Improving Students' Long-Term Retention Performance: A Study on Personalized Retention Schedules

Xiaolu Xiong
Worcester Polytechnic Institute
100 Institute Rd
Worcester, MA 01609
(508) 831-5000
xxiong@wpi.edu

Yan Wang
Worcester Polytechnic Institute
100 Institute Rd
Worcester, MA 01609
(508) 831-5000
ywang14@wpi.edu

Joseph Barbosa Beck
Worcester Polytechnic Institute
100 Institute Rd
Worcester, MA 01609
(508) 831-5000
josephbeck@wpi.edu

ABSTRACT
Traditional practices of spacing and expanding retrieval practices have typically fixed their spacing intervals to one or few predefined schedules [5, 7]. Few have explored the advantages of using personalized expanding intervals and scheduling systems to adapt to the knowledge levels and learning patterns of individual students. In this work, we are concerned with estimating the effects of personalized expanding intervals on improving students' long-term mastery level of skills. We developed a Personalized Adaptive Scheduling System (PASS) in ASSISTments' retention and relearning workflow. After implementing the PASS, we conducted a study to investigate the impact of personalized scheduling on long-term retention by comparing results from 97 classes in the summer of 2013 and 2014. We observed that students in PASS outperformed students in traditional scheduling systems on long-term retention performance (p = 0.0002), and that in particular, students with medium level of knowledge demonstrated reliable improvement (p = 0.0209) with an effect size of 0.27. In addition, the data we gathered from this study also helped to expose a few issues we have with the new system. These results suggest personalized knowledge retrieval schedules are more effective than fixed schedules and we should continue our future work on examining approaches to optimize PASS.

Categories and Subject Descriptors
H.4 Information Systems Applications; K.3.1 Computer Uses in Education; J.4 Social and Behavioral Sciences

General Terms
Algorithms, Measurement, Performance, Design, Theory.

LAK '15, March 16 - 20, 2015, Poughkeepsie, NY, USA

ACM 978-1-4503-3417-4/15/03…$15.00
http://dx.doi.org/10.1145/2723576.2723636

Keywords
Knowledge retention, retrieval practice, spacing effect, intelligent tutoring system, personalization

1. INTRODUCTION
1.1 Automatic Reassessment and Relearning System
Based on a robust memory phenomenon known as the *spacing effect* [4], expanding retrieval practice is often regarded as a superior technique for promoting long-term retention relative to equally spaced retrieval practice [3, 8]. Expanding retrieval practice works by, after the student learns a skill, having the student perform the skill at gradually increasing spacing intervals between successful retrieval attempts. Research has shown that spacing practice has a cumulative effect so that each time an item is practiced it receives an increment of strength [10]. This effect is specifically crucial to subjects such as mathematics: we are more concerned with students' capability to recall the knowledge that they acquired over a long period of time. What is more, the ability to retain a skill long-term is one of the three indicators of robust learning [2].

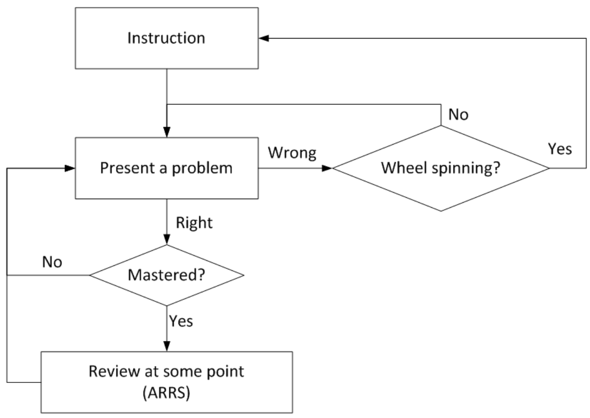

Figure 1. The enhanced ITS mastery learning cycle

Inspired by the importance of long-term retention and the design of the enhanced ITS mastery cycle in Figure 1 proposed by Wang and Beck [11], we developed and deployed a system called the Automatic Reassessment and Relearning System (ARRS) [13] to make decisions about when to review skills that students have mastered in ASSISTments, a non-profit, web-based tutoring system. ARRS is an implementation of expanding retrieval in the ITS environment. Unlike most ITS systems in which the tutoring stops if the student masters a given skill, ARRS assumes that if a student masters a skill with three correct responses in a row, such mastery is not necessarily an indication of long-term retention. Therefore, ARRS will present the student with retention tests on the same skill at expanding intervals spread across a schedule of at least 3 months. The default setting of the ARRS scheduling system uses a spacing interval of 7-14-28-56, and this indicates that each skill requires 4 level tests: the first level of retention tests takes place 7 days after the initial mastery; the second level of retention tests 14 days after successfully passing the first retention test, and so on. If a student answers incorrectly in one of these retention tests, ASSISTments will give him an opportunity to relearn this skill before redoing the same level of test.

Table 1. Retention performance by mastery speed and retention interval from pilot study

Retention test delay	# tests	% correctness
Mastery speed 3 – 4		
1 day	1186	84.4%
4 days	1169	82.2%
7 days	1171	81.7%
14 days	1233	81.2%
Mastery speed 5 – 7		
1 day	467	77.9%
4 days	432	76.2%
7 days	362	77.1%
14 days	420	73.1%
Mastery speed > 7		
1 day	280	67.5%
4 days	320	62.8%
7 days	267	59.6%
14 days	243	54.8%

In our previous studies [13, 14] of modeling student retention performance, we found that the number of problems required achieving mastery, which we referred to as the *mastery speed*, is an extremely important feature for predicting students' retention performance. We observed that, in general, the slower the mastery speed, the lower the probability that the student can answer the problems in the retention test correctly. Students who mastered a skill in 3 or 4 problems had approximately an 82% chance of responding correctly on the first retention test, while students who took over 7 attempts to master a skill only had a 62% chance [13]. Based on these results, we conclude that students with different mastery speeds have different retention patterns, so we began searching for the optimal retrieval schedules for different levels of student knowledge.

In order to find the optimal retention schedule for students and the best way to boost their performance in long-term mathematics learning, we conducted a pilot study by setting up four different interval schedules (1 day, 4 days, 7 days, and 14 days) and examined the impact on retention performance by comparing results across different groups of students. The results are shown in Table 1 and [12]. We saw a consistent decrease in retention performance with the longer retention intervals across in all students, no matter if they fell into the high mastery level, medium mastery level or low mastery level category. The results from Table 1 also demonstrated a main effect of mastery speed on retention performance: students with slower mastery speed had lower performance than students with a faster mastery speed; this statement is true even when we compared a 1-day performance of students with a mastery speed of over 7 (67.5% correct) speed versus a 14-day performance of students with a mastery speed of 3 or 4 (81.2% correct). A sizeable and interesting effect is that students with slower mastery speeds had bigger decreases in retention performance as retention intervals lengthened. For example, a high mastery level student had a decrease of 3.2% between 1 day tests and 14 days tests but the retention performance of low mastery level students dropped 12.7%. These results suggest retention intervals probably should vary, rather than be fixed, based on the student's knowledge of the skill.

1.2 Personalized Adaptive Scheduling System

Although ARRS helps students review knowledge after a time period, it neither knows a student's knowledge level, nor does it have the mechanism to change the retention schedule based on a particular student's performance. Here we formed a hypothesis that we can improve students' long-term retention levels by adaptively assigning students with gradually expanding and spacing intervals over time and we proposed to design and develop such a system, called Personalized Adaptive Scheduling System (PASS), as shown in Figure 2. PASS enables ARRS to schedule retention tests for students based on their knowledge levels. In the spring of 2014, we enhanced the traditional ARRS with the PASS and deployed it in ASSISTments.

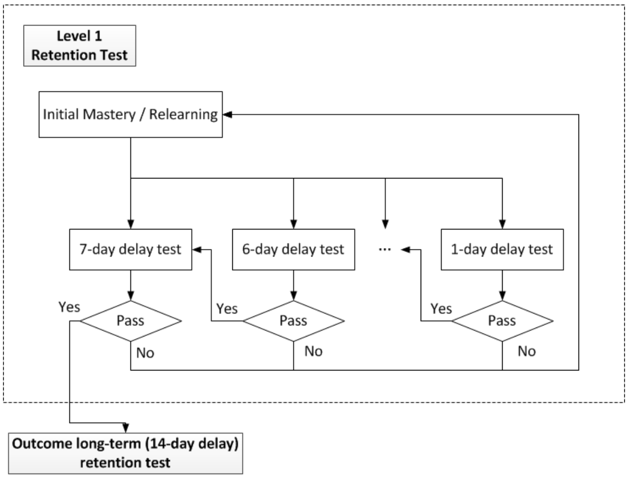

Figure 2. Design of Personalized Adaptive Scheduling System (PASS)

The current workflow of PASS aims to improve students' long-term retention performance by setting up personalized retention test schedules based on their knowledge levels. Here we rely on the mastery speed of a skill as an estimate of the student's knowledge and, consequently, predictor of retention performance. We retained the ARRS design of 4 expanding intervals of retention tests for each skill; however, PASS alters how the first interval behaves. When a student finishes initially learning a skill, we use his mastery speed to decide when to assign his first level 1 retention test. The mapping between mastery speed and retention delay intervals of the level 1 test is shown in Table 2. When a student passes the first test, PASS will schedule another test with a 1-day longer delay. Once the student passes the 7-day test, he is promoted to level 2 with a delay of 14 days. From that point on the intervals are the same as in the ARRS system. Note that mastery speed can be extracted from both students' initial learning and relearning processes. Therefore, when a student fails a retention test, a relearning assignment will be assigned to the student immediately. How quickly the student relearns this assignment will be used to set the interval for his next test. The mechanism of level 2 to level 4 tests is simpler. When a student fails a retention test, the retention delay will be reduced to the previous level (e.g., from 56 days to 28 days). It will be increased to the next level if the student passes the delayed retention test.

Table 2. Mapping between mastery speed and level 1 retention delays

Mastery Speed	Retention Delay
3	7
4	6
5	5
6	4
7	3
> 7	1

Here is an example of a student working with PASS in ASSISTments. Let's assume he needed 4 attempts to achieve three correct responses in a row in an initial learning assignment, so his mastery speed on this skill was 4. PASS then scheduled the first level 1 retention test for him to complete 6 days after the initial mastery. 6 days later, the student passed the retention test and PASS scheduled a 7-day retention test. Then a week later, the student passed the 7-day retention test and moved to the level 2 retention tests.

2. A STUDY ON IMPACT OF PERSONALIZED EXPANDING RETENTION INTERVALS

After the deployment of PASS in ASSISTments, several key issues were revealed that needed to be explored in order to realize the potential benefits of personalized expanding retention intervals and scheduling for students. We first conducted a study in ASSISTments to compare the new PASS with the traditional ARRS without PASS. In addition, this study explored the influence of personalized scheduling on students' long-term performance, student learning patterns and how they interact with the ASSISTments.

There were several objectives for this study. A central goal was to investigate potential long-term retention performance improvement to the benefit of personalized spacing schedules. We enabled PASS for all classes that were using ARRS on May 15, 2014; we expected students in these classes might be assigned homework during the next few months and thereby become the participants in the study. We ended this study on September 1, 2014 and found that 2,052 students from 40 classes were using PASS in the summer of 2014. Teachers of these classes assigned 93 different homework assignments to their students. Since traditional ARRS had been deployed in ASSISTments for over two years and a lot of data have been accumulated in the system, we extracted previous summer's ARRS-enabled classes that used the same assignments as the historical control group. 2,541 students from 57 classes in the summer of 2013 were qualified to act as historical control group.

During these two summer periods, students consistently received mathematics problem sets as homework assignments from their teachers. Once they answered three consecutive questions correctly in a problem set, students in the PASS condition would be given retention tests based on their mastery speed. If a student answered a retention test correctly, he was then given another retention test with a longer delay until he passed the level 1 test with a 7-day delay. On the other hand, students in traditional ARRS condition got 7-day delay retention tests after the mastery and went on with the 14-day tests if they answered the 7-day tests correctly. In this study, we defined how students performed on the 14-day retention tests (14 days after passing the level 1 test and at least 21 days after the initial mastery learning) as the outcome *long-term retention tests*. It is important to note that students usually receive several homework assignments and they may perform differently in these assignments, which means a student would have multiple tests that should be accounted for in the long-term performance. However, it is also possible that students do not complete assignments. Specifically, if a student has not finished the outcome retention test of a homework assignment by the end of this study, we cannot take this record into account.

3. RESULTS AND ANALYSIS

Retention test completion rate was calculated based on the number of homework assignments that had outcome tests answered divided by the total number of homework assignments. Days spent is the time interval between the start time of level 1 retention tests and the start time of outcome tests in days. Test count accounts for how many level 1 retention tests a student has to answer before this student can proceed to outcome tests. Students' long-term performance was calculated as the ratio of number of questions answered correctly in outcome tests to number of all questions answered in outcome tests.

3.1 Retention Test Completion Rate, Day Spent and Test Count

At the end of this study, the first result we noticed was that a lot of homework assignments in both groups did not have the records for associated outcome tests. In other words, a lot of students did not reach the 14-day retention tests. In the traditional ARRS condition, a total of 8404 homework assignments had been assigned to students but only 1,558 (18.5%) of these assignments had 14-days retention tests answered. When looking at the PASS condition, the retention test completion rate was even lower, only 1,029 (13.6%) of total 7,589 homework assignments had outcome tests answered. In one sense these low completion rates could result from the fact these homework and retention tests were assigned to students during the summer vacation so that perhaps many students did not treat these assignments seriously. The data also indicated the difference in the completion rates of the two conditions were statistically significant ($p < 0.0001$). We hypothesized that this was due to the fact that students in the

PASS condition took more tests in order to pass the 7-day delay tests. Remember, some medium- and low-knowledge students had to pass a number of shorter-delay tests to even reach the 7-day and then 14-day retention tests. To address this hypothesis, we investigated how many days were needed to reach the 14-day test from the beginning of level 1 retention tests. The data was grouped by the three identified mastery speed bins to represent high-, medium- and low-knowledge students on their homework assignments

Table 3. Average day spent of each knowledge level by conditions

Initial mastery performance	ARRS	PASS	*p*-value
Mastery Speed 3 - 4	16.80	18.96	0.0002
Mastery Speed 5 - 7	17.67	33.24	0.0001
Mastery Speed > 7	17.34	32.33	0.0001

Table 3 describes the differences in average days spent between ARRS and PASS conditions. The minimum possible delay is 14 days, achievable for ARRS students who answer the 7-day test correctly, and then take their ARRS test when it is immediately available. Students who failed the first ARRS test would have to take one or more additional 7-day tests until they responded correctly and could be promoted to the 14-day test. For the PASS condition, 14 days is a lower bound only for those students with an initial mastery speed of 3, as slower mastery speeds would require multiple first-level tests before being promoted to the 14-day interval. As expected, students in the PASS condition spent more time in the practices of level 1 retention tests; especially for medium- and low-knowledge students who spent nearly two more weeks in the process of passing the 7-day delay tests relative to ARRS students. Table 4 demonstrates that students in the PASS condition had more tests to answer by showing the average test count of the two conditions therefore it took them more days to reach 14-day tests.

Table 4. Average test count of each knowledge level by conditions

Initial mastery performance	ARRS	PASS	*p*-value
Mastery Speed 3 - 4	1.34	1.21	0.0003
Mastery Speed 5 - 7	1.44	3.25	0.0001
Mastery Speed > 7	1.59	3.69	0.0001

3.2 Long-Term Retention Performance

After it was observed that PASS made students take more practice in the retention tests, we became more curious about the impact of PASS on long-term retention performance. It is important to emphasize that students were balanced with respect to proficiency in the ARRS and PASS conditions given their close homework performance level: 71.0% correct versus 71.2%. An initial analysis on long-term retention performance across all students showed the PASS condition (83.4%) outperformed the ARRS condition (77.2%) with a reliable but small improvement (p = 0.0002, effect size = 0.15). When considering the performance changes in different knowledge level of students, we again grouped the data by three identified mastery speed bins; then we examined students' long-term retention performance with p-values and effect sizes.

Table 5. Long-term (14-day) retention performance comparison and sample size (in parenthesis)

Initial mastery performance	ARRS	PASS	*p*-value	Effect size
Mastery Speed 3 – 4	81.79% (978)	83.91% (889)	0.2266	0.06
Mastery Speed 5 – 7	73.08% (327)	84.53% (97)	0.0209	0.27
Mastery Speed > 7	64.82% (253)	70.59% (51)	0.4301	0.12

The comparison of long-term retention performance shows that all three groups of students in the PASS condition outperformed those in the ARRS condition, although the improvements were not all statistically significant; only students with medium-knowledge on skills performed reliably better with an effect size of 0.27. For students with high knowledge on skills, the benefit of using PASS was limited; this suggests that solely relying on 7-day delay tests is sufficient for this population. A previous study [12] also suggested that high-knowledge students have high resistance against forgetting. On the other hand, providing low-knowledge students with more spaced retention tests and relearning assignments did not stop the decay of retention even after these students had approximately 3 additional relearning assignments on the same skill, and we only noticed a small effect size (0.12) improvement on the retention performance. Because PASS employs a higher stand of mastery and retention, thus few low-knowledge students reached outcome tests; we in fact noticed that only 51 tests had been completed, so this also prevented us from achieving a higher effect size in PASS condition. Another notable result was when we compared Table 5 vertically: we could see that PASS helped to close the performance gap between different groups of students. In fact, in the PASS condition, the long-term performance of medium-knowledge students even outperformed the high-knowledge students. Of course, the small sample size tells us we need more studies to validate this result.

4. CONTRIBUTIONS, FUTURE WORK AND CONCLUSIONS

The paper makes three contributions. First, the work behind this paper designed and deployed a personalized expanding interval scheduling system that utilizes spacing effect in the field. Through the participation of thousands of students, we carried out a study to test the idea of assigning students with different delays of retention tests to help them better retain skills. As the first study on this system, the paper explores the path of improving ITS to help students achieve robust learning via personalized expanding retrieval practices. The second contribution of this paper is a validation of the hypothesis that students' long-term performance can be improved by giving them tests that are well spaced out and scheduled appropriately, before gradually expanding the spacing between these tests. Most importantly, this study demonstrates the importance of individualization in scheduling retention tests, as it shows that students with medium knowledge can match up their long-term performance with high-knowledge students by using PASS. The third contribution of this paper is the confirmation of concept of finding the optimal retention interval by using mastery speed as a measurement of students' knowledge level. By using mastery speed to group students, we can distinguish different learning and retention patterns among students with different knowledge levels. In the process of work, we have noticed that there has been other work on retention, such as the personalized spaced review system [6]; however, this work focuses on fact retrieval and is able to make far stronger assumptions of when students are exposed to content. Our work examines a procedural skill, in a classroom context where we cannot be sure what material teachers cover in class and we are not aware of all homework assignments, thus we cannot be sure when students last saw a skill.

This PASS and its implementation in ASSISTments have been introduced to the field for just a few months, so we are still at the initial phase of study. Our goal is to find the optimal spacing schedules for students and the best way to boost their performance in long-term mathematics learning. There are many further problems that we are interested in: What should we do to help low-knowledge students, considering the improvement we saw in the study was so small, particularly given the increased amount of practice they received? From the data we collected, it was obvious that there were some areas that required improvement. For example, we simulated a scenario to improve the retention performance of low-knowledge students to match up to the performance level of high-knowledge students (83.91%) and also improve completion rates to the level of ARRS condition so we could collect 228 data points. Given these optimistic assumptions, there intervention would have an effect size of 0.45. Thus, in this scenario, achieving a medium effect size (0.5) is not feasible. What is the fundamental cause of mistakes? Lack of effort or interest on the student's part, or a genuine lack of knowledge [1]? How can we increase the completion rate? Most importantly, how can we solve the optimization problem to balance time cost and performance improvement [9]? Is there a better way than just assigning high-frequency retention tests to students?

This paper presents the initial study of using the personalized adaptive scheduling system to explore a solution to the optimal spacing schedule problem. With the experiment data we collected, we are excited to see that the PASS can help to improve long-term retention performance across all three groups of students and become the backbone of future development for promoting student robust learning.

5. ACKNOWLEDGMENTS

We acknowledge funding for ASSISTments from NSF (# 1440753, 1316736, 1252297, 1109483, 1031398, and 0742503), ONR's 'STEM Grand Challenges' and IES (# R305A120125 & R305C100024) grants.

6. REFERENCES

[1] Anderson, J. R. (2014). *Rules of the mind*. Psychology Press.

[2] Baker, R. S., Gowda, S. M., Corbett, A. T., & Ocumpaugh, J. (2012, January). Towards automatically detecting whether student learning is shallow. In *Intelligent Tutoring Systems* (pp. 444-453). Springer Berlin Heidelberg.

[3] Crowder, R. G. (1976). *Principles of learning and memory.*

[4] Hintzman, D. L. (1974). *Theoretical implications of the spacing effect*.

[5] Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition, 15(3), 257-280.*

[6] Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological science,* 25(3), 639-647.

[7] Karpicke, J. D., & Roediger III, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 704.

[8] Melton, A. W. (1967). Repetition and retrieval from memory. Science (New York, NY), 158(3800), 532-532.

[9] Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology*: Applied, 14(2), 101.

[10] Thalheimer, W. (2006). *Spacing learning events over time: What the research says.*

[11] Wang, Y., & Beck, J. E. (2012). Using Student Modeling to Estimate Student Knowledge Retention. International Educational Data Mining Society.

[12] Xiong, X., & Beck, J. E. (2014). A Study of Exploring Different Schedules of Spacing and Retrieval Interval on Mathematics Skills in ITS Environment. In *Intelligent Tutoring Systems* (pp. 504-509). Springer International Publishing.

[13] Xiong, X., Li, S., & Beck, J. E. (2013). Will You Get It Right Next Week: Predict Delayed Performance in Enhanced ITS Mastery Cycle. In FLAIRS Conference.

[14] Xiong, X., Adjei, S. A., & Heffernan, N. T. (2014) Improving Retention Performance Prediction with Prerequisite Skill Features. The 7th International Conference on Educational Data Mining