

Blocking vs. Interleaving: Examining Single-Session Effects Within Middle School Math Homework

Korinn Ostrow¹(✉), Neil Heffernan¹, Cristina Heffernan¹, and Zoe Peterson²

¹ Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, USA
{ksostrow,nth,ch}@wpi.edu

² Carleton College, 1 North College Street, Northfield, MN 55057, USA
petersonz@carleton.edu

Abstract. The benefit of interleaving cognitive content has gained attention in recent years, specifically in mathematics education. The present study serves as a conceptual replication of previous work, documenting the interleaving effect within a middle school sample through brief homework assignments completed within ASSISTments, an adaptive tutoring platform. The results of a randomized controlled trial are presented, examining a practice session featuring interleaved or blocked content spanning three skills: Complementary and Supplementary Angles, Surface Area of a Pyramid, and Compound Probability without Replacement. A second homework session served as a delayed posttest. Tutor log files are analyzed to track student performance and to establish a metric of global mathematics skill for each student. Findings suggest that interleaving is beneficial in the context of adaptive tutoring systems when considering learning gains and average hint usage at posttest. These observations were especially relevant for low skill students.

Keywords: Interleaving · Blocking · Adaptive tutoring system · Mathematics education · Randomized controlled trial

1 Introduction

The benefit of interleaving cognitive content has gained attention in recent years. A simple intervention rooted in kinesthetic research pertaining to the acquisition of motor skills [18], interleaving has since evolved into a powerful tool for the modern classroom. Specifically, significant effects have been verified in the realm of mathematics education in classroom trials and through simulated studies [9, 17, 6, 19, 7]. Research within this realm has examined the interleaving effect by mixing or alternating the delivery of skill or problem content, such that similar problems are no longer ‘blocked’ or presented in uniform segments. The benefits observed when interleaving mathematics content are often credited to the discriminative-contrast hypothesis [1], which purports that the effect is rooted in a student’s enhanced ability to pinpoint differences in problem content. As such, interleaving provides an obvious tool within a domain that relies largely on problem type identification and solution strategy choice [15].

Despite this clarity, the details of interleaving remain somewhat obscure. It is heavily documented that interleaving is confounded by an inherent spacing effect [15], yet few researchers effectively isolate interleaving by examining a single session or controlling for the spacing of content [19]. Researchers have also added complexity to the issue, questioning which dimension of cognitive content (i.e., the skill, the task type, the representation, etc.) to interleave for optimal results [12, 13]. Further, despite continued reports of significant learning gains observed at posttest after interleaved practice, policymakers and educational designers fail to interleave mass-produced content, claiming that it is detrimental to the student's learning *experience* [16, 19, 5]. Essentially, the practice has earned a bad reputation for making the learning process more complex, or for adding what Bjork terms 'desirable difficulty' [2].

The present study serves as a conceptual replication of Rohrer & Taylor's work on shuffling mathematics practice problems [17]. While replications are rare in general [14], a recent analysis of leading education journals found that less than 0.13% of publications were replications [8]. However, repeated observations of significant educational findings, especially within different contexts, have the power to produce systemic change. While not a direct replication, we similarly aim to assess the interleaving effect within mathematics skills amidst a single practice session, considering delayed posttest measures as dependent metrics. More uniquely, we seek to document the effect using a brief homework assignment completed within ASSISTments, an online adaptive tutoring system. We also consider a global metric of mathematics skill for each student, in an attempt to gauge how the effect differs across skill level.

ASSISTments is a fast growing platform offered as a free service of Worcester Polytechnic Institute and used for homework and classwork by over 50,000 students around the world [4]. The system offers teachers a library of prebuilt content, primarily with a focus on mathematics skills aligned to the Common Core State Standards, as well as the ability to build content to match their curriculum or course goals. Simultaneously, students benefit from correctness feedback and tutoring strategies within an adaptive environment that advances skill practice beyond that achieved through traditional classroom practices. ASSISTments also serves as a shared scientific tool for education research [4]. Adaptive tutoring systems provide a natural learning environment from which to assess best practices, and yet, to our knowledge, little work has been done to examine interleaving within these settings. Thus, a randomized controlled trial was designed within ASSISTments to examine the subtleties of interleaving, as guided by the following research questions:

1. When controlling for student skill level, do learning gains (as measured by average posttest score) differ when practice session content is interleaved?
2. When controlling for student skill level, does interleaving practice session content lead students to interact differently with the system at posttest (as measured by average hint usage and average attempt count)?

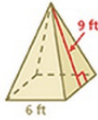
It was hypothesized that interleaving skill content in the practice session would have a beneficial effect on student performance as measured at posttest, leading to increases in posttest score and reductions in the average number of hints and attempts used during posttest problems.

2 Methods

This study was conducted with five classes spanning three teachers at a suburban middle school in Massachusetts. All teachers and students within the sample population were familiar with ASSISTments, having used the system for classwork and homework throughout the school year. Researchers worked with a participating teacher to design problem content for two homework assignments (i.e., a practice session and a delayed posttest). In April 2014, the teacher isolated three mathematics skills that her students had learned earlier in the year to serve as review while allowing for the observation of relearning via hint usage. The skills covered were Complementary/Supplementary Angles (Skill A; originally covered in February/March 2014), Surface Area of a Pyramid (Skill B; originally covered in November/December 2013), and Probability of Compound Events without Replacement (Skill C; originally covered in January 2014). A problem exemplifying Skill B with all available hint feedback is provided in Figure 1; problems exemplifying Skills A and C can be accessed at Ostrow [11] for further reference.

Problem ID: PRAX8WT [Comment on this problem](#)

Find the surface area of the regular pyramid.



The pyramid has one base and 4 equal faces. The base is a square and the other faces are made of triangles each with a base (b) and a height (h).

Find Area of Base = s^2

Find 4 * Area of side face = $4(b \cdot h / 2)$

[Comment on this hint](#)

Area of Base = $6^2 = 36$

Area of 4 faces = $4(6 \cdot 9 / 2) = 4 \cdot 27 = 108$

Total surface area = $36 + 108$

[Comment on this hint](#)

The answer is 144.

[Comment on this hint](#)

Type your answer below (mathematical expression):

Fig. 1. Example of Skill B, Surface Area of a Pyramid

For the practice session, four problems were created for each skill, resulting in a single assignment with twelve problems. These problems were isomorphic in structure, but designed such that problem difficulty would increase with each practice opportunity. Hence, a student's first experience with Skill A was relatively easy, while her fourth experience with the skill was more challenging. One additional problem

was created for each skill, matching the highest difficulty level presented during practice, to establish a separate, three-problem assignment that would serve as a delayed posttest. Practice and posttest sessions were both assigned as homework, establishing an authentic learning experience and reducing the potential for immediate assistance from an adult. Settings for homework completion were ultimately unknown and were likely differential across students.

Further, although straying from the conventions of a ‘formal’ posttest, permitting the use of hints and multiple attempts during the posttest assignment allowed researchers to investigate variables of student performance extending beyond average posttest score (i.e., an average of the student’s accuracy on their first attempt at solving each problem).

3 Procedure

After the creation and release of content, five teachers assigned this study to an initial sample of 226 7th grade students. Students were randomly assigned to either the experimental condition, in which skill problems within the practice assignment were presented in an interleaved or mixed pattern, or to the control condition, in which skill problems within the practice assignment were presented using a blocked approach. Random assignment was accomplished using a pseudo-random number generator within the ASSISTments tutor, and occurred at the student level rather than the class level to control for potential teacher and class effects.

Regardless of condition, students received the same twelve problems during the practice session, with the only difference being presentation order. Problem delivery patterns for each group are depicted in Figure 2. Using this design, the effects of interleaving were not specifically isolated from the effects of spacing. For instance, students in the interleaved condition experienced problem A₄ at a later point in time than students in the blocked condition. However, the practice session was delivered as a single assignment in an attempt to minimize the effects of spacing [16, 3].

Regardless of condition, all students received a second homework assignment consisting of three problems in a static delivery pattern, serving as a delayed posttest. Participating teachers assigned this posttest anywhere from two to five days following the practice session. Details pertaining to the design of this study, including access to question content and the student experience can be found at Ostrow [11].

Blocked	A ₁ , A ₂ , A ₃ , A ₄ , B ₁ , B ₂ , B ₃ , B ₄ , C ₁ , C ₂ , C ₃ , C ₄
Interleaved	A ₁ , A ₂ , B ₁ , B ₂ , C ₁ , C ₂ , A ₃ , B ₃ , C ₃ , B ₄ , C ₄ , A ₄
Posttest	A ₅ , B ₅ , C ₅

Fig. 2. Experimental Design: Skill Problem Delivery Across Groups

Tutor log files were retrieved from the ASSISTments database and problem level data, including correctness, hint usage, and attempt count was isolated for each student. Using previously logged data, it was also possible to calculate a global metric of mathematics skill for each student based on the average accuracy of all problems he or she had ever completed within the system. This measure was then discretized using a median split to bin students as generally ‘high’ or ‘low’ skill.

Within the initial sample of 226 students assigned the practice session, one participating teacher failed to assign the posttest, resulting in the removal of 68 students from final analysis. Of the remaining 158 students, three students failed to complete enough of the practice session to verify their condition based on logged data, and were therefore excluded from analysis. Additionally, nine low-skill students failed to start the posttest assignment. Further assessment of these nine students revealed that six had experienced the blocked condition during the practice session, while three had experienced the interleaved condition. Only five of these students completed the practice session, with four students failing to complete the blocked session and one student failing to complete the interleaved session. A two-tailed independent t-test was performed to compare the number of practice session problems completed by these students across groups, revealing that condition was not a significant factor in disparate completion rate, $t = 0.048$, $p = .963$. These nine students were therefore excluded from posttest analysis without introducing an obvious bias.

A Chi-squared test of independence of the remaining 146 students did not indicate a significant relationship between condition and student skill level, $\chi^2(1, N = 146) = 0.195$, $p > .05$. However, the distribution across conditions was not equivalent (Blocked, $n = 60$; Interleaved, $n = 86$) due to the pseudo-random number generator that conducted student level randomization. Given the successful use of this assignment method in previous research, the authors had no reason to believe that a selection effect had occurred or that this process was in any way biased (i.e., affected by specific student characteristics). Thus, the skewed distribution observed here was not regarded as a threat to validity. The log files discussed herein have been stripped of identifiers and are available at Ostrow [11] for further reference.

4 Results

To examine our first research question, an ANCOVA was performed to analyze average posttest score across conditions when controlling for student skill level. Within 146 students, after controlling for the effect of student skill level, the effect of condition on posttest score trended toward significance, $F(1,143) = 2.69$, $p = 0.103$, $\eta^2 = 0.02$, Hedge's $g = 0.22$. As a covariate, student skill level was significantly related to posttest score, $F(1, 143) = 29.308$, $p < .001$, $\eta^2 = 0.17$. Levene's test was not significant, $p > .05$, and thus error variance was assumed to be equal across conditions. A summary of the effects of condition on average posttest score is depicted in Table 1. Analysis of means revealed that students in the interleaved condition ($M = 0.67$, $SD = 0.27$, $n = 86$) outperformed those in the blocked condition ($M = 0.61$, $SD = 0.27$, $n = 60$).

Split file ANOVAs were conducted to further examine the effect of condition across student skill level. For low skill students, condition had a significant effect on average posttest score, $F(1, 62) = 5.59$, $p < .05$, $\eta^2 = 0.08$, Hedge's $g = 0.60$. Levene's test was significant, $F(1, 62) = 5.16$, $p < .05$ suggesting the assumption of equivalent variance has been violated. Analysis of means revealed that students in the interleaved condition ($M = 0.58$, $SD = 0.29$, $n = 39$) significantly outperformed those in the blocked condition ($M = 0.42$, $SD = 0.23$, $n = 25$). Within high skill students, condition no longer had a significant effect on posttest score, $F(1, 80) = 0.01$, $p > .05$. Students in the interleaved condition ($M = 0.74$, $SD = 0.23$, $n = 47$) performed quite similarly

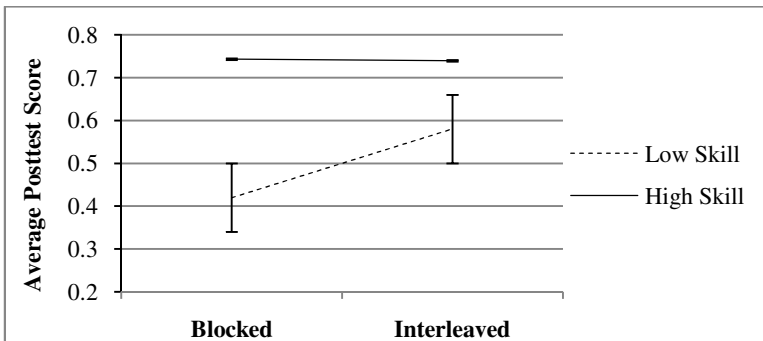
to those in the blocked condition ($M = 0.74$, $SD = 0.21$, $n = 35$). Summaries of the effects of condition on average posttest score for both skill levels are presented in Table 2. Figure 3 depicts the interaction of condition and student skill level observed in average posttest score.

Table 1. ANCOVA of the Effects of Condition on Average Posttest Score

Source	df	SS	MS	F	p	η^2
Skill Level	1	1.80	1.80	29.21	.000	0.17
Condition	1	0.17	0.17	2.69	.103	0.02
Error	143	8.80	0.06			
Total	146	71.25				

Table 2. ANOVA of the Effects of Condition on Average Posttest Score by Skill Level

Source	df	SS	MS	F	p	η^2
Low Skill						
Condition	1	0.41	0.41	5.59	0.021	0.08
Error	62	4.51	0.07			
Total	64	22.19				
High Skill						
Condition	1	0.00	0.00	0.01	0.945	0.00
Error	80	4.06	0.05			
Total	82	49.06				



Note. Standard Error for high skill students is not visible at this scale.

Fig. 3. Means for Average Posttest Score Across Conditions and Student Skill Levels

To examine our second research question, a MANCOVA was used to analyze the dependent measures of average posttest hint usage and average posttest attempt count as a function of condition after controlling for student skill level. Pillai's Trace is reported throughout, as the assumption of equality of covariance matrices was violated and this parameter offers a more robust understanding of variance. Within 146 students, after controlling for the effect of student skill level, there was a significant main effect of condition, Pillai's Trace = 0.06, $F(2, 142) = 4.81$, $p < 0.05$. At the multivariate level, student skill level was significant as a covariate, Pillai's Trace = 0.36, $F(2, 142) =$

39.25, $p < .001$, explaining approximately 36% of the total variance. Tests of between subjects effects revealed that condition had a significant effect on average posttest hint usage, $F(1, 143) = 6.24$, $p < .05$, $\eta^2 = 0.03$, Hedge's $g = -0.29$. Students in the interleaved condition used significantly less hints on average ($M = 0.33$, $SD = 0.57$, $n = 86$) than those in the blocked condition ($M = 0.50$, $SD = 0.64$, $n = 60$). However, condition did not significantly affect average posttest attempt count, $F(1, 143) = 0.10$, $p > .05$, with those in the interleaved condition ($M = 1.75$, $SD = 1.08$, $n = 86$) and those in the blocked condition ($M = 1.68$, $SD = 0.57$, $n = 60$) using a similar amount of attempts. A summary of univariate results is presented in Table 3.

Split file analyses revealed that the effects of interleaving were more impressive when low skill students were considered in isolation. Within 64 low skill students, condition had a significant multivariate effect, Pillai's Trace = 0.12, $F(2, 61) = 4.20$, $p < 0.05$. Univariate analyses revealed that condition had a significant effect on posttest hint usage, $F(1, 62) = 5.38$, $p < .05$, $\eta^2 = 0.08$, Hedge's $g = -0.59$, with students in the interleaved condition using less hints on average ($M = 0.64$, $SD = 0.70$, $n = 39$) than those in the blocked condition ($M = 1.04$, $SD = 0.62$, $n = 25$). Condition did not significantly affect posttest attempts, $F(1, 62) = 0.08$, $p > .05$, with those in the interleaved condition ($M = 2.12$, $SD = 1.42$, $n = 39$) and those in the blocked condition ($M = 2.04$, $SD = 0.58$, $n = 25$) using a similar amount of attempts.

Within high skill students, condition no longer had a significant multivariate effect, Pillai's Trace = 0.02, $F(2, 79) = 0.84$, $p > 0.05$. Summaries of the effects of condition on the dependent variables for both skill levels are presented in Table 4. Figure 4 depicts the interaction of condition and student skill level observed in average posttest hint usage.

Table 3. Univariate Summaries of the Effects of Condition on Dependent Variables

Source	df	Ave. Posttest Hints					Ave. Posttest Attempts				
		SS	MS	F	p	η^2	SS	MS	F	p	η^2
Skill Level	1	18.24	18.24	79.06	.000	0.35	15.13	15.13	21.07	.000	0.13
Condition	1	1.44	1.44	6.24	.014	0.03	0.07	0.07	0.10	.749	0.00
Error	143	32.98	0.23				102.67	0.72			
Total	146	75.75					552.06				

Table 4. ANOVA of the Effects of Condition on Dependent Variables by Skill Level

Source	df	Ave. Posttest Hints					Ave. Posttest Attempts				
		SS	MS	F	p	η^2	SS	MS	F	p	η^2
Low Skill											
Condition	1	2.43	2.43	5.38	.024	0.08	0.10	0.10	0.08	.785	0.00
Error	62	27.93	0.45				84.32	1.36			
Total	64	71.00					363.94				
High Skill											
Condition	1	0.06	0.06	1.09	.299	0.01	0.01	0.01	0.03	.865	0.00
Error	80	4.01	0.05				18.31	0.23			
Total	82	4.75					188.13				

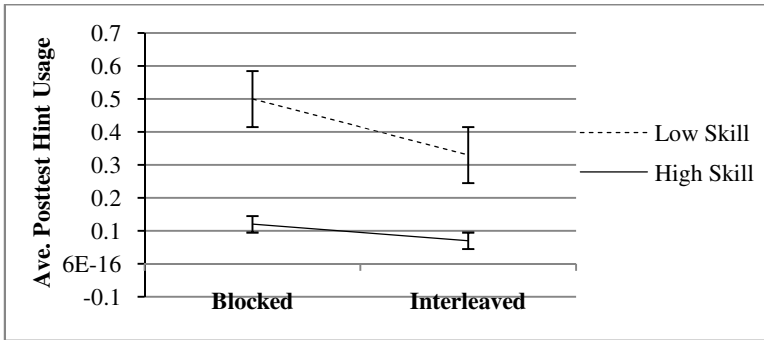


Fig. 4. Means for Average Posttest Hint Usage Across Conditions and Student Skill Levels

5 Discussion

The findings herein highlight the promising effects of interleaving skill content within brief mathematics homework assignments in the context of adaptive tutoring systems. Despite failing to achieve an effect size as large as that observed by Rohrer & Taylor [17] (Cohen's $d = 1.34$), we observed trends toward significance aligning with past work, serving as further evidence that interleaving skill content enhances learning gains as measured at a delayed posttest. This study also expanded upon interleaving literature to examine how these learning gains differ across student skill level. Further, the findings of the present study extended beyond binary measures of correctness to consider students' differential use of hints and attempts within an informal posttest setting. While this approach was somewhat novel, adaptive tutoring systems allow for the comparison of a variety of rich features within the learning experience that may provide deeper insight than accuracy alone. The observation of significantly different hint usage across conditions suggested that the consideration of feedback utilization, perhaps through a partial credit metric, may offer a more robust explanation for differential learning gains in future research.

The findings observed for low skill students were especially impressive and could prove groundbreaking for future design of adaptive tutoring content. Systems like ASSISTments already provide educational resources in a manner that has been shown to produce significantly greater learning gains than those found using traditional classroom practices [10]. This study suggests that learning outcomes can be further enhanced simply by adding support for a dynamic approach to content delivery through interleaving.

A major limitation of this study was the loss of a large portion of the original sample due to the failure of a participating teacher to assign the posttest to her students. It is possible that a larger sample would better reveal subtleties in the interaction between condition and student skill level. The sample distribution was also suboptimal, with random assignment resulting in more students in the interleaved condition than in the blocked condition. Further, analyses may have been weakened by the discretization of students as generally 'high' or 'low' skill. Departing from the use of a median split should be examined in future work.

Future iterations of this work should incorporate a pretest assignment and use novel skill content rather than skills intended for review. Future work should also examine variables pertaining to student performance within the *practice session* (i.e., average problem time, hint usage, and attempt count) to investigate Bjork's theory of desirable difficulties [2]. Additionally, future research should investigate more robust measures of learning, including extended retention rates following interleaved assignments and the effects on far transfer application.

6 Contribution

While many studies have examined the effect of interleaving, we offer a significant contribution to the field of artificial intelligence in education in that our work replicates the effect of interleaving within a brief homework assignment delivered using an adaptive tutoring system. Emphasized significance was observed for low skill students. Further, the use of homework assignments as both intervention and posttest resulted in the observation that rich features common to adaptive tutoring systems may allow researchers to pinpoint effects in variables other than correctness. The ease with which interleaving can be conducted within adaptive tutoring systems offers a low-cost, high-benefit approach to enhancing student learning outcomes.

Acknowledgments. We acknowledge funding from the NSF (1316736, 1252297, 1109483, 1031398, 0742503, 1440753), the U.S. Dept. of Ed. GAAN (P200A120238), ONR's "STEM Grand Challenges," and IES (R305A120125, R305C100024). Thanks to S.O. & L.P.B.O.

References

1. Birnbaum, M.S., Kornell, N., Bjork, E.L., Bjork, R.A.: Why interleaving enhances inductive learning: the roles of discrimination and retrieval. *Mem Cogn* **41**, 392–402 (2013)
2. Bjork, R.A.: Memory and metamemory considerations in the training of human beings. In: Metcalfe, J., Shimamura, A.P. (eds.) *Metacognition: Knowing about knowing*, pp. 185–295. MIT Press, Cambridge, MA (1994)
3. Cepeda, N.J., Pashler, H., Vul, E., Wixted, J.T., Rohrer, D.: Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psy Bulletin* **132**, 354–380 (2006)
4. Heffernan, N., Heffernan, C.: The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. *Int. J. of AI. in Ed.* **24**(4), 470–497 (2014)
5. Kornell, N., Bjork, R.A.: Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science* **19**, 585–592 (2008)
6. LeBlanc, K., Simon, D.: Mixed practice enhances retention and JOL accuracy for mathematical skills. 49th Annual Meeting of the Psychonomic Society, Chicago, IL (2008)
7. Li, Nan, Cohen, William W., Koedinger, Kenneth R.: Problem order implications for learning transfer. In: Cerri, Stefano A., Clancey, William J., Papadourakis, Giorgos, Panourgia, Kitty (eds.) *ITS 2012. LNCS*, vol. 7315, pp. 185–194. Springer, Heidelberg (2012)

8. Makel, M.C., Plucker, J.A.: Facts Are More Important Than Novelty: Replication in the Education Sciences. *Educational Researcher*. AERA (2014)
9. Mayfield, K.H., Chase, P.N.: The effects of cumulative practice on mathematics problem solving. *J. of Applied Behavior Analysis* **35**, 105–123 (2002)
10. Mendicino, M., Razzaq, L., Heffernan, N.T.: Comparison of Traditional Homework with Computer Supported Homework. *J. of Research on Tech in Ed.* **41**(3), 331–358 (2009)
11. Ostrow, K.: Materials for Study on Blocking vs. Interleaving, January 13 2015. <http://tiny.cc/AIED-2015-Interleaving>
12. Rau, M.A., Alevin, A., Rummel, N.: Interleaved practice in multi-dimensional learning tasks: Which dimension should we interleave? *Learning and Instr.* **23**, 98–114 (2013)
13. Rau, M.A., Alevin, V., Rummel, N., Pacilio, L., Tunc-Pekkan, Z.: How to schedule multiple graphical representations? A classroom experiment with an intelligent tutoring system for fractions. In: *The Future of Learning: Proceedings of the 10th ICLS* (2012)
14. Roediger, H.L.: Psychology's woes and a partial cure: the value of replication. *The Academic Observer*, The Association for Psychological Science (2012). <http://tiny.cc/RoedigerReplication>
15. Rohrer, D.: Interleaving helps students distinguish among similar concepts. *Educational Psychology Review* **24**, 355–367 (2012)
16. Rohrer, D., Pashler, H.: Recent research on human learning challenges conventional instructional strategies. *Educational Researcher* **39**(5), 406–412 (2010)
17. Rohrer, D., Taylor, K.: The shuffling of mathematics practice problems boosts learning. *Instructional Science* **35**, 481–498 (2007)
18. Shea, J.B., Morgan, R.L.: Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory* **5**(2), 179–187 (1979)
19. Taylor, K., Rohrer, D.: The effects of interleaved practice. *Applied Cognitive Psychology* **24**, 837–848 (2010)