

Tests Enhance the Transfer of Learning

Doug Rohrer

Kelli Taylor

Brandon Sholar

University of South Florida

To appear in *Journal of Experimental Psychology: Learning, Memory, & Cognition*

Acknowledgements

Address correspondence to Doug Rohrer, Psychology, PCD4118G, University of South Florida, Tampa, FL 33620, USA. E-mail: drohrer@cas.usf.edu. We thank Sean Kang for his invaluable help with the literature review. This research was supported by the Institute of Education Sciences, U.S. Department of Education (R305B070537). The opinions expressed are those of the authors and do not represent the views of the Institute of Education Sciences.

Abstract

Numerous learning studies have shown that if the period of time devoted to studying information (e.g., casa-house) includes at least one test (casa-?), performance on a final test is improved – a finding known as the *testing effect*. In most of these studies, however, the final test is identical to the initial test. If the final test that requires a novel demonstration of learning (i.e., transfer), prior studies suggest that a greater degree of transfer reduces the size of the testing effect. We tested this conjecture. In two experiments, fourth- or fifth-grade students learned to assign regions or cities to map locations and returned one day later for two kinds of final tests. One final test required exactly the same task seen during the learning session, and the other final test consisted of novel, more challenging questions. In both experiments, testing effects were found for both kinds of final tests, and the testing effect was no smaller, and actually slightly larger, for the final test requiring transfer.

Studies have shown that material is better remembered if the period of time devoted to learning includes one or more tests on the material – a finding known as the *testing effect* (for a review, see Roediger & Karpicke, 2006). The testing effect is demonstrated by comparing the efficacy of two learning procedures: a *test-study* condition that combines at least one test (e.g., casa-?) with opportunities to study (e.g., casa-house), and a *study-only* condition devoted solely to study. A testing effect is obtained if performance on a final test is greater in the test-study condition than in the study-only condition.

Transfer

With very few exceptions, however, testing effect studies rely on final test questions that are identical to the initial test questions, and this raises a troubling question: does the testing effect diminish in size if the final test requires subjects to do more than merely answer the same question they have already seen in the test-study condition? In more formal terms, the present question of interest is whether the testing effect will dissipate if the final test requires a novel demonstration of learning known as *transfer* (e.g., Salomon & Perkins, 1989). The importance of transfer is not easily overstated;

indeed, transfer is sometimes said to be the aim of learning. In brief, if the need to demonstrate transfer reduces the size of the testing effect, the benefits of test-enhanced learning strategies are far more limited than previously suggested.

One theoretical account of the testing effect specifically predicts that the size of the testing effect will diminish as the final test requires a greater degree of transfer. By this account of the testing effect, study alone is less effective than a combination of study and test because this initial test provides *transfer-appropriate processing* (e.g., Morris, Bransford, & Franks, 1977), a theoretical view holding that successful retrieval is more likely if the cognitive processes invoked during encoding are similar to those invoked during retrieval. In other words, this account attributes the testing effect to the similarity between the learning strategy (initial test) and the manner of assessment (final test). Indeed, in many testing effect studies, the initial test and final test are *identical*. By the same reasoning, if the final test requires transfer, which inherently introduces a difference between the initial and final tests, the transfer appropriate processing account predicts a concomitant reduction in the size of the testing effect.

Though parsimonious, the transfer-appropriate processing account of the testing effect is at odds with several findings. For instance, McDaniel and Masson (1985) reported a series of studies in which subjects received semantic or phonemic cues during both the initial and final tests, and recall on the final test was better when the initial test cues and final test cues were *mismatched* (e.g., phonemic/semantic) rather than *matched* (e.g., semantic/semantic). Likewise, in several studies with a manipulation of the test format (e.g., recognition vs. recall), a match of initial and final test formats did *not* necessarily optimize final test scores (e.g., Carpenter & DeLosh, 2006; Kang, McDermott, & Roediger, 2007).

An alternative explanation of the testing effect attributes the testing effect to the act of retrieval required by the initial test, a view supported by several lines of evidence. For instance, the testing effect is reduced or even eliminated when the initial test consists of multiple-choice questions (which require little or no retrieval) rather than short-answer questions (e.g., Kang et al., 2007). Also, several studies (from outside the testing effect literature) have shown that final test performance is greater when the initial test led to *greater* retrieval difficulty. For instance, in a study reported by Gardiner, Craik, and Bleasdale (1973), college students heard definitions and tried to recall the defined words, and the words requiring longer response times were *more* likely to be recalled on a subsequent free recall test. Similarly, in two studies reported by Pyc and Rawson (2009), college students learned Swahili-English pairs by repeatedly cycling through the list of Swahili words and trying to recall each word's English translation, and a manipulated increase in the duration of time between successive presentations of the same Swahili word, which presumably increased retrieval difficulty, produced greater recall on the final test. Exactly why retrieval might benefit subsequent retrieval is less clear, though. One possibility is that the combination of study and retrieval provides greater encoding variability than does study alone, an account put forth by McDaniel and Masson (1985). By this explanation, test-study strategies ensure that encoding occurs in different contexts (i.e., study and test), and, metaphorically, this increases the number of effective retrieval routes.

Retrieval-based accounts appear to predict that the size of the testing effect does *not* depend on the degree of transfer required by the final test, although one might argue instead that these accounts simply do not make any prediction in this case. It is difficult to be sure because these accounts are not highly specified. Still, retrieval-based accounts do *not*

predict a decline in the size of testing effect resulting from increasing transfer demands, unlike the transfer-appropriate processing view.

Testing Effect Studies with a Final Test Requiring Transfer

In our search of the literature, we defined transfer as broadly as possible so that our search would err on the side of inclusion. Most notably, we included studies in which the initial and final tests required the same association but in a different order (i.e., $A \rightarrow B$ during initial test, $B \rightarrow A$ during final test). However, we excluded testing effect studies in which the same question (e.g., “Who assassinated Abraham Lincoln?”) appeared as a short-answer question on the initial test and reappeared as a multiple-choice question during the final test (or vice-versa). Finally, for a study to be included, the initial and final tests, though necessarily different, must have assessed the *same* fact or concept. (This inclusion criterion eliminated two nominally related studies: Chan, McDermott, & Roediger, 2006; Foos & Fisher, 1988.) With these search criteria, we found three relevant studies.

One of these studies, reported by McDaniel, Anderson, Derbish, and Morrisette (2007), was specifically designed to assess transfer. For this experiment, which was conducted in conjunction with an online college psychology course, test questions were written in pairs, and the two questions within a pair were based on the same fact. For example, the fact, “All preganglionic axons release acetylcholine,” was the focus of the following two questions (p. 499):

1. All preganglionic axons, whether sympathetic or parasympathetic, release ___ as a neurotransmitter.

[Answer: acetylcholine]

2. All ___ axons, whether sympathetic or parasympathetic, release acetylcholine as a neurotransmitter.

[Answer: preganglionic]

Thus, these two questions represent a forward and reverse association (i.e., $A \rightarrow B$ vs. $B \rightarrow A$), but the associations are embedded with complexity that provides ecological validity. One question within each pair appeared on the initial test given in the test-study condition, and the other question within each pair appeared on the final test seen a few weeks later. (This study included several other dimensions not described here.) Final test scores revealed a statistically significant testing effect, which is to say that tests enhanced learning even though the final test required transfer.

But does the size of the testing effect diminish with the degree of transfer required by the final test? This question was the focus of two nearly identical experiments reported by Carpenter, Pashler, and Vul (2006). Subjects studied associated word pairs (e.g., chain-fence) by either a study-only procedure ($A-B$) or a test-study procedure ($A-?$). One day later, subjects sat for one of the four kinds of final tests. One of these final tests was identical to the initial tests ($A-?$), but transfer was required by the remaining three kinds of tests: reverse association ($?-B$), recall of all A words, or recall of all B words. For all four kinds of tests, final test scores were greater in the test-study condition than in the study-only condition, but it appears that the necessity of demonstrating transfer reduced the size of this testing effect. Specifically, although the authors did not separately analyze the results for each kind of test, the difference between the mean final test scores achieved in the test-study and study-only conditions was largest for the final test requiring no transfer ($A-?$). However, large differences in means do not always translate to large effect sizes (because effect size depends on variability), and, for this reason, this characterization of these data is only tentative.

Finally, a similar finding was reported by Agarwal, Roediger, McDaniel, and McDermott (May, 2008). This study was one in a series of not-yet-published studies in

which middle school students received an initial test on material presented during an immediately previous class meeting. The final test was given as much as eight months later. In one of these studies, which took place in a science class, both the initial test and the final test included “definition” and “application” questions. Performance on each kind of final test question was best if the corresponding initial test question was the *same* kind of question. Moreover, if the initial test question and final test question were *different* kinds, final test performance in the test-study condition was scarcely any better than that in the study-only condition. However, this characterization of these findings is again based solely on a comparison of mean test scores, because the results of statistical tests were not provided in the reprint of the conference presentation (which was replete with findings from several studies). Thus, a comparison of effect sizes might tell a different story.

In summary, a testing effect has been observed when a final test requires transfer (McDaniel et al.), but, in studies with a manipulation of transfer (Carpenter et al., Agarwal et al.), the size of the testing effect seemingly diminishes as the degree of required transfer increased. This latter interpretation is only tentative, though, because it is based on a comparison of means rather than effect sizes. In brief, although the relevant data are ostensibly consistent with the possibility that the testing effect is diminished when the final test requires transfer, this conclusion is only conjecture.

We tested this conjecture in two experiments. Subjects in each experiment completed two kinds of final tests. One required subjects to do exactly what they had done during the learning session, and one required transfer. Rather than rely on initial test questions and final transfer test questions of equal difficulty (e.g., $A \rightarrow B$ vs. $B \rightarrow A$), the transfer test was inherently more challenging than the initial test. Effect sizes

were calculated for each kind of final test in order to assess whether the size of the testing effect depended on the degree of transfer required by the final test. Finally, because most testing effect studies rely solely on adult subjects, we chose to use child samples (ages 10 – 12) in both experiments in order to assess whether tests, which have been championed as an underutilized learning strategy, enhance the learning of young students.

Experiment 1

In the first study, fourth-grade students learned to assign region names to map locations by either a test-study (TS) procedure or a study-only (SO) procedure, and they returned one day later for two final tests. The so-called standard final test required subjects to perform the same task they had practiced during the learning session: assign each of 10 presented region names to its correct location. On the transfer final test, however, region names were not provided.

Method

Subjects. Both sessions were completed by 28 fourth-grade students (57% girls) at a private grammar school (K-8) in St. Petersburg, Florida. All were 9 or 10 years of age ($M = 9.1$, $SD = 0.26$).

Materials. Each of two fictional maps included 20 regions and 10 named regions (Figure 1). We included 10 unnamed regions so that subjects could not rely heavily on the process of elimination. Two additional maps similar to those in Figure 1 served as sample maps for use during tutorials. No region name appeared on more than one of the four maps.

Procedure. Subjects attended a learning session and returned for a final test one day later. During the learning session, subjects participated in groups of two to four subjects. A computer-driven audiovisual presentation was projected onto a large screen, and subjects wrote their responses in individual booklets. Subjects first read instructions and then completed both a TS and SO phase. Both the order of the TS and SO stages and the pairing

of condition and map were counterbalanced, so that each subject was randomly assigned to one of four different schedules. The TS stage and SO stage each began with a tutorial, followed by the opportunity to practice the learning procedure with a sample map and then take a sample test requiring subjects to assign region names to locations.

Shortly after the test on the sample map (within the TS and SO stage), subjects saw one of the two scored maps. They first observed an “initial exposure” cycle in which the unlabeled map remained onscreen while each region name appeared one at a time (3 s) in its correct location. Subjects then cycled five times through the 10 regions (6 s per region) using either the TS or SO learning procedure. The order in which the regions appeared varied systematically across cycles so that no two regions appeared consecutively in more than one cycle. The map boundaries remained onscreen throughout each cycle.

For the TS procedure, the 6-s duration devoted to each region included a 4-s test phase and 2-s study phase, and a tone signaled the start of each phase (Figure 2A). At the outset of each test phase, a region name appeared onscreen just *above* the map, adjacent to a numeral indicating the serial position within that particular cycle. For example, if Bond was the seventh region to appear within a cycle, “7” and “Bond” appeared above the onscreen map (Figure 2A), prompting subjects to write “7” in what they believed was the correct region within the unlabeled map in their booklet. During the immediately following study phase, the region name appeared onscreen in its correct region, and subjects checked their answer but did not write anything. Each booklet page included one unlabeled map, and subjects wrote the 10 responses for each cycle on the same map before turning to the next page at the outset of the next cycle.

With the SO procedure, a region name appeared onscreen in its correct location throughout the *entire* 6-s duration devoted to

that region (Figure 2B). Thus, subjects were able to view the correct location of the given region *while* they wrote the numeral on their booklet map. Otherwise, the SO and TS procedures were identical.

Subjects returned one day later for a standard final test and a transfer final test on each map (2 min each). Both transfer final tests were administered first, followed by both standard final tests. For both kinds of final test, one half of the subjects first saw the map learned with the TS procedure. Thus, the two final tests for a particular map were never immediately successive. The one-page transfer final test included an unlabeled map, and subjects were asked to write each region name in its correct location. The standard final test was the same as the transfer final test except that the 10 region names appeared in a column adjacent to the unlabeled map.

Results and Discussion

Learning. The mean proportion of region names assigned to the correct location during each learning cycle is shown in Figure 3. Averaged across all cycles, SO accuracy was nearly perfect and far greater than TS accuracy (98% vs. 59%), $t(27) = 8.96$, $p < .001$, $d = 1.69$. TS accuracy improved with each subsequent cycle, yet SO accuracy nevertheless exceeded TS accuracy on the final cycle as well (98% vs. 73%), $t(27) = 4.58$, $p < .001$, $d = 0.86$.

Final Tests. A testing effect was observed for both final tests (Figure 3). On the transfer final test, the test-study procedure enhanced learning by more than a factor of two (26% vs. 11%), $t(27) = 4.05$, $p < .01$, $d = 0.76$. For the standard final test, the testing effect was slightly smaller (56% vs. 34%), $t(27) = 3.36$, $p < .01$, $d = 0.64$. (For both kinds of final tests, the testing effect was larger if the TS learning procedure preceded, rather than followed, the SO learning procedure, but this order effect was not reliable.) In brief, the size of the testing effect, as measured by Cohen’s d , was no smaller – and actually larger – for the final test requiring transfer.

Experiment 2

The second experiment was nearly identical to the first, as the subjects learned to assign city names to map locations, and they returned one day later for a standard final test and a transfer final test. The transfer task, however, was more akin to conventional assessments of transfer. Specifically, the transfer test required subjects to identify the city they would pass through as they drove along the shortest possible route between two given cities.

Method

Subjects. Both sessions were completed by 28 students (57% girls) in the fourth grade ($n = 17$) or fifth grade ($n = 11$) of a private grammar school (K-8) in St. Petersburg, Florida. We recruited both fourth- and fifth-grade students in order to obtain enough subjects. (The subjects in Experiment 1 attended a different school.) Subjects were 9, 10, or 11 years of age ($M = 10.1$, $SD = 0.79$). One additional student completed the first session but not the second, and her data were excluded from all analyses.

Materials. Each of two maps depicted 15 cities, including 10 named cities, along with the roads connecting these cities (Figure 4). The maps depict the location of cities and roads in areas within Iraq (panel A) and Afghanistan (panel B), but each original city name was replaced by a 4-letter name beginning with the same letter. For example, Boyd replaced Baghdad. We also created two fictional maps to serve as sample maps.

Procedure. The procedures differed from that of Experiment 1 in minor ways. For both the TS and SO procedure, the duration devoted to each location was increased from 6 s to 7 s because the subjects in Experiment 1 appeared to be flustered by the pace. The 7-s period was divided into a 5-s test phase and a 2-s study phase in the TS procedure (Figure 5). For the final tests, the standard final test immediately preceded the transfer final test on the same map. For the standard final test, subjects

received an unlabeled map and an adjacent list of the 10 city names, and they attempted to write each city name in its correct location. For the transfer final test, subjects received the same map seen during the standard test, along with five questions like the following: “If you drive from Ross to Boyd along the shortest possible path, which city do you drive through?” (Answer: Ford). Each of the 10 city names appeared in exactly one of the five questions. Immediately prior to the first of the two transfer tests, subjects observed a brief computer-driven visual presentation that provided instructions about the transfer task.

Results and Discussion

Learning. The mean proportion of city names assigned to the correct location during each learning cycle is shown in Figure 6. Averaged across all cycles, SO accuracy exceeded TS accuracy (99% vs. 55%), $t(27) = 12.28$, $p < .001$, $d = 2.32$. TS accuracy increased with each subsequent cycle, but SO accuracy was nevertheless superior on the final cycle (99.6% vs. 71%), $t(27) = 5.81$, $p < .001$, $d = 1.10$.

Final Tests. A testing effect was observed for both final tests (Figure 6). On the standard final test, the initial tests enhanced learning by about a third (58% vs. 42%), $t(27) = 2.86$, $p < .01$, $d = 0.54$. On the transfer final test, the initial test nearly doubled performance (47% vs. 25%), $t(27) = 3.03$, $p < .01$, $d = 0.57$. (Averaged across conditions and final tests, fifth-grade students outscored fourth-grade students [50% vs. 39%]. However, the final test score *difference* between the TS and SO conditions was about the same for both fifth- and fourth-grade students [21% vs. 17%, $t < 1$]. Also, as in Experiment 1, order effects were not statistically significant.) The take-home story is the same as that in Experiment 1: the size of the testing effect was unaffected when transfer was required.

General Discussion

As detailed in the Introduction, the studies reported here were motivated by the

troubling possibility that the size of the testing effect might fade if the final test included novel questions rather than questions previously seen only in the test-study condition. By this possibility, the practical utility of the testing effect would be limited to scenarios in which the precise nature of the final assessment was known in advance. Far from being a straw man hypothesis, this conjecture was suggested by previous findings, as detailed in the Introduction. However, this possibility is entirely inconsistent with the results presented here. In Experiments 1 and 2, the testing effect for the final test requiring transfer was no smaller, and actually slightly larger, than that observed for the final test requiring no transfer. Finally, the present findings also demonstrated that tests can enhance the learning of young children – a nontrivial finding in light of the potential educational applications of the testing effect.

Children

Only a handful of studies have assessed the effects of tests on learning by children, yet nearly all have found positive effects of tests. For instance, Gates (1917), in a study of “pupils from homes of business men and artisans of moderate means” (p. 24), found that reading alone was inferior to a combination of reading and recitation (though the effect disappeared for students younger than eight years of age). Glover (1989) also reported two studies that purportedly showed a testing effect with seventh grade students, but that study conflated the effects of testing and spacing.

Very recently, two studies observed testing effects with children in the classroom. In a study reported by Carpenter, Pashler, and Cepeda (2009), review questions improved eighth-grade students’ recall of material they had learned in their U.S. history course. Similarly, in the previously cited series of studies with students in grades six through eight (Agarwal et al., May, 2008), tests provided a variety of benefits to students in

science and social studies courses. In these studies, the learning material was presented by the subjects’ regular teacher, and retention intervals were as long as eight months. In summary, the literature provides nearly uniform support for the use of test-enhanced learning strategies for children older than eight years of age.

Map Learning

It appears that only one prior testing effect study used visual-spatial materials, and that experiment, reported by Carpenter and Pashler (2007), also used a map learning task. College students studied maps with 12 features, such as a lake or golf course. During the study-only procedure, subjects studied the map for 120 s. The equally long test-study procedure required subjects to cycle through repeated presentations of the map, each with one feature missing, and attempt to “covertly retrieve” the missing feature (i.e., provide no overt response). One half hour later, subjects tried to recall the location of each feature, and a testing effect was observed.

The results of Carpenter and Pashler, as well as those reported here, provide empirical support for the test-enhanced map learning activities found on several educational websites. For example, in the game *Geospy*, which appears on a website sponsored by National Geographic, users learn the locations of countries in a particular continent by trying to recall the location of each country. With Europe, for instance, each of 37 country names appears one at a time, in a random order, to the left of an unlabeled map of Europe. As each name appears, users select a region with a pointing device, and an incorrect response immediately triggers the shading of the correct region. Total errors and total response time are provided so that users can track improvement in accuracy and speed. We informally asked a few students to play this game, and all reported that it was more interesting than a study-only procedure.

Generality and Utility of the Testing Effect

The results reported here suggest that the benefits of test-enhanced learning are not compromised when transfer is required, and this represents another demonstration of the generality of the testing effect. Indeed, it appears that only one notable boundary condition has been identified: an initial test consisting of multiple-choice questions sometimes often fails to produce a testing effect, presumably because such questions require little or no retrieval (e.g., Kang et al., 2007). Moreover, the benefits of test-enhanced learning have been demonstrated in settings with high ecological validity (e.g., Butler & Roediger, 2007; Carpenter et al., 2009; McDaniel et al., 2007; Metcalfe, Kornell, & Son, 2007). Moreover, tests can improve learning by a dramatic degree (e.g., Karpicke & Roediger, 2008).

Nevertheless, test-enhanced learning strategies remain underutilized except in disciplines that inherently require retrieval. Mathematics problems, for example, intrinsically require retrieval of previously learned procedures, and writing requires one to recall the rules of syntax and the proper spelling of words (with immediate corrective feedback provided by word processing software). Yet the use of test-enhanced learning is far more limited in courses requiring students to little other than listen to lectures and read outside of class. In these disciplines, the benefits of test-enhanced learning have yet to be fully exploited.

References

- Agarwal, P. K., Roediger, H. L., McDaniel, M. A., & McDermott, K. B. (May, 2008). *Improving student learning through the use of classroom quizzes*. Poster presented at the 20th Annual Meeting of the Association for Psychological Science, Chicago, IL.
- Butler, A.C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*, 514–527.
- 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. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14*, 474–478.
- Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U. S. history facts. *Applied Cognitive Psychology, 23*, 760-771.
- 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.
- Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*, 553-571.
- Foos, P. W., & Fisher R. P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology, 80*, 179-183.
- Gardiner, J. M., Craik, F. I. M., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition, 1*, 213–216.
- Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*(40), 1–104.
- Glover, J. A. (1989). The “testing” phenomenon: Not gone but nearly

- forgotten. *Journal of Educational Psychology*, 81, 392–399.
- Kang, S. H. K., McDermott, K. B. & Roediger, H. L. (2007). Test format and corrective feedback modulate the effect of testing on memory retention. *The European Journal of Cognitive Psychology*, 19, 528–558.
- Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, 319, 966–968.
- 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., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 11, 371–385.
- Metcalf, J., Kornell, N., & Son, L.K. (2007). A cognitive-science based programme to enhance study efficacy in a high and low-risk setting. *European Journal of Cognitive Psychology*, 19, 743–768.
- Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behavior*, 16, 519–533.
- Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, 60, 437–447.
- Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181–210.
- Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. *Educational Psychology*, 24, 113–142.

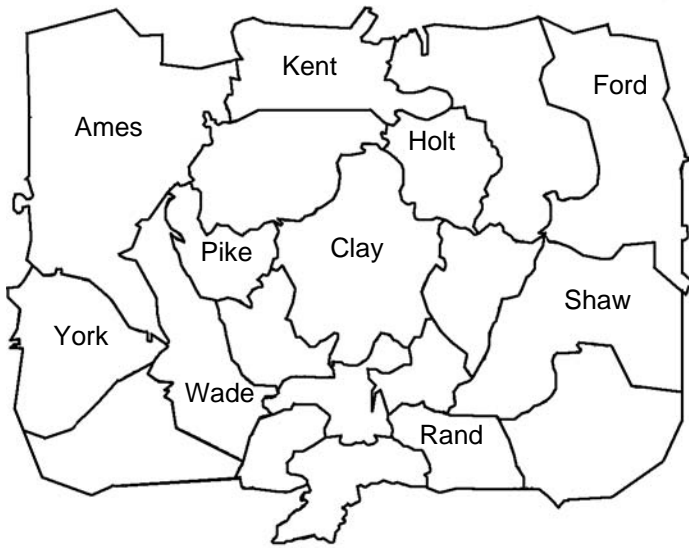
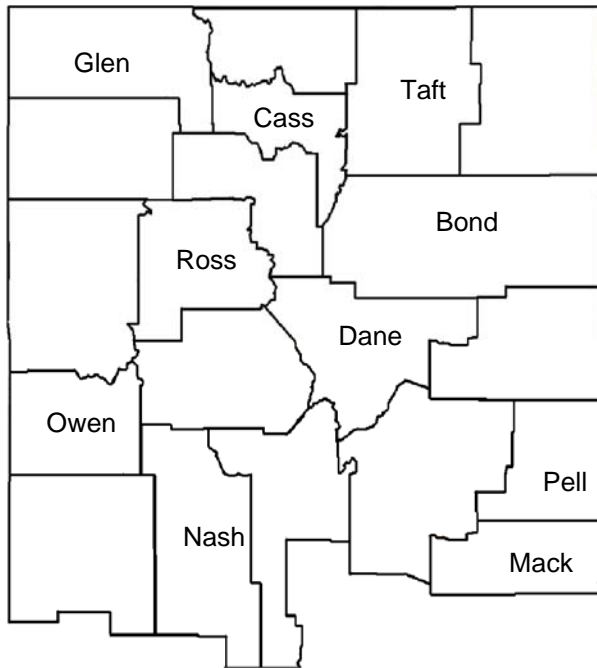
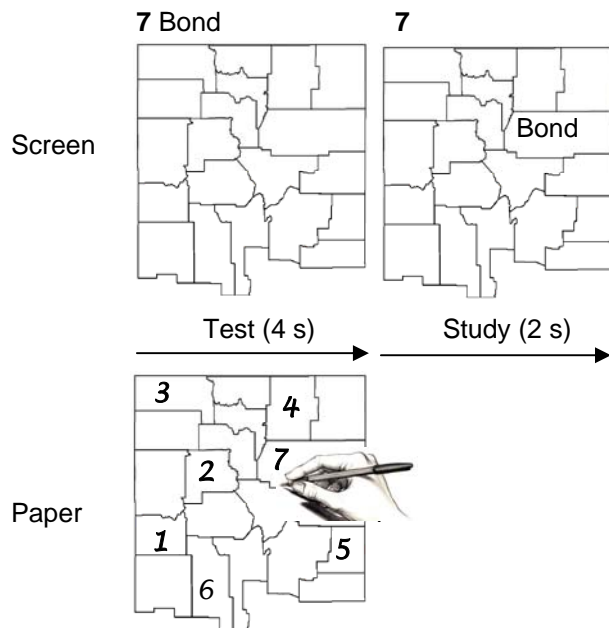


Figure 1. Maps for Experiment 1. Each map includes 20 regions and 10 named regions.

Figure 1

A Test-Study (TS)



B Study-Only (SO)

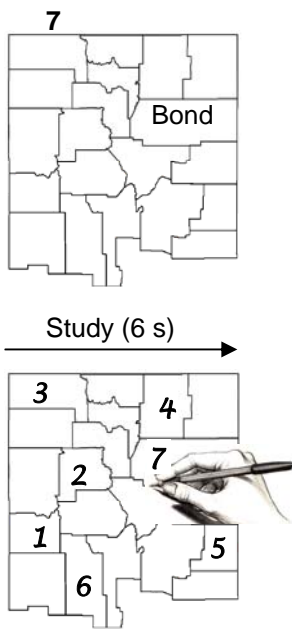


Figure 2. Learning Procedure for Experiment 1. Children learned to assign each region name to its correct location by repeatedly cycling through the list of 10 regions. In the cycle illustrated here, Bond is the seventh region to appear. In the test-study condition (A), subjects wrote a “7” in the region of their booklet map they believed was named Bond *before* the name appeared onscreen in its correct location. In the study-only condition (B), the name “Bond” appeared onscreen in its correct region *while* subjects wrote a “7” in the same region of their booklet map.

Figure 2

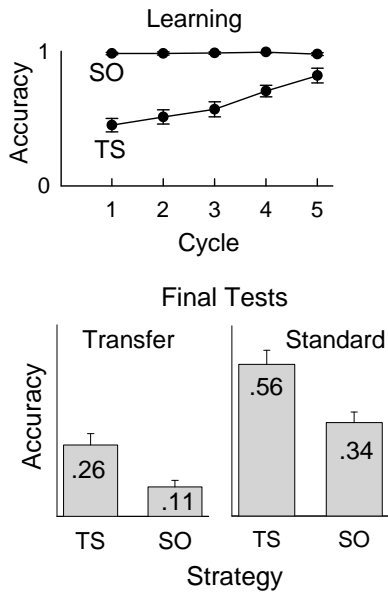
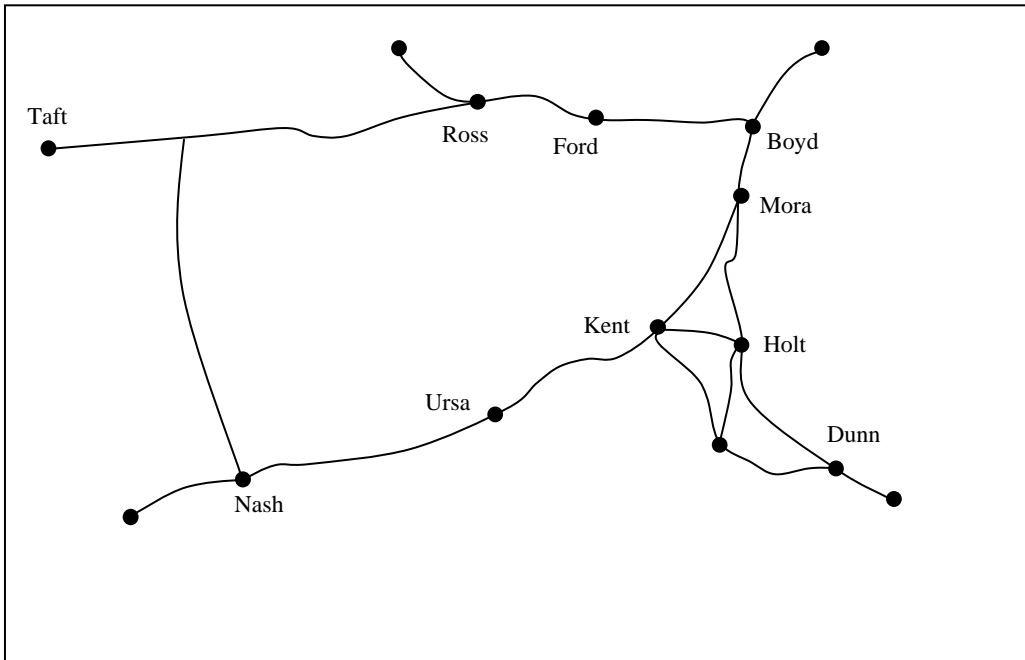


Figure 3. Results of Experiment 1. One day after the learning session, subjects sat for a transfer final test and a standard final test (in that order) for each map. For the transfer final test, subjects were asked to write each region name in its correct location within an unlabeled map. The standard final test was the same as the transfer final test except that the 10 region names were provided.

Figure 3

A



B

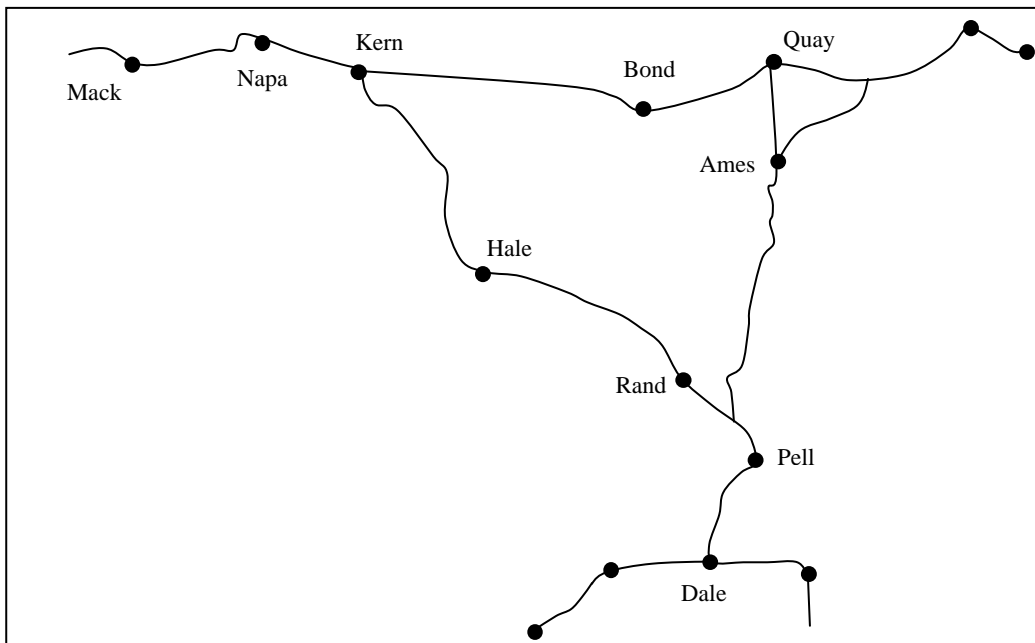
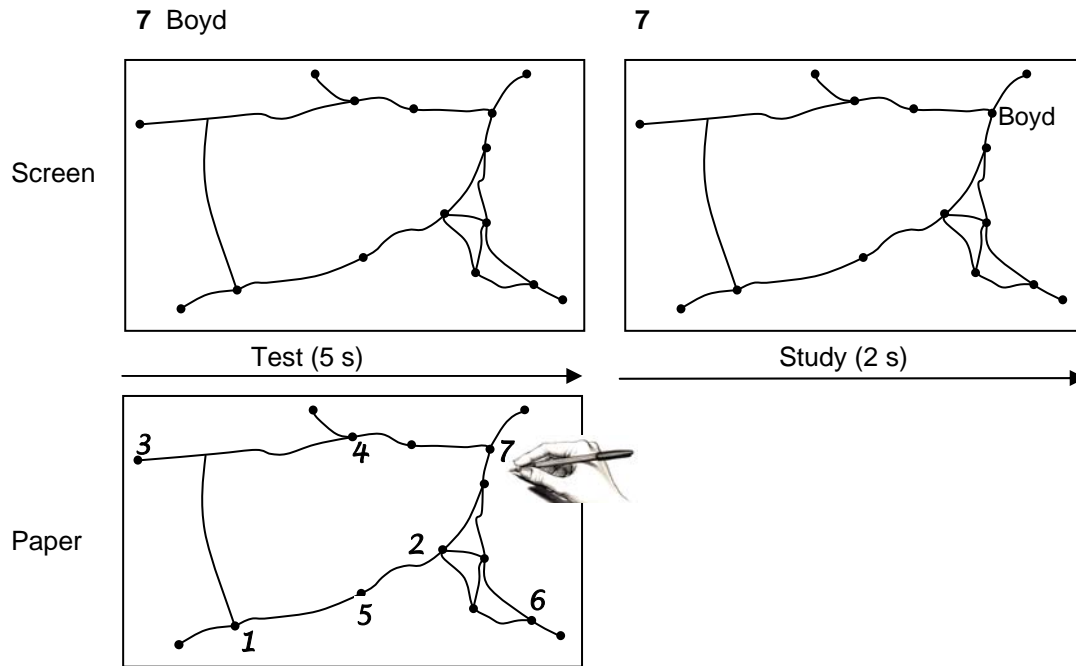


Figure 4. Maps for Experiment 2. Each map includes 15 cities and 10 named cities. The maps depict areas within central Iraq (A) and northern Afghanistan (B), but the original city names were replaced. For example, Boyd replaced Baghdad.

Figure 4

A Test-Study (TS)



B Study-Only (SO)

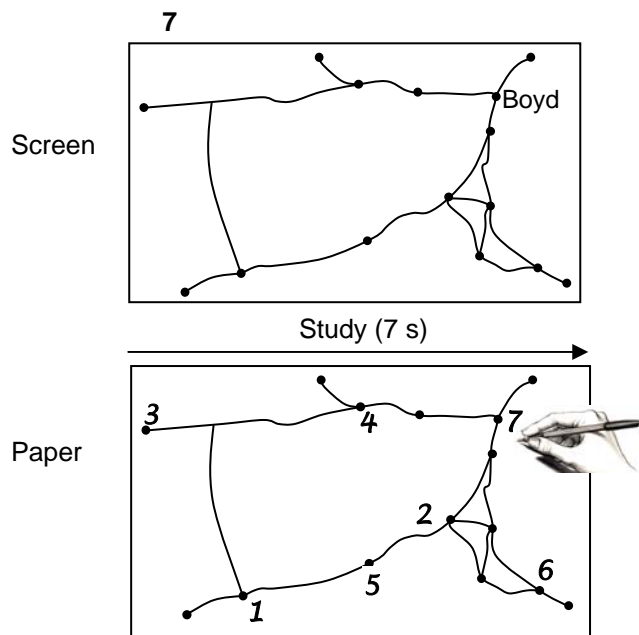


Figure 5. Learning Procedure for Experiment 2. Children learned to assign each city name to its correct location by repeatedly cycling through the list of 10 cities. In the cycle illustrated here, Boyd is the seventh city to appear. In the test-study condition (A), subjects wrote a “7” in the location on their booklet map they believed was named Boyd *before* the name appeared onscreen in its correct location. In the study-only condition (B), the name “Boyd” appeared onscreen in its correct location *while* subjects wrote a “7” in the same location of their booklet map.

Figure 5

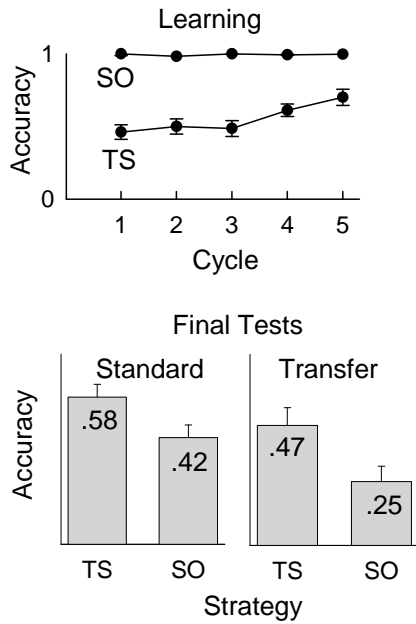


Figure 6. Results of Experiment 2. One day after the learning session, subjects sat for a standard final test and a transfer final test (in that order) for each map. For the standard final test, subjects received an unlabeled map and a list of the 10 city names, and they tried to write each city name in its correct location. For the transfer final test, subjects received the same map seen during the standard test and five questions like the following: “If you drive from Ross to Boyd along the shortest possible path, which city do you drive through?” The size of testing effect was larger for the test requiring transfer. Error bars depict one standard error.

Figure 6