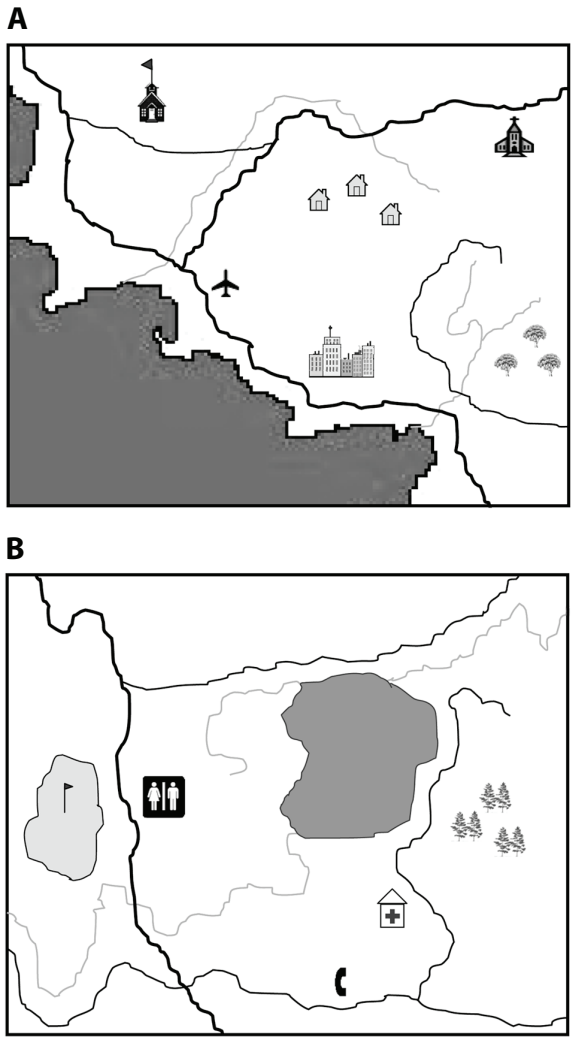

Figure 1. We created two maps, one of a town (upper panel), and one of a recreational area (lower panel). Each map contained four roads, two rivers, and six additional features such as a school or golf course. The original maps contained colors such as blue (for the lake, ocean, and rivers), green (for the trees and golf course), and red (for the school and first aid station). Each subject learned both of the maps, one through completing a test with feedback (test/study) that lasted 120 sec, and the other through engaging in additional study time (study) for 120 sec. Each subject was randomly assigned to one of four different counterbalancing conditions to learn both of the maps.

awarded for each correct feature placed in the correct location, and one point was deducted for each extra feature, or a feature that did not belong in the map at all. The location of features was scored according to both absolute and relative accuracy.

Absolute accuracy. A point was awarded if the feature was placed within the correct quadrant (NE, NW, SE, SW), without respect to other features included in the drawing. For example, in Map A, a point was given for the church if it was drawn in the NE quadrant. In Map B, a point was given for the telephone if it was drawn in the SE quadrant.

According to liberal quadrant (LQ) scoring, roads and rivers were considered correct if they were drawn in any of their correct quadrants. For example, a point was given for the main north–south road in Map A if it was drawn in either the NW or SE quadrant. A point was given for the main north–south road in Map B if it was drawn in either the NW or SW quadrant. According to stringent quadrant (SQ) scoring, roads and rivers were considered correct only if they were located in *all* of their correct quadrants, and only if the roads were connected correctly. For example, a point was given for the main north–south road in Map A only if it was drawn through both the NW and SE quadrants, and only if it was connected to both east–west roads. A point was given for the main north–south road in Map B only if it was drawn through both the NW and SW quadrants, and only if it was connected to both east–west roads. Withholding points for failing to connect roads applied only if subjects included those roads in their drawings.

Relative accuracy. A point was awarded if the feature was placed in the correct position relative to other features included in the drawing. For example, in Map A, a point was given for the houses if they were drawn south of the main east–west road and river, southwest of the church, and northeast of the airport. In Map B, a point was given for the telephone if it was drawn southwest of the first aid station, south of the lake, west of the lesser north–south road, and north of the southernmost east–west road.

According to liberal relative position (LRP) scoring, roads and rivers were scored as correct if they were located in the correct position relative to other features included in the drawing. For example, a point was given for the main east–west road in Map A if it was drawn north of the church, houses, or airport. A point was given for the main north–south road in Map B if it was drawn west of the restrooms, or east of the golf course. According to stringent relative position (SRP) scoring, roads and rivers were scored as correct only if they were located in the correct position relative to *all* of the other features included in the drawing, and only if the roads were connected correctly. For example, a point was given for the main east–west road in Map A only if it was drawn from the north side of the airport to the north side of the church, and only if it was connected to the main north–south road and the lesser east–west road. A point was given for the lesser north–south road in Map B only if it was drawn west of the trees but east of the telephone, first aid station, and lake, and only if it was connected to the southernmost east–west road. Withholding points for failing to connect roads applied only if subjects included those roads in their drawings.

RESULTS

Performance During Test/Study

Map B was easier to learn than Map A, on the basis of subjects' self-scoring. This was true whether performance was based on the first 12 test trials—when each feature was tested for the first time—or on *all* trials during the test/study phase. For the first 12 trials, the subjects scored an average of 82% on Map B ($SD = 14\%$) and 73% on Map A ($SD = 14\%$) [$t(48) = 2.11, p < .05$]. For all of the trials combined, the subjects scored an average of 87% on Map B ($SD = 12\%$) and 77% on Map A ($SD = 14\%$) [$t(48) = 2.54, p < .02$]. The subjects completed a higher number of test trials for Map B ($M = 22.15$, $SD = 8.32$) than for Map A ($M = 18.65$, $SD = 8.60$), but this difference was not significant ($t = 1.46$).

Accuracy of the Final Map Drawings

To estimate interrater agreement, we had the final map drawings from 10 subjects randomly selected and scored by an independent rater who was instructed on the scoring system but was blind to the conditions to which the maps were assigned. The correlation between the accuracy scores of the two raters was .91 for test/study and .96 for study, according to LQ scoring; .81 for test/study and .76 for study, according to SQ scoring; .81 for test/study and .96 for study, according to LRP scoring; and .94 for test/study and .81 for study, according to SRP scoring (all ps $< .02$). The following analyses were based on the accuracy of the final map drawings according to one rater.

The test/study condition was significantly more beneficial than the study condition by all four scoring procedures. According to a 2×4 mixed ANOVA with test condition (test/study vs. study) as the within-subjects factor and counterbalancing condition as the between-subjects factor, there was a significant main effect of test condition according to LQ scoring [$F(1,46) = 10.85$, $p < .005$, $MS_e = .026$], SQ scoring [$F(1,46) = 7.14$, $p < .02$, $MS_e = .023$], LRP scoring [$F(1,46) = 4.12$, $p < .05$, $MS_e = .024$], and SRP scoring [$F(1,46) = 4.82$, $p < .05$, $MS_e = .022$].

Counterbalancing condition did not affect accuracy for any of the scoring procedures (all Fs < 1), nor did it interact with test condition for LQ scoring ($F = 1.08$), SQ scoring ($F = .78$), LRP scoring ($F = 2.10$), or SRP scoring ($F = 1.18$). Table 1 shows the means and standard errors for the test/study and study conditions by all four scoring procedures.

DISCUSSION

Testing produced a significant enhancement of map learning. To our knowledge, this is the first demonstra-

Table 1
Mean Proportion Correct on Final Test As a Function of Scoring Procedure, Counterbalancing Condition, and Test Condition

Scoring Procedure	Counterbalancing Condition 1		2		3		4		Total	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Test/Study										
LQ	.82	.04	.68	.06	.73	.05	.75	.06	.74	.03
SQ	.69	.04	.56	.05	.58	.05	.63	.05	.61	.02
LRP	.76	.05	.61	.06	.65	.05	.72	.06	.68	.03
SRP	.68	.04	.55	.06	.58	.05	.63	.06	.61	.03
Study										
LQ	.64	.06	.66	.07	.62	.06	.62	.07	.63	.03
SQ	.54	.05	.52	.07	.53	.06	.53	.07	.53	.03
LRP	.63	.05	.62	.07	.67	.06	.56	.07	.62	.03
SRP	.56	.05	.54	.07	.57	.06	.51	.07	.54	.03

Note—The maps were scored according to a liberal quadrant (LQ) procedure, a stringent quadrant (SQ) procedure, a liberal relative position (LRP) procedure, and a stringent relative position (SRP) procedure. All four scoring procedures yielded a significant benefit of test/study over study.

tion that the testing effect can be found with a visuospatial task that does not require the production of an overt verbal response. It is encouraging to know that the effect is not limited to memory tasks that require writing, typing, or speaking a verbal response aloud, as have been used in many of the past studies on this topic (see, e.g., Dempster, 1996).

The map-learning task differs from other tasks involving nonverbal components—for example, face–name learning—for which testing effects have been demonstrated in past studies (e.g., Carpenter & DeLosh, 2005). Although a face contains properties that might be difficult to verbalize, face–name learning is still similar to paired-associate verbal learning in that it requires a verbal response (i.e., the name). In the map-learning task, although subjects might use verbal labels to code the presence of some geographic features (e.g., *church*, *golf course*), verbal descriptions are not likely to underlie coding of the spatial properties of those features.

These results help to broaden the boundaries of the testing effect beyond just verbal memory tasks; they should also encourage future theoretical work to explore the means by which tests are beneficial for nonverbal learning as well as verbal learning. Although some hypotheses considered for verbal materials might be pertinent to nonverbal materials as well (for reviews of hypotheses, see Carpenter & DeLosh, 2006; Carpenter et al., 2006; Carrier & Pashler, 1992), others would seem less applicable. Understanding these boundaries may also help in formulating and testing neurocomputational models of the phenomenon (see, e.g., Mozer, Howe, & Pashler, 2004).

The present study appears to be the first demonstration of the testing effect in which subjects—in both the test and study conditions—were offered a monetary incentive to learn the material as well as they could. In past studies comparing testing with additional studying, subjects may not always have been motivated to learn material in the study condition, and thus these studies may have underestimated what would be learned from additional study in real-world contexts in which learners are motivated. By offering a cash bonus for learning the maps in both conditions, the present study provides a stronger basis for concluding that testing is likely to offer a true learning advantage over and above the gains attributable to incentives (which may often already be present in practical learning contexts). It may be a good idea to incorporate similar incentives into future testing-effect studies in all domains.

Optimizing the Effect

It may be possible to refine these methods to produce even larger advantages of testing. On the basis of the number of test trials completed, subjects were tested less than twice on each feature, on average. It seems conceivable that whereas the benefits offered by additional study might soon reach diminishing returns as more study time was provided, testing would show an even greater advantage if time was increased so that each feature could be tested more often.

It is also conceivable that the testing advantage could be amplified further by allowing subjects to have additional tests over just those features they have not yet mastered, as in the drop-out (or "learning to criterion") method used in studies of verbal learning (e.g., Atkinson, 1972). Such a method would capitalize on the indirect benefits of testing by optimizing subsequent study time.

The testing advantage may also be greater when retention is tested after longer intervals (as has been reported for verbal materials; cf. Roediger & Karpicke, 2006a; Runquist, 1983; Wheeler et al., 2003). Manipulations that prevent subjects from basing their covert retrievals on working memory—as they were probably able to do, to some degree, in the present experiment—may further enhance the effect. For example, Whitten and Bjork (1977) found that an intervening test was more beneficial to retention if it was administered several seconds after the material was first presented, rather than immediately after the material was first presented.

Future Directions for Practical Applications

The need to learn information represented in maps arises commonly—not only in geography education, but also in other practical fields, such as military operations and transportation. Currently, the standard practice is for a learner to try to memorize the map by scrutinizing it. Given that the testing procedure described here was significantly more effective, and that the procedure can be implemented readily with computers, practical applications of these findings would seem to hold some potential.

It would be worthwhile to examine testing with other visuospatial materials, such as faces (e.g., law enforcement officials studying "wanted" posters), and the graphs and diagrams that are commonplace in scientific and technical training. Although the benefits of testing for discrete verbal materials have been exploited through the simple but effective technology of flash cards, computerized covert retrieval practice of the type used in this study may offer similar potential for enhancing a diverse range of nonverbal learning tasks.

AUTHOR NOTE

This work was supported by the Institute of Education Sciences (Grant R305H040108 from the U.S. Department of Education) and the National Institute of Mental Health (Grant R01-MH61549 to H.P.). We thank David Cun for his expert programming assistance and Noriko Coburn for her assistance with scoring. Correspondence concerning this article should be addressed to S. K. Carpenter, Department of Psychology, 0109, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0109 (e-mail: scarpenter@ucsd.edu).

REFERENCES

ALLEN, G. A., MAHLER, W. A., & ESTES, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning & Verbal Behavior*, **8**, 463-470.

ATKINSON, R. C. (1972). Ingredients for a theory of instruction. *American Psychologist*, **27**, 921-931.

BJORK, R. A. (1988). Retrieval practice and the maintenance of knowledge. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues. Vol. 1: Memory in everyday life* (pp. 396-401). Chichester, U.K.: Wiley.

CARPENTER, S. K., & DELOSH, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, **19**, 619-636.

CARPENTER, S. K., & DELOSH, E. L. (2006). Impoverished cue support

enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, **34**, 268-276.

CARPENTER, S. K., PASHLER, H., & VUL, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, **13**, 826-830.

CARPENTER, S. K., PASHLER, H., WIXTED, J. T., & VUL, E. (2007). *The effects of tests on learning and forgetting*. Manuscript submitted for publication.

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

CULL, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, **14**, 215-235.

DEMPSTER, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review*, **4**, 309-330.

DEMPSTER, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Handbook of perception and cognition: Memory* (pp. 317-344). San Diego: Academic Press.

DRUCKMAN, D., & BJORK, R. A. (EDS.) (1991). *In the mind's eye: Enhancing human performance.* Washington, DC: National Academy Press.

DRUCKMAN, D., & BJORK, R. A. (EDS.) (1994). *Learning, remembering, believing: Enhancing human performance.* Washington, DC: National Academy Press.

GLOVER, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, **81**, 392-399.

HEALY, A. F., KING, C. L., & SINCLAIR, G. P. (1997). Maintenance of knowledge about temporal, spatial, and item information: Memory for course schedules and word lists. In D. G. Payne & F. G. Conrad (Eds.), *Intersections in basic and applied memory research* (pp. 215-230). Mahwah, NJ: Erlbaum.

IZAWA, C. (1992). Test trials contributions to optimization of learning processes: Study/test trials interactions. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in honor of William K. Estes: Vol. 2. From learning processes to cognitive processes* (pp. 1-33). Hillsdale, NJ: Erlbaum.

KUO, T. M., & HIRSHMAN, E. (1996). Investigations of the testing effect. *American Journal of Psychology*, **109**, 451-464.

LACHMAN, R., & LAUGHERY, K. R. (1968). Is a test trial a training trial in free recall learning? *Journal of Experimental Psychology*, **76**, 40-50.

LANDAUER, T. K., & BJORK, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues. Vol. 2: Clinical and educational implications* (pp. 625-632). New York: Academic Press.

MCDANIEL, M. A., & EINSTEIN, G. O. (2005). Material appropriate difficulty: A framework for determining when difficulty is desirable for improving learning. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications* (pp. 73-85). Washington, DC: American Psychological Association.

MCDANIEL, M. A., & FISHER, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, **16**, 192-201.

MOZER, M. C., HOWE, M., & PASHLER, H. (2004). Using testing to enhance learning: A comparison of two hypotheses. In K. Forbus, D. Gentner, & T. Reiger (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 975-980). Mahwah, NJ: Erlbaum.

NUNGESTER, R. J., & DUCHASTEL, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology*, **74**, 18-22.

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

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

ROEDIGER, H. L., III, & MARSH, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 1155-1159.

RUNQUIST, W. N. (1983). Some effects of remembering on forgetting. *Memory & Cognition*, **11**, 641-650.

WHEELER, M. A., EWERS, M., & BUONANNO, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory*, **11**, 571-580.

WHEELER, M. A., & ROEDIGER, H. L., III (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, **3**, 240-245.

WHITTEN, W. B., & BJORK, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning & Verbal Behavior*, **16**, 465-478.

WITTMAN, W. T., & HEALY, A. F. (1995). A long-term retention advantage for spatial information learned naturally and in the laboratory. In A. F. Healy & L. E. Bourne, Jr. (Eds.), *Learning and memory of knowledge and skills: Durability and specificity* (pp. 170-205). Thousand Oaks, CA: Sage.

(Manuscript received April 17, 2006;
revision accepted for publication July 19, 2006.)