Instr Sci (2007) 35:481–498
DOI 10.1007/s11251-007-9015-8

The shuffling of mathematics problems improves learning

Doug Rohrer · Kelli Taylor

Received: 29 August 2006 / Accepted: 3 January 2007 / Published online: 19 April 2007
© Springer Science+Business Media, Inc. 2007

Abstract In most mathematics textbooks, each set of practice problems is comprised almost entirely of problems corresponding to the immediately previous lesson. By contrast, in a small number of textbooks, the practice problems are systematically shuffled so that each practice set includes a variety of problems drawn from many previous lessons. The standard and shuffled formats differ in two critical ways, and each was the focus of an experiment reported here. In Experiment 1, college students learned to solve one kind of problem, and subsequent practice problems were either massed in a single session (as in the standard format) or spaced across multiple sessions (as in the shuffled format). When tested 1 week later, performance was much greater after spaced practice. In Experiment 2, students first learned to solve multiple types of problems, and practice problems were either blocked by type (as in the standard format) or randomly mixed (as in the shuffled format). When tested 1 week later, performance was vastly superior after mixed practice. Thus, the results of both experiments favored the shuffled format over the standard format.

Keywords Mathematics · Practice · Distribute · Mass · Block · Mix · Interleave · Spacing

Introduction

The effort to improve mathematics learning has focused primarily on the manner in which material is taught, with far less attention given to the role of practice problems. Yet, for many students, the majority of their mathematics learning effort is devoted to practice problems (rather than, say, reading). While many aspects of practice are worthy of investigation, the two experiments presented here focused primarily on the effects of varying either the temporal distribution of practice problems or the order in which

D. Rohrer (✉) · K. Taylor
Department of Psychology, PCD 4118G, University of South Florida, Tampa, FL 33620, USA
e-mail: drohrer@cas.usf.edu

problems are solved. Neither manipulation required an increase in the total number of practice problems, yet both experiments revealed large boosts in subsequent test performance. That is, merely altering the timing of practice led to large gains in test performance.

The arrangement of practice problems in most mathematics textbooks is one that most readers will recognize. Each set of practice problems, or *practice set*, consists almost entirely of problems corresponding to the immediately preceding lesson (e.g., Glencoe, 2001). For example, a lesson on the addition or subtraction of fractions (e.g., 5/6–4/5) is followed immediately by perhaps a few dozen problems, all of which require the addition or subtraction of fractions. In brief, each set of practice problems is devoted to the most recent lesson. Moreover, problems of the same type are usually in blocks (e.g., 12 fraction addition problems, followed by 12 fraction subtraction problems). This format also is the modal format of computer-aided instructional packages, and, therefore, the data reported herein apply to this instructional medium as well.

The standard practice format has two features that are examined here. First, most or all of the problems relating to a given lesson are concentrated or *massed* into the immediately following practice set instead of being distributed or *spaced* across multiple practice sets. For example, in the standard format, virtually all of the quadratic formula problems within the textbook appear in the practice set that appears immediately after the lesson on the quadratic formula. The second feature of the standard format is that the problems within each practice set are usually *blocked* by topic and not *mixed* across topics. For example, after a lesson explaining how to find the least common multiple and the greatest common factor of two integers, a practice set includes a block of least common multiple problems followed by a block of greatest common factor problems. Notably, it is possible for a textbook to use massed practice but not blocked practice, but, in our experience, these two features usually co-occur.

By contrast, a very small number of mathematics textbooks use what we call a shuffled format (e.g., Saxon, 1997). A textbook with a shuffled format may have lessons identical to those in the standard format, and moreover, the two formats need not differ in either the number of practice sets within the text or the number of practice problems per practice set. But, with the shuffled format, the practice problems are systematically arranged so that practice problems are both distributed and mixed. For example, after a lesson on the quadratic formula, the immediately following practice set would include no more than a few quadratic formula problems, with other quadratic formula problems appearing in subsequent practice sets with decreasing frequency. Thus, the practice problems of a given type are systematically spaced throughout the textbook. This spacing intrinsically ensures that the problems within each practice set include a mixture of different types, as there are no more than one or two practice problems of each kind within each practice set. In order to achieve such variety in the early portion of the textbook, the first several practice sets can include problems relating to topics covered in previous *years*.

In summary, virtually all mathematics textbooks use one of two formats that differ with regard to two variables. First, the problems of a given type are either massed in a single practice set (as in the standard format) or spaced across multiple practice sets (as in the shuffled format). Second, problems of different types are either blocked by type (as in the standard format) or randomly mixed (as in the shuffled format). The massed vs. spaced variable was examined in Experiment 1, and the blocked vs. mixed variable was examined in Experiment 2. A third variable—light versus heavy massed practice—was also examined in Experiment 1, for reasons described below. The remainder of the Introduction is devoted to the relevant literature.

Massed versus spaced practice

In an experiment comparing the benefits of massed and spaced practice, a given amount of practice is either massed into a single session or spaced across multiple sessions. For example, four practice problems (relating to the same skill or concept) might be assigned in a single session or divided evenly across two sessions separated by 1 week. The *retention interval* equals the period of time between the *last* practice problem and the test. For example, if a skill is practiced on Monday and Tuesday and tested on Friday, the retention interval equals 3 days.

Test performance is generally superior after practice that is spaced rather than massed—a finding known as the *spacing effect* (e.g., Bahrick, Bahrick, Bahrick, & Bahrick, 1993; Bjork, 1979, 1988, 1994; Bloom & Shuell, 1981; Carpenter & DeLosh, 2005; Reynolds & Glaser, 1964; Smith & Rothkopf, 1984). Exactly how spacing of practice produces this benefit is the focus of much unresolved debate (for a review, see Dempster, 1989), but, for the present purposes, it is sufficient to simply note that spaced practice boosts test performance. For this reason, many previous authors have advocated that learners space their study (Bahrick et al., 1993; Bjork, 1979, 1988, 1994; Bloom & Shuell, 1981; Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dempster, 1989; Pashler, Rohrer, Cepeda, & Carpenter, 2007; Reynolds & Glaser, 1964; Schmidt & Bjork, 1992; Smith & Rothkopf, 1984).

While only a few of the hundreds of spacing experiments have used mathematics tasks, these few findings have shown benefits of spacing mathematics practice. For instance, Smith and Rothkopf (1984) observed a spacing effect if several statistics lectures were spaced across 4 days rather than massed into one session. More recently, Rohrer and Taylor (2006) found a benefit of spacing mathematics practice for students who were tested 4 weeks after their last practice problem. Finally, Rea and Modigliani (1985) found a spacing effect with young children who were asked to memorize five multiplication facts (e.g., $8 \times 5 = 40$), although this kind of task is better described as verbal memory rather than mathematical learning (which is not to say that such facts are not sometimes useful). Incidentally, several mathematics learning experiments that purport to show a spacing effect were, in fact, confounded in favor of the spacing effect. In Grote (1995), for instance, students either massed their practice on Day 1 or spaced their practice across Days 1 through 22, but every student was tested on Day 36. Thus, the spaced practice condition benefited from a far shorter retention interval. Nevertheless, the results of the few non-confounded studies support the view that the long-term retention of mathematical knowledge is enhanced by distributing the corresponding practice problems across multiple practice sessions. This effect is revisited in Experiment 1.

Light versus heavy massed practice

One explanation for the preponderance of massed practice within mathematics textbooks is the oft-cited belief that material is retained longer if study or practice continues immediately after the material is understood. This kind of massed practice is formally known as an *overlearning strategy*. For example, after a student has correctly solved one mathematics problem (or perhaps two problems of the same type in order to rule out the possibility that the first correct answer was due to chance), additional problems *of the same type, if attempted immediately*, constitute an overlearning strategy. It must be clarified, incidentally, that the term overlearning describes a *strategy* and not the degree of learning.

In fact, one can achieve a very high degree of learning without using an overlearning strategy. For example, most everyone has mastered the names of the calendar months, but few did so by the use of an overlearning strategy (i.e., immediate post-criterion practice). Thus, we are not evaluating the utility of knowing material very well but rather the utility of learning by the strategy of post-criterion practice.

Overlearning experiments include a condition that ensures overlearning and a condition in which overlearning is avoided or at least minimized. The great majority of these experiments have found that the overlearning condition produces greater subsequent test performance (e.g., Gilbert, 1957; Krueger, 1929; Postman, 1962), and such a benefit was confirmed by a meta-analysis reported by Driskell et al. (1992). In brief, although a few studies have found little or no benefit of overlearning (e.g., Reynolds & Glaser, 1964; Rohrer, Taylor, Pashler, Wixted, & Cepeda, 2005), most results find overlearning to boost subsequent test performance. These empirical findings perhaps explain the widespread support for overlearning as a learning strategy (e.g., Fitts, 1965; Foriska, 1993; Hall, 1989; Jahnke & Nowaczyk, 1998; Radvasky, 2006).

Yet there is reason to be cautious about the utility of overlearning in the mathematics classroom. Only one previous overlearning experiment has used a mathematics task, and it found no effect of overlearning on subsequent test performance. In an experiment reported by Rohrer and Taylor (2006), students learned a single procedure and then immediately worked either three or nine practice problems. The threefold increase in practice had no effect on test scores at either the 1-week or 4-week tests.

Thus, this single experiment raises the possibility that mathematics overlearning is a waste of time, and the implications of this finding are troubling because many mathematics assignments demand a large degree of overlearning. For example, in the standard (massed-blocked) format described at the outset of this Introduction, practice sets often include as many as a dozen or more problems of the same kind. Thus, if overlearning is ineffective, most mathematics students are devoting a sizeable proportion of their practice to a learning strategy with little or no benefit. The benefits of overlearning are revisited in Experiment 1.

Blocked versus mixed practice

Practice problems within mathematics textbooks are usually blocked by topic and not mixed together, as described at the outset of this Introduction, but there appears to be little direct evidence supporting either strategy for mathematics tasks. For motor tasks, the data suggest that subsequent test performance is greater after mixed practice (see Bjork, 1994, for a review). In Carson and Wiegand (1979), for instance, young children learned to throw bean bags of different weights at a target, and their subsequent test performance was greater when the practice throws for each particular weight were intermixed and not blocked by weight.

For mathematics learning, however, we are unaware of any experiments comparing mixed and blocked practice. Some previous studies have compared practice schedules that differ with regard to the extent of mixture, but these experimental comparisons have been confounded. For example, in an experiment reported by Mayfield and Chase (2002), one group of subjects relied on mixed, spaced practice while another group underwent blocked, massed practice. Thus, it was impossible to assess the specific effect of mixture. In Experiment 2 of the present paper, students are randomly assigned to either a mixed or blocked practice schedule, and the practice problems for both groups are spaced across two

sessions. This way, we were able to assess whether mixed practice provides benefits *above and beyond* the benefit of spaced practice.

There is good reason to expect that a mixture of problem types will benefit subsequent test performance. If a practice set includes a randomly arranged variety of problem types, students learn to pair each kind of problem with the appropriate procedure. In other words, a mixed practice schedule requires that students learn not only *how* to perform each procedure but also *which* procedure is appropriate for each kind of problem (e.g., Kester, Kirschner, & Van Merriënboer, 2004). For example, when a lesson on the repeated-measures *t*-test is followed immediately by a practice set comprised solely of repeated-measures *t*-test problems, the choice of procedure is obvious to students. Thus, they can complete this block of practice problems without learning why each problem requires this particular procedure. Consequently, when these students receive a repeated-measures *t*-test problem on a later exam that includes a variety of problem types, each requiring that they ''assess statistical significance,'' they are faced with a task they have not practiced: knowing *which* statistical test is appropriate for each type of problem. In fact, knowing which procedure is appropriate is arguably more important than knowing how to perform the procedure.

Learning to pair problem types and procedures is especially challenging in mathematics because different problem types are often superficially similar. For example, the solution of a single equation with a single variable is a rather narrow subset of problems, but even this subset of problem types subsumes different procedures. For example, the equation, $x^3-3x^2-2x = 0$, is solved by factoring the left-hand expression, but the equation, $x^2-x-1 = 0$, cannot be solved by factoring and instead requires the quadratic formula. Likewise, integral problems share a similar appearance, but students must learn which integration technique is appropriate for each of the subtly different kinds. Such superficial similarity is ubiquitous in mathematics, and this is why students need discrimination training.

The link between superficial similarity and the importance of this discrimination learning has been demonstrated by VanderStoep and Seifert (1993). In their first experiment, for instance, students learned to solve two kinds of mathematics problems that were either similar or different in appearance. Some students saw a tutorial emphasizing *how* to solve each kind of problem, and others saw a tutorial emphasizing *which* of two procedures was appropriate for each kind of problem. The *learning-which* tutorial proved more effective than the *learning-how* tutorial when the two kinds of problems were similar, but the tutorials were equally effective when the kinds of problems did not resemble each other. Thus, discrimination training proved useful when problems were similar in appearance.

In summary, while the importance of discrimination training provides one reason to suspect that the mixture or interleaving of problem types will produce better subsequent test performance, it appears that no prior experiments have directly compared mixed and blocked practice. This was the aim of Experiment 2. If mixed practice is, in fact, superior to blocked practice for mathematics learning, it would suggest that the widespread reliance on blocked practice needs reevaluation.

Experiment 1

The first experiment assessed the effects of temporal distribution (spaced vs. massed practice) and overlearning (massed practice vs. light massed practice) of mathematics practice. College students were taught how to calculate the number of permutations of a

letter sequence with at least one repeated letter (e.g., *aabccc*), and they then practiced this procedure according to one of three schedules. Spacers worked two practice problems in each of two sessions separated by 1 week; Massers worked the same four practice problems in a single session; and Light Massers worked just two practice problems in one session. All students were tested 1 week after their final practice problem. The procedure is summarized in Fig. 2a.

Two critical comparisons are made. First, we assessed the effect of spacing practice by comparing the test performance of Spacers and Massers. Second, we assessed the effect of overlearning by comparing the test performance of Massers and Light Massers. As detailed in the Introduction, the standard format relies predominantly on practice sets that are massed, and the sheer number of problems within these practice sets ensures overlearning. By contrast, the shuffled format incorporates spaced practice.

Method

Participants

All three sessions were completed by 66 undergraduates (51 women) at the University of South Florida. An additional 14 students completed the first session but failed to attend either the second or third session.

Task

Students calculated the number of unique orderings (i.e., permutations) of a letter sequence with at least one repeated letter. For example, the sequence *abbccc* has 60 permutations, including *abccbc, accbcb, bbaccc*, and so forth. Every letter sequence was four to eight letters in length, and the number of *unique* letters in each sequence equaled two (*a* and *b*) or three (*a*, *b*, and *c*). No sequence had more than 90 permutations. The number of permutations for any sequence is given by a formula that is illustrated in the Appendix, but students were not shown this formula because we believed that it would prove too complex for some of our students Instead, we taught students with examples that were presented exactly as shown in Fig. 1.

Base rate survey

Although we were confident that this particular kind of permutation problem was unknown to our participant pool, we verified this by testing a sample of 50 students (with 43 women) from the same participant pool, none of whom participated in either Experiments 1 or 2. Each student was given 3 min to find the number of permutations for three of the practice problems used in Experiment 1.

None of the surveyed students correctly answered any of the problems, and none of their written solutions exhibited any evidence of the appropriate procedure. Some attempted to simply list every permutation, but none succeeded, probably because of the time constraint. Hence, this survey showed that this task is virtually, if not entirely, unknown to our participant pool. Furthermore, to the extent that any relevant pre-experimental knowledge did exist, it would not confound the experiment because of random assignment and the law of large numbers.

Problem
In how many ways can the letters *abbccc* be arranged?

Solution

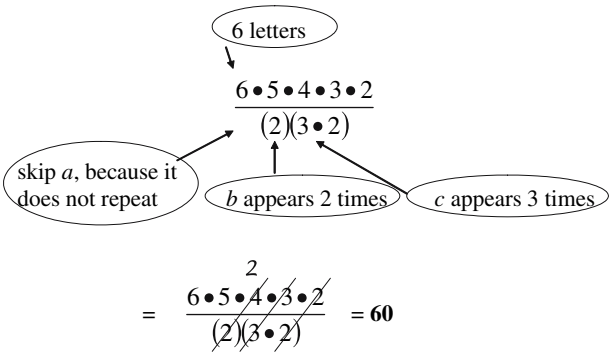


Fig. 1 Permutation task. This example illustrates the format of the solutions presented to students during the tutorial and the feedback after each example

Procedure

Each student attended three sessions spaced 1 week apart. At the beginning of the first session, each student was assigned to the group of Spacers, Massers, or Light Massers. At no point were students told what to expect in subsequent sessions.

All students simultaneously observed a 3-min tutorial at the beginning of the first session. The tutorial included a single projected visual slide with some explanatory information and a sample problem, accompanied by oral explanation. The slide also included the solution to the sample problem, which was presented exactly as shown in Fig. 1. Immediately after the tutorial, every student completed the first practice set. The Light Massers worked only the first practice set. The Massers worked both practice sets in session one. The Spacers worked the first practice set in session one and the second practice set in session two.

Each practice set included two examples and two practice problems, all of which were presented in a test booklet. Students were given 45 s to solve each example, and each example was followed immediately by a 15-s visual projection of its solution (which, like the tutorial sample problem, was presented as shown in Fig. 1). The two practice problems were also allotted 45 s each but were not followed by feedback. The selection and order of the example and practice problems did not vary across students.

The test was given to the Massers and Light Massers in session two (1 week after their final practice problem), and the Spacers were tested in session three (1 week after their final practice problem), as illustrated in Fig. 2a. The test consisted of a single piece of paper with five novel problems, and all students saw the same five problems in the same order. Students were asked to solve all five problems in 225 s (which averages to 45 s per problem). Students were required to sit for the entire time period, and feedback was not provided.

Critically, although the Massers and Light Massers were tested in the second session, they were required to attend the third session. If they had been allowed to skip the third

A Practice Procedure

	week 1	week 2	week 3
Spacers	2 problems	2 problems	test
Massers	4 problems	test	filler task
Light Massers	2 problems	test	filler task

B Test Performance

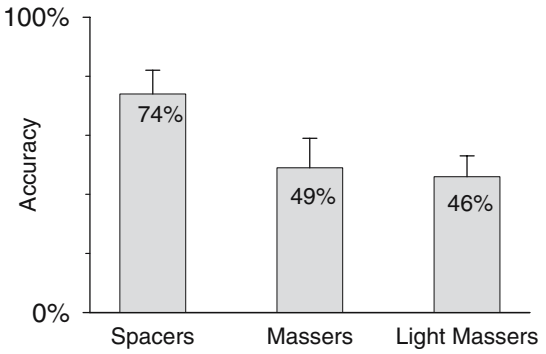


Fig. 2 Experiment 1. **a** Practice procedure. Each pair of practice problems was preceded by two examples. Students saw a single tutorial immediately before the first example. Practice session performance did not differ reliably between groups. **b** Test performance. *Error bars* reflect ±1 SE

session, the test scores of the Massers and Light Massers would have included subjects who might have not attended the third session *if* it had been required. This would have confounded the experiment because subjects who fail to show for a follow-up session perform worse, on average, than those who show. Thus, allowing Massers and Light Massers to skip the third session would have confounded the experiment in favor of the Spacers. Indeed, the present experiment included three Massers who failed to attend the third session, and their average test score was, in fact, lower than the average score of the Massers who attended all three sessions. Thus, by requiring every student to attend the third session, the observed spacing effect was not exaggerated.

Resultsand discussion

Inclusion criterion

Because one aim of this study was to assess the benefits of overlearning by comparing the Massers and Light Massers, it was important that Light Massers provide at least one correct response during practice. This is because overlearning requires that students continue practice beyond criterion, and, consequently, the benefits of overlearning cannot be assessed unless the control group reaches criterion. Therefore, we restricted our analyses to those students who correctly answered at least one of the first two practice problems (which were the only two practice problems attempted by all students). This eliminated six of the 66 students. The exclusion of these six students slightly increased the mean test scores of each group, but it had no effect on the findings.

Practice performance

Mean accuracy for the first two problems equaled 95% (SE = 2%). Naturally, there were no reliable differences between the three groups on these first two problems ($p > 0.05$) because these two practice problems were completed *before* the procedures for the three groups diverged.

For the second set of two practice problems, the timing *was* manipulated, as it was begun immediately after the first practice set (Massers) or 1 week later (Spacers). Yet despite the delay imposed upon the Spacers, their second practice set mean accuracy of 83% (SE = 6%) was about equal to the Massers' average of 82% (SE = 7%), $t < 1$. Thus, a 1-week delay did not impair performance on the second practice set, and this was probably due to the fact that each practice set began with two solved examples. Notably, though, this was not a confound because both Massers and Spacers saw the same two examples just before the second practice set. In summary, practice strategy did not significantly affect practice performance.

Test performance

Practice strategy affected test performance. As shown in Figure 2a, the Spacers' mean test accuracy of 74% (SE = 8%) exceeded both the Massers' average of 49% (SE = 10%) and the Light Massers' average of 46% (SE = 7%). An analysis of variance revealed a reliable difference between the groups, $F\ (2, 57) = 3.59$, $p < 0.05$, $\eta_p^2 = 0.11$. Subsequent Holm–Sidak comparisons revealed that the Spacers outscored both the Massers ($p < 0.05$) and the Light Massers ($p < 0.05$), but the Massers and the Light Massers did not differ reliably ($p = 0.8$).

Summary

Two key findings were observed. First, despite a twofold different in the amount of massed practice assigned to Massers and Light Massers, there was not detectable difference in their test scores. Thus, because the Light Massers correctly answered at least one practice problem (as all analyses excluded subjects who did not correctly answer any practice problems), this finding constitutes a null effect of overlearning (i.e., immediate post-criterion study). Admittedly, overlearning might have significantly boosted test scores if the number of massed practice problems had varied by a factor of, say, 10 and not just two. However, any such effect would need to be extremely large before it would justify the tenfold increase in study time. This is because learners have a finite amount of study time, and they should invest this time in strategies that provide a good return on their investment. Thus, while an extremely large amount of overlearning might boost test scores, it would probably not be efficient. Finally, and as noted in the Introduction, a null effect of mathematics overlearning was observed previously (Rohrer & Taylor, 2006). However, the present finding is the first in which the null effect cannot be attributed to an artificial constraint on test performance. That is, the inability of the Massers to outscore the Light Massers cannot be attributed to an inherent ceiling effect because the Massers were vastly outscored by the Spacers. This superiority of Spacers over Massers–a spacing effect—is the second key finding of this study. Both findings—the null effect of overlearning and the superiority of spacing over massing—favor the shuffled format, which uses spaced practice, over the more commonly used standard format, which induces massing and overlearning.

Experiment 2

In the second experiment, students worked a set of practice problems that were either blocked by problem type or mixed together. College students were taught how to find the volume of the four obscure geometric solids shown in Fig. 3a and then completed one of two randomly assigned practice schedules. Each group worked the same practice problems, but the practice problems were either blocked (e.g., four problems for one solid, then four problems for another solid) or systematically mixed. Both the Mixers and the Blockers completed two practice sessions, separated by 1 week, and were tested 1 week after their second practice session, as shown in Fig. 4a. As detailed in the Introduction, mixed practice requires that students learn to pair a type of problem with its appropriate procedure, and, for that reason, we suspected that the Mixers would outscore Blockers at test.

Method

Participants

Three sessions were completed by 18 undergraduates (13 women) at the University of South Florida. An additional 15 students completed the first session but failed to attend either the second or third session. None participated in Experiment 1. Although the sample size was small, statistical power was not a concern because of effect sizes were large.

Task

The students learned to calculate the volume of four geometric solids. Formal definitions of the four solids are given in the Appendix, but students instead saw the illustrations and descriptions shown in Fig. 3a. The volume of each solid depends solely on its radius (r) and height (h). In every problem presented during practice or test, the radius and height equaled a positive integer of seven or less. Problems and solutions were presented in the format shown in Fig. 3b. Of note, students were asked to write the appropriate formula in a preprinted box and write the volume in a preprinted oval.

Base rate survey

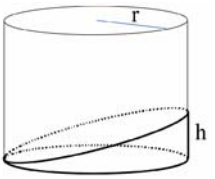
To verify that the volume formulas were virtually unknown to the participant pool used in Experiment 2, we tested a sample of 25 students (14 women) from the same pool, none of whom participated in either experiment. Each student was given 8 min to solve the eight test problems given in Experiment 2, and these included two problems for each of the four solids. None of the students correctly answered any of the problems. As in Experiment 1, concerns about pre-experimental knowledge are further tempered by random assignment and the law of large numbers.

Procedure

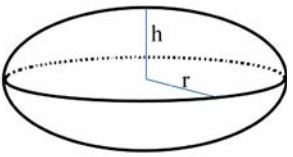
The students attended three sessions spaced 1 week apart. At the beginning of the first session, each student was randomly assigned to the group of Mixers or Blockers. For both groups, the first and second sessions were practice sessions, and the third session included the test.

A

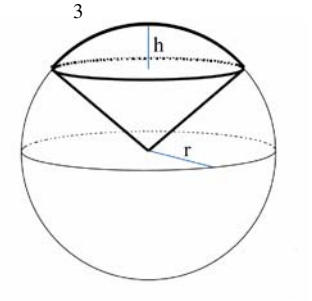
A **wedge** is the boldfaced portion of the tube.
Its bottom is a circle, and its top is a slanted oval.
Its volume equals $\frac{r^2 h \pi}{2}$



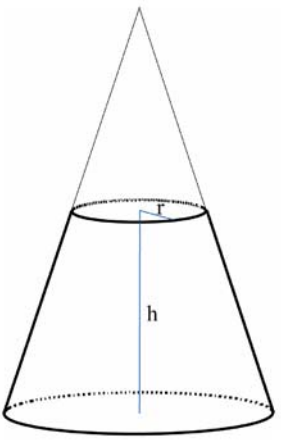
A **spheroid** is similar to a sphere.
But its height has been squeezed or stretched.
Its volume equals $\frac{4 r^2 h \pi}{3}$



A **spherical cone** is the boldfaced part of the sphere.
Its bottom is at the center of the sphere.
The rim of the cone is on the surface of the sphere.
Its volume equals $\frac{2 r^2 h \pi}{3}$



A **half cone** is the bottom half of a cone.
Both its top and bottom are circles.
Its volume equals $\frac{7 r^2 h \pi}{3}$



B

Problem
Find the volume of a wedge with $r = 2$ and $h = 3$.
Write the formula in the box; write the answer in the oval.

Solution

$$\boxed{\frac{r^2 h \pi}{2}}$$

$$= \frac{2 \cdot 2 \cdot 3\pi}{2}$$

$$= \boxed{6\pi}$$

Fig. 3 Volume task. **a** The illustrations and descriptions are identical to those shown to the students. Formal definitions of each shape are given in the Appendix. **b** A sample problem. This example illustrates the format of the solutions presented during the tutorial and the feedback after each practice problem

Each of the two practice sessions included four tutorials and 16 practice problems. The Mixers read all four tutorials before beginning the practice problems, and the 16 practice problems were randomly ordered with the constraint that each set of four practice problems (e.g., 1-4, 5–8, etc.) included one problem for each of the four solids. For the Blockers, each tutorial on a given solid was followed immediately by the four problems relating to that solid (e.g., the wedge tutorial was followed by four wedge problems, the spherical

cone tutorial was followed by four spherical cone problems, and so forth). Within each condition, the order of the problems did not vary across students, and no problem appeared in both practice sessions. Most importantly, both groups saw the same tutorials and the same practice problems in each session.

Students were given 45 s to read each tutorial, which consisted of the illustration and written description in Fig. 3a and one solved example like that shown in Fig. 3b. Students were allotted 40 s for each practice problem, and each practice problem was followed immediately by a 10-s visual presentation of the solution. Each practice problem and its subsequent solution were presented in the format shown in Fig. 3b.

One week after the second session (and their last practice problem), students were tested. Eight novel problems, with two problems for each solid, were presented simultaneously in a random order. All students saw the same problems in the same order. Students were allotted 8 min and were required to sit for the entire time period. Feedback was not provided.

Results and discussion

Inclusion criterion

Every student correctly answered at least one practice problem in *each* practice session. Consequently, every student was included in all further analyses.

Practice performance

Practice session performance was impeded by mixture (Fig. 4b), as the Blockers' average of 89% (SE = 4%) statistically exceeded the Mixers' average of 60% (SE = 7%), $t(16) = 3.14$, $p < 0.01$, $d = 1.06$. This superiority of Blockers was due primarily to the difference in their scores during the first session (87 vs. 43%), $t(16) = 3.88$, $p < 0.01$, $d = 0.53$. In the second practice session, the Blockers' superiority was more moderate and not statistically significant (91 vs. 78%), $t(16) = 1.58$, $p > 0.05$.

Test performance

By contrast, the mean test performance of Mixers (63%, SE = 12%) was far greater than that of the Blockers (20%, SE = 9%), $t(14) = 2.64$, $p < 0.05$, $d = 1.34$, as shown in Fig. 4c. Thus, mixed practice produced superior test performance and inferior practice performance (compared to blocked practice), as evidenced by a statistically significant interaction between practice strategy (mixed vs. blocked) and experiment phase (practice vs. test), $F(1, 16) = 35.08$, $p < 0.001$.

In a secondary analysis of test performance, we tabulated the number of test problems for which students provided the correct formula but not the correct answer. Across all students and all test problems, this happened only twice: once for a Mixer and once for a Blocker. Thus, if the correct formula was recalled, the correct answer was almost always found. This means that Blockers (and Mixers) knew *how* to solve each kind of problem at the time of test, and, consequently, their poor performance was due to their inability to recall the correct formula for each problem. Thus, as fully detailed in the Introduction, it appears that students received the necessary discrimination training only when practice problems were mixed by type.

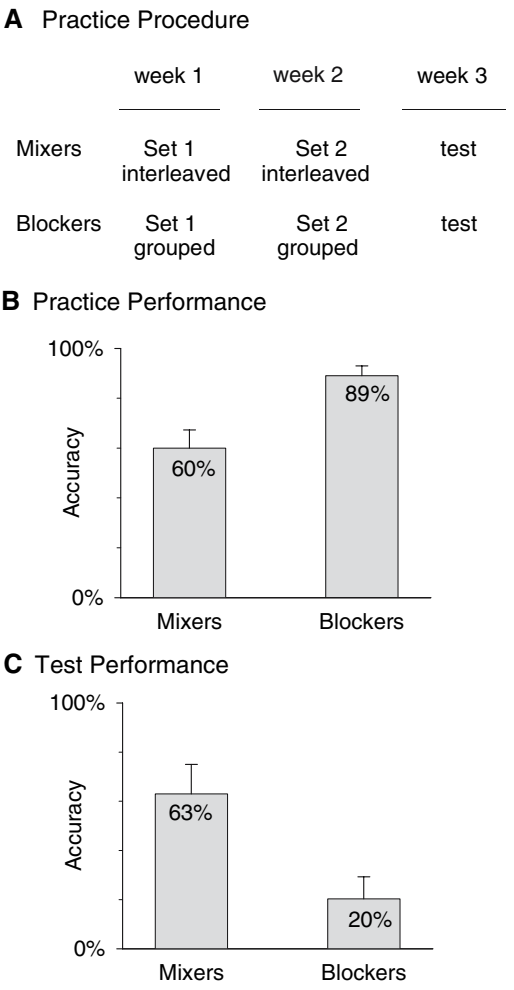


Fig. 4 Experiment 2 **a** Practice procedure. **b** Practice session performance. *Error bars* reflect ±1 SE. Data are averaged across the two practice sessions. See text for details about performance on each specific practice session. **c** Test performance. *Error bars* reflect ±1 SE

Finally, although it might seem that the superior test scores of Mixers could be attributed to the fact that the test problems were mixed rather than blocked, we believe this is unlikely for two reasons. First, if it is assumed that the Blockers' poor test performance stemmed from their inability to pair each kind of problem with the appropriate formula, as suggested by the analysis in the paragraph immediately above, the order of the test problems is logically inconsequential. Second, because the test included only two problems of each type, the difference between a blocked and mixed format would have been slight.

Summary

While blocked practice proved superior to mixed practice during the *practice* session, subsequent test scores were much greater when practice was mixed rather than blocked. The superior test performance after mixed practice is, in our view, attributed to the fact that students in this condition were required to know not only *how* to solve each kind of problem but also *which* procedure (i.e., formula) was appropriate for each kind of problem

(i.e., solid). This possibility is also consistent with the finding that virtually every test error was due to the selection of the wrong formula.

General discussion

Test performance in both experiments benefited from altering either the timing or the serial order of practice problems. In Experiment 1, test performance increased sharply if a given set of practice problems was spaced across two sessions separated by 1 week, as compared to the massing of these problems within a single session. In addition, there was no decrement in test performance when the number of massed practice problems was reduced by half, which is to say that there was a null effect of the strategy known as overlearning. In Experiment 2, test performance improved 250% when practice problems of different types were mixed together and not blocked by type. In brief, while an increase in the number of massed practice problems did not reliably affect test scores (Experiment 1), large gains in test performance were achieved by the use of spacing or mixing, even though neither of these strategies required additional practice problems.

The two experiments also demonstrated that a learning strategy which provides superior *test* performance is not necessarily the one that optimizes *practice* performance. In Experiment 1, the spacing of practice, which boosted test performance, had no effect on practice performance. In Experiment 2, the mixture of problem types, which boosted test performance, actually impeded practice performance. Bjork and his colleagues have observed similar dissociations between practice and test performance, leading them to describe these initially costly but ultimately beneficial strategies as *desirable difficulties*(e.g., Bjork, 1994; Christina & Bjork, 1991; Schmidt & Bjork, 1992).

Caveats

Several limitations apply to the generality of these findings. First, our subjects were college students, and it is possible that the effects observed here might be muted or even absent with much younger students. Second, the experiments reported here relied on a test that required students to solve problems exactly like those shown in practice, and it is not known whether our findings would obtain with measures requiring transfer. Third, our experiments were laboratory based, and future research will be needed to determine if the findings will replicate in a classroom setting. Fourth, the tasks used in our experiments are procedural rather than conceptual (e.g., Rittle-Johnson & Alibali, 1999; Rittle-Johnson, Siegler, & Alibali, 2001), and it remains unknown whether the benefits of spaced and mixed practice would hold for more abstract, conceptual tasks. In brief, our results leave open the possibility that our findings may not generalize to different subjects, tasks, and settings, yet, at the same time, we know of no reason why they would not.

Practical implications

The present results cast doubt on the utility of the standard practice format used in most mathematics textbooks because this format is characterized by massed practice and blocked practice—the very two strategies that proved here to be deficient long-term learning strategies. Likewise, the present findings suggest that the shuffled format, with its

reliance on spaced and mixed practice, deserves further consideration by researchers, teachers, educators, and authors.

We should emphasize that the shuffled format can be adopted without any change in the nature or the order of the *lessons*. It does mean, however, that, if a lesson is omitted, one must be careful to also omit corresponding problems throughout the remainder of the textbook. Fortunately, this task is made easy if the textbook includes an index listing every practice problem and its corresponding lesson, allowing the instructor to easily avoid assigning problems relating to omitted topics. Such an index also means that a student can find the lesson corresponding to a problem that he or she cannot solve. In fact, the lesson number for each problem could be provided immediately adjacent to each practice problem.

Perhaps the most well known example of the shuffled format is the Saxon line of mathematics textbooks (e.g., Saxon, 1997). In these textbooks, no more than two or three problems within each practice set are drawn from the immediately preceding lesson, and the remaining one or two dozen problems are drawn from many different lessons. We are not aware of any published, controlled experiments comparing a Saxon and non-Saxon textbook, but such an experiment may not be particularly informative because it would be confounded by the numerous differences between any two such texts. That is, regardless of the outcome of an experimental comparison of a shuffled textbook and a standard textbook, any observed differences in, say, final exam performance might reflect differences in the lessons rather than practice format.

Such confounds would be avoided, however, if two groups of students were presented with the same lessons and different practice sets. For example, each group of students could receive a packet that includes the lessons from a traditional textbook, and these lessons would appear in the same order for both groups. Both groups would also see the same practice problems, but the problems would be arranged in either a standard format or shuffled format. By way of disclosure, neither author has an affiliation with a publishing company or mathematics textbook, although the first author is a former mathematics teacher who has taught with textbooks from many different publishers, including Saxon.

Additional advantages of a shuffled format

There may be additional benefits of a shuffled format not addressed by Experiments 1 and 2. For example, when practice problems relating to a given topic are spaced across multiple practice sets, a student who fails to understand a lesson (or fails to attend a lesson) will still be able to solve most of the problems within the following practice set, whereas a massed practice set ensures that this student will have little or no success. Likewise, if that student achieves better understanding of the topic in a subsequent class meeting (perhaps by observing other students solve the previously assigned practice problems in class), a shuffled format provides opportunities to practice these new skills in the future.

Finally, the logistical demands and the financial costs of adopting a shuffled practice format are relatively small. Instructors can incorporate a shuffled format regardless of their adopted textbook by merely shuffling practice problems from multiple practice sets. Ideally, though, the shuffled format would be incorporated by textbooks and instructional software packages. Notably, the adoption of this new format could be accomplished with little trouble or expense, as authors and publishers could merely rearrange the practice problems in the next edition.

Acknowledgments This research was supported by a grant from the Institute of Education Sciences, US Department of Education. We thank Kristina Martinez and Erica Porch for their assistance with data collection.

Appendix

Permutations

If a sequence of items includes n items and k unique items, the number of permutations of the sequence equals $n!/(n_1!\ n_2!\ ...\ n_k!)$, where n_i equals the number of occurrences of item i. Thus, for the sequence *abbccc*, the number of permutations equals 6!/(1! 2! 3!), or 60.

Wedge

A wedge is obtained by the truncation of a cylinder by two planes if exactly one of the planes is perpendicular to the cylinder and if the linear intersection of the two planes includes exactly one point on the cylindrical surface. If the latter constraint is relaxed so that the linear intersection may intersect the cylindrical surface *at either one or two points*, the solid is a *cylindrical wedge*. This is the shape shown in Fig. 3a. We chose the term wedge for this specific case because we do not know of an accepted term. Its volume equals $r^2h\pi/2$, where r equals the radius of its circular base and h equals its maximum height

Spherical cone

A spherical cone is obtained by removing a conical section of a sphere provided that the vertex of the cone is at the sphere's center and the base of the cone is on the sphere's surface, as shown in Fig. 3a. Its volume is given by $2r^2h\pi/3$, where r equals the radius of the sphere and h equals the difference of the sphere's radius and the cone's height

Spheroid

A spheroid is obtained by the rotation of an ellipse about one of its axes. The spheroid in Fig. 3a, for example, is rotated about its vertical axis. Its volume equals $4r^2h\pi/3$, where r equals the ''equatorial radius'' and h equals the ''polar radius.'' The values of r and h also equal one-half of the major and minor lengths of the rotated ellipse.

Half cone

A half cone is a cone truncated by a plane parallel to its base so that the truncation reduces the cone's height by half. Its volume equals $7r^2h\pi/3$, where r equals the radius of the upper base and h equals the height of the truncated cone, as illustrated in Fig. 3a. The half cone is a specific instance of a *conical frustum*, which has a height equal to any proportion of the cone's height. We chose the term ''half cone'' to describe a conical frustrum with height equal to exactly half of the cone's height.

References

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign-language vocabulary and the spacing effect. *Psychological Science, 4*, 316–321.

Bjork, R. A. (1979). Information-processing analysis of college teaching. *Educational Psychologist, 14*, 15–23.

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 II* (pp. 391–401). London: Wiley.

Bjork, R. A. (1994). Memory and meta-memory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge: MIT.

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research, 74*, 245–248.

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

Carson, L. M., & Wiegand, R. L. (1979). Motor schema formation and retention in young children: A test of Schmidt's schema theory. *Journal of Motor Behavior, 11*, 247–251.

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.

Christina, R. W., Bjork, R. A. (1991). Optimizing long-term retention and transfer. In D. Druckman & R. A. Bjork (Eds.), *In the mind's eye: Enhancing human performance* (pp. 23–56). Washington DC: National Academy Press.

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review, 1*, 309–330.

Driskell, J. E., Willis, R. P., & Copper, C. (1992). Effect of overlearning on retention. *Journal of Applied Psychology, 77*, 615–622.

Fitts, P. M. (1965). Factors in complex skill training. In R. Glaser (Ed.), *Training research and education* (pp. 177–197). New York: Wiley.

Foriska, T. J. (1993). What every educator should know about learning. *Schools in the Middle, 3*, 39–44.

Gilbert, T. F. (1957). Overlearning and the retention of meaningful prose. *Journal of General Psychology, 56*, 281–289.

Glencoe (2001) *Mathematics: Applications and Connections—Course 1*. New York: Glencoe-McGraw Hill.

Grote, M. G. (1995). Distributed versus massed practice in high school physics. *School Science and Mathematics, 95*, 97–101.

Hall, J. F. (1989). *Learning and memory*, 2nd Ed. Boston: Allyn & Bacon.

Jahnke, J.C., & Nowaczyk, R. H. (1998). *Cognition*. Upper Saddle River: Prentice Hall.

Kester, L., Kirschner, P. A., & Van Merriënboer, J. J. G. (2004). Timing of information presentation in learning statistics. *Instructional Science, 32*, 233–252.

Krueger, W. C. F. (1929). The effect of overlearning on retention. *Journal of Experimental Psychology, 12*, 71–78.

Mayfield, K. H., & Chase, P. N. (2002). The effects of cumulative practice on mathematics problem solving. *Journal of Applied Behavior Analysis, 35*, 105–123.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review* (in press).

Postman, L. (1962). Retention as a function of degree of overlearning. *Science, 135*, 666–667.

Radvasky, G. (2006). *Human memory*. Boston: Pearson Education Group.

Rea, C. P., & Modigliani, V. (1985). The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. *Human Learning: Journal of Practical Research & Applications, 4*, 11–18.

Reynolds, J. H., & Glaser, R. (1964). Effects of repetition and spaced review upon retention of a complex learning task. *Journal of Educational Psychology, 55*, 297–308.

Rittle-Johnson, B. & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology, 91*, 175–189.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*, 346–362.

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology, 20*, 1209–1224.

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology, 20*, 1209–1224.

Saxon, J. (1997). *Algebra I* (3rd Ed.). Norman: Saxon Publishers.
Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*, 207–217.
Smith, S. M., & Rothkopf, E. Z. (1984). Contextual enrichment and distribution of practice in the classroom. *Cognition and Instruction, 1*, 341–358.
VanderStoep, S. W., & Seifert, C. M. (1993). Learning 'how' versus learning 'when': Improving transfer of problem-solving principles. *Journal of the Learning Sciences. 3*, 93–111.