Comparing Adaptive and Random Spacing Schedules during Learning to Mastery Criteria

Everett Mettler (mettler@ucla.edu)[1]
Christine M. Massey (cmassey@psych.ucla.edu)[1]
Timothy Burke (mizerai@ucla.edu)[1]
Philip J. Kellman (kellman@cognet.ucla.edu)[1]

[1]Department of Psychology, University of California, Los Angeles
Los Angeles, CA 90095 USA

Abstract

Adaptive generation of spacing intervals in learning using response times improves learning relative to both adaptive systems that do not use response times and fixed spacing schemes (Mettler, Massey & Kellman, 2016). Studies have often used limited presentations (e.g., 4) of each learning item. Does adaptive practice benefit learning if items are presented until attainment of objective mastery criteria? Does it matter if mastered items drop out of the active learning set? We compared adaptive and non-adaptive spacing under conditions of mastery and dropout. Experiment 1 compared random presentation order with no dropout to adaptive spacing and mastery using the ARTS (Adaptive Response-time-based Sequencing) system. Adaptive spacing produced better retention than random presentation. Experiment 2 showed clear learning advantages for adaptive spacing compared to random schedules that also included dropout. Adaptive spacing performs better than random schedules of practice, including when learning proceeds to mastery and items drop out when mastered.

Keywords: adaptive learning; spacing effect; memory; optimal practice; mastery learning

Introduction

The spacing effect is a simple and powerful driver of gains in learning, requiring only changes in the temporal distribution of practice across time (Pashler, Bain, Bottge, Graesser, Koedinger, McDaniel, & Metcalfe, 2007). A prominent theoretical account of spacing effects - the retrieval effort hypothesis (Pyc & Rawson, 2009; c.f. Bjork, 1994) - suggests that benefits from spacing arise because of the difficulty of retrievals after a spacing delay. According to this theory difficult retrievals are due to partial forgetting of information across widely spaced presentations. ARTS (Adaptive Response-Time-based Sequencing) is a technique for adaptively adjusting the spacing delays between trials during a learning session to encourage the greatest possible benefits of spacing (Mettler, Massey & Kellman, 2016; Mettler & Kellman, 2014). ARTS produces spacing delays that tend to optimize learning in light of ongoing learning strength for individual items and learners, using reaction time along with accuracy as a proxy for learning strength. ARTS attempts to stretch spacing up to, but just short of, the point of forgetting, an approach that has been shown to increase retention of information in a variety of domains including fact learning and perceptual learning (Mettler & Kellman, 2014).

Mettler, Massey & Kellman (2016) showed that in factual learning, adaptive spacing in ARTS outperformed fixed spacing schedules in which the total number of presentations per item in a session was limited (e.g. 4 presentations per item), consistent with other studies of the spacing effect (Karpicke & Roediger, 2007). In the studies presented here we asked whether the benefits of adaptive spacing apply under conditions more consistent with "real-world" learning, where learning sessions are not limited to a few presentations per item. Instead, presentations continued until learning reached a standard of proficiency, i.e., mastery learning. We compared adaptive schedules to random schedules and observed how mastery learning affects learning gains.

Mastery Learning

Mastery learning (Bloom, 1974) treats learners as individuals having different learning requirements, especially in terms of the amount of time needed to achieve similar performance standards, unlike contemporary instruction where learners are graded on units of instruction whether or not successful mastery has been accomplished. Mastery learning motivated the development of adaptive learning techniques and adaptive curricula such as "programmed instruction" (Holland and Skinner, 1961; Keller, 1967), and curricula implementing mastery learning have resulted in learning gains superior to traditional instruction (Kulik, Kulik & Bangert-Drowns, 1990).

Mastery learning is ripe for revival. In computer-based adaptive learning, technological advances have made tracking of individual learning items and assessment of instructional objectives easier to achieve; mastery has become a goal of many adaptive learning systems (Ritter et al., 2016). At the same time, advances in our understanding of spacing effects are revealing crucial interactions between spacing and learning conditions including criterion learning levels that mediate effects of spacing (Vaughn, Dunlosky & Rawson, 2016).

Learning Criteria How do learning (mastery) criteria affect learning? Studies show learning criteria positively

affect associative retrieval and that increases in the stringency of learning criteria have a logarithmic relationship with later recall (Vaughn & Rawson, 2011). Stronger criteria result in greater learning, but with diminishing returns on learning for every unit increase in the strictness of the criterion. Criterion level also interacts with spacing. Pyc and Rawson (2009) had participants learn Swahili-English word pairs under two spacing conditions (long and short spacing) and a number of criterion levels - roughly 1 to 10 correct retrievals. Long spacing intervals generally improved learning and increases in the strictness of the criterion level produced increasing but diminishing learning gains. The greatest learning gains came from the highest criterion levels and longest spacing. In these studies, however spacing was not adaptive. Learning criteria need to be evaluated in relation to the kinds of ongoing schedule dynamics during learning interventions using adaptive spacing of learning events.

In ARTS, learning criteria include both accuracy and speed. Further, accurate and fast responses cause future items to have longer spacing between presentations. The result is that the criteria enforce widely spaced, difficult retrievals, not just correct or fast ones.

Dropout How does retirement or "dropout" - the removal of well-learned items during learning - affect performance? Criteria that poorly estimate learning strength may harm learning when combined with dropout. For instance, Kornell and Bjork (2007) found that dropout was a common strategy that students employed while studying, but that students' mis-estimations of their learning resulted in dropout lowering their overall learning gains. Dropping items based on objective learning criteria seems to fare better. Pyc and Rawson (2011) showed that dropout based on learning criteria improved learning as measured by efficiency (the amount retained at a test per trial of training invested) compared to fixed schedules of practice where there was no dropout. This result was not without caveats. Recall accuracy was better for fixed schedules than for dropout schedules possibly due to a weak criterion (1 correct response), and non-dropout schedules may have undergone overlearning - a condition where learners benefit from further practice even when performance is at ceiling (Underwood, 1964). Unsophisticated dropout algorithms may lead to either difficult items dropping out prematurely or well-learned items being overpracticed without undergoing overlearning (Vaughn, Rawson, & Pyc, 2013). In adaptive learning, dropout would likely benefit from criteria that also include reaction time. We know of no previous studies or learning systems other than ARTS that have used response-time criteria in dropout, or any kind of dropout criteria that relate to adaptive spacing. In the following studies we attempt to compare and contrast reasonable combinations of such features across adaptive spacing vs. fixed spacing schedules.

Current Studies

We compared adaptive and random schedules. In the adaptive schedule condition, using ARTS, learning continued until each item met mastery criteria, after which individual items were dropped from the learning set. In the random schedule condition, learning items were presented randomly and items were not dropped from the session; instead the session was terminated after every item had met mastery criteria. In a second experiment, we compared adaptive presentation with random presentation, where the adaptive and random conditions had identical learning criteria and both schedules included dropout. We compared the efficiency of learning across the two scheduling conditions at both immediate and delayed tests of retention.

Random Schedules Despite the focus in the experimental literature on organized retrieval practice (e.g. adaptive or fixed expanding spacing), it bears mentioning that random schedules naturally implement a form of spaced practice. Spacing intervals for items in random schedules are on average as large as the number of items in the learning set. In addition, random practice increases encoding context variability relative to more constrained fixed schedules. That is, each practice with an item is usually preceded by and followed by a different set of items; conditions that some theories of spacing claim are beneficial for learning (Howard & Kahana, 1999; Maddox, 2016). Interestingly, few studies in the literature on scheduling and spacing include random practice as a comparison. Random schedules are understudied and may provide a window into conditions of practice that are beneficial for learning.

Dependent Measures and Data Analysis Our primary measure of learning performance was learning *efficiency*, defined as accuracy gain from pretest to posttest divided by the number of trials invested in learning and multiplied by the number of learning items. Efficiency gives a way of measuring learning that incorporates variations in both posttest performance and the number of learning trials required to reach mastery criteria. It may be thought of as a rate measure, indicating performance improvement per item per learning trial, with a maximum value of 1. We also examined learning at equivalent points during the learning phase, and raw accuracy change scores between pre and posttests. All measures were assessed using standard parametric statistics such as ANOVA and planned comparisons between conditions. All statistical tests were two-tailed, with a 95% confidence level; all effect sizes are Cohen's d; and all error bars in graphs show +/- 1 standard error of the mean.

Experiment 1: Random vs. Adaptive Scheduling

Method

Participants Participants were 48 UCLA undergraduates who participated for course credit.

Design The experiment used a pretest/posttest design. There was a pretest, training phase, immediate posttest and a delayed posttest administered after 1 week. There were 2 between-subjects conditions, Adaptive and Random, that manipulated the scheduling of items during the training phase. Adaptive scheduling was determined dynamically for each participant using the ARTS algorithm. After every response, ARTS calculates a priority score for each learning item and compares scores across items to determine which item will be presented next. Equation 1 shows the priority score calculation.

$$(1) \qquad P_i = a(N_i - D)[b(1 - \alpha_i)\, Log(RT_i / r) + \alpha_i W]$$

Detailed description of the ARTS algorithm can be found in previous work (Mettler, Massey & Kellman, 2011, 2016). The parameters of the adaptive algorithm were the same as those in Mettler, Massey & Kellman (2016).

Random scheduling consisted of purely random presentation where on each trial for each participant, a random item was selected for presentation.

Materials The learning items were 24 African countries participants were required to identify on a map of Africa. There were no filler items. All material was presented on a computer within a web-based application. Participants saw a 500 x 800-pixel map of Africa on the left side of the screen and a two-column list of African countries alphabetically organized by column then row. Each list label was a software button that could be selected independently.

Procedure In all sessions of the experiment, learning items were presented singly, in the form of interactive test trials. Participants were shown a map of Africa featuring an outlined country and were asked to select, using a mouse, from a list of 24 names the name matching the highlighted country. In the Adaptive condition each item was learned until it reached mastery criteria and was then dropped from the set. In the Random condition, each item was tracked so that the experiment session ended after every item had reached mastery criteria, or after the learning session reached 45 minutes, whichever came first. There was no dropout of items during the learning session in the Random condition. The learning criterion was five out of the last five presentations of an item correct with all five response-times less than seven seconds.

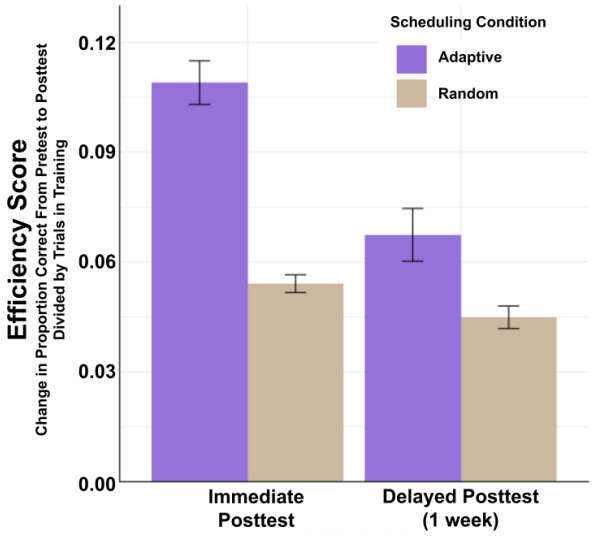

Figure 1: Learning efficiency in pretest and delayed posttest by scheduling condition in Experiment 1.

Results

Pretest Accuracy Pretest accuracies were roughly equal across conditions (Adaptive: M=0.078, SD=0.051; Random: M=0.083, SD=0.117) and not significantly different (t(46)=0.2, p=.84; d=0.06).

Learning Efficiency Results for efficiency at immediate and delayed posttests are shown in Figure 1. A 2x2 mixed factor ANOVA comparing efficiency across scheduling condition (adaptive vs. random) and posttest phase (immediate vs. delayed) found significant main effects of condition ($F(1,46)=33.83$, $p<.001$, $\eta_p^2=0.424$), a main effect of posttest phase ($F(1,46)=89.69$, $p<.001$, $\eta_p^2=0.661$), and a significant scheduling condition by test phase interaction ($F(1,46)=36.6$, $p<.001$, $\eta_p^2=0.443$). At immediate posttest, efficiencies were higher in the adaptive condition (M=0.109, SD=0.03) than the random condition (M=0.054, SD=0.012) a significant difference ($t(46)=8.53$, $p<.001$, $d=2.68$). This outcome represents 102% greater efficiency in the adaptive condition at immediate posttest. At delayed posttest, efficiencies were also higher in the adaptive condition (M=0.067, SD=0.04) than the random condition (M=0.045, SD=0.015), a significant difference ($t(46)=2.87$, $p=.006$, $d=0.892$). These differences comprise 50% greater efficiency in the adaptive condition at delayed posttest. Comparing means between the two test phases, the difference between efficiencies at each test phase for both the adaptive and random condition were significant (adaptive imm. vs. delayed, $t(23)=5.88$, $p<.001$, $d=0.68$; random immediate vs. delayed, $t(23)=8.11$, $p<.001$, $d=1.29$). The interaction appeared to be the result of declining efficiency in the adaptive condition from immediate to

delayed posttest, but smaller decline in the random condition.

Trials to Criterion Learning took differing amounts of time across participants and conditions, so the number of trials to criterion (Figure 2) was analyzed. The random condition took on average 429 trials to reach the end of the session (SD=97.7). The adaptive condition took on average 197 trials (SD=43.5). This difference was significant ($t(46)=10.6$, $p<.001$, $d=3.29$). Two participants in the random condition and one participant in the adaptive condition did not retire all items (16, 9, and 1 items were retired respectively).

Equivalent Learning Trials Analysis To assess whether differing numbers of learning trials were the only driver of learning differences between conditions, we carried out an additional analysis comparing conditions at points when they had the same number of learning trials. Specifically, we determined the mean number of trials to criterion in the adaptive condition (197) and compared accuracy between conditions at that number of trials and at earlier points. Using blocks of trials consisting of 3 trials per item, we performed the equivalent trials analysis starting at 5 blocks prior to trial 197. As can be seen in Figure 3, results showed higher accuracies at all points in the adaptive condition.

A 2X4 mixed factor ANOVA on schedule condition and trial block was conducted (trial block 5 was not included since some participants did not have a 5th trial block). The ANOVA found main effects of scheduling condition ($F(1,46)=11.9$, $p<.01$) and trial block ($F(3,138)=128.5$, $p<.001$), but no condition by trial block interaction ($F(3,138)=1.88$, $p=.14$). Independent t-tests were conducted at each trial block. Accuracies were reliably higher for the adaptive condition than the random condition at blocks 1, 2, and 3 ($ts(46)=5.83$, 2.5, and 3.43 respectively, all $ps<.05$) but not at blocks 4 and 5 (block 4, $t(46)=1.16$, $p=.25$; block 5, $t(35)=0.354$, $p=.73$).

Accuracy Change Between Pretest and Posttests Accuracy was compared across conditions using a change score between pre and posttests (Posttest accuracy minus pretest accuracy). A 2X2 mixed factor ANOVA on scheduling condition and posttest phase showed a significant main effect of condition ($F(1,46)=17.75$, $p<.001$, $\eta_p^2=0.28$), a main effect of posttest phase ($F(1,46)=105.49$, $p<.001$, $\eta_p^2=0.70$), and a significant test phase by condition interaction ($F(1,46)=10.92$, $p=.001$, $\eta_p^2=0.19$). At immediate posttest change scores were higher for the random condition ($M=0.93$, $SD=0.09$) than for the adaptive condition ($M=0.85$, $SD=0.12$), a significant difference ($t(46)=2.55$, $p=.01$, $d=0.746$). At delayed posttest, change scores were also higher for the random condition ($M=0.76$, $SD=0.16$), than for the adaptive condition ($M=0.52$, $d=2.68$). Comparing conditions across test phases, the difference between immediate and delayed change scores for the random condition was significant ($t(23)=5.38$, $p<.001$, $d=1.374$) as was the difference for the adaptive condition ($t(23)=8.9$, $p<.001$, $d=1.967$).

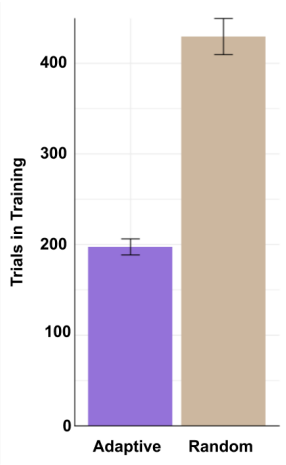

Figure 2: Learning trials by condition in Experiment 1.

Experiment 1 Discussion

We found significantly greater learning efficiency in conditions where learning was scheduled using ARTS with learning to criterion and dropout than in a random presentation condition that had no adaptive, mastery or dropout features. Efficiencies were higher in the adaptive condition for both immediate and delayed posttests. Efficiencies were 102 percent higher for adaptive than random at immediate posttest and 50 percent higher at delayed posttest, and the effect sizes of these differences were large. In addition, when compared at equivalent points during learning, average accuracies were higher in the adaptive than in the random condition. Delayed gains persisted in the adaptive condition despite a degree of

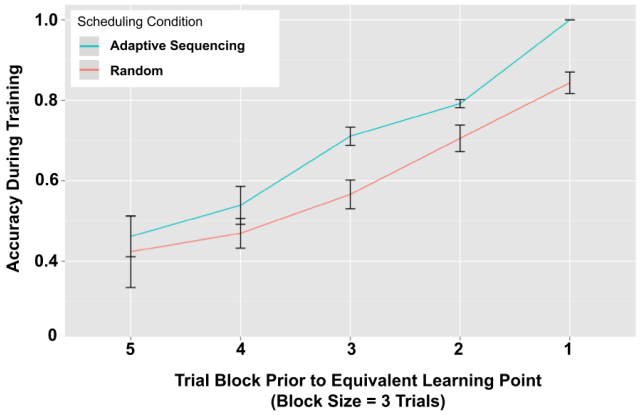

Figure 3: Learning session accuracy at equivalent points by blocks of 3 presentations of each item in Experiment 1.

overlearning that occurred in random schedules - nearly twice as many presentations of each item, and overall higher accuracies in the random condition. Experiment 1 shows that adaptive scheduling with dropout provides a robust advantage in learning efficiency over random scheduling.

In a second experiment we compared adaptive sequencing to a random schedule of practice that also included dropout.

Experiment 2: Random vs. Adaptive Presentation (Random With Dropout)

Method

Participants Participants were 48 UCLA undergraduates, some of whom participated for course credit and some of whom were recruited and paid $16 for their time.

Design The design was identical to Experiment 1, except that the random condition was altered to include dropout. After each item reached a learning criterion it was removed from the set of learning items. The learning criteria were the same as in Experiment 1.

Materials & Procedure The materials and procedure were identical to Experiment 1.

Results

Learning Efficiency Results for efficiency at immediate and delayed posttests are shown in Figure 4. A 2x2 mixed factor ANOVA comparing efficiencies across scheduling condition (Adaptive vs. Random) and posttest phase (Immediate vs. Delayed) found significant main effects of condition ($F(1,46)=10.6$, $p=.002$, $\eta_p^2=0.188$), a main effect of posttest phase ($F(1,46)=163.28$, $p<.001$, $\eta_p^2=0.78$), but no significant scheduling condition by test phase interaction ($F(1,46)=1.83$, $p=.18$, $\eta_p^2=0.039$). At immediate posttest, efficiencies were higher in the Adaptive condition ($M=0.12$, $SD=0.02$) than the Random condition ($M=0.09$, $SD=0.02$) a significant difference ($t(46)=4.07$, $p<.001$, $d=1.17$). At delayed posttest, efficiencies were also higher in the Adaptive condition ($M=0.084$, $SD=0.03$) than the Random condition ($M=0.064$, $SD=0.064$), a significant difference ($t(46)=2.31$, $p=.025$, $d=0.67$). The difference in efficiency between the two phases (immediate vs. delayed) was significant for both adaptive ($t(23)=9.05$, $p<.001$, $d=1.35$) and random scheduling ($t(23)=9.14$, $p<.001$, $d=1.207$).

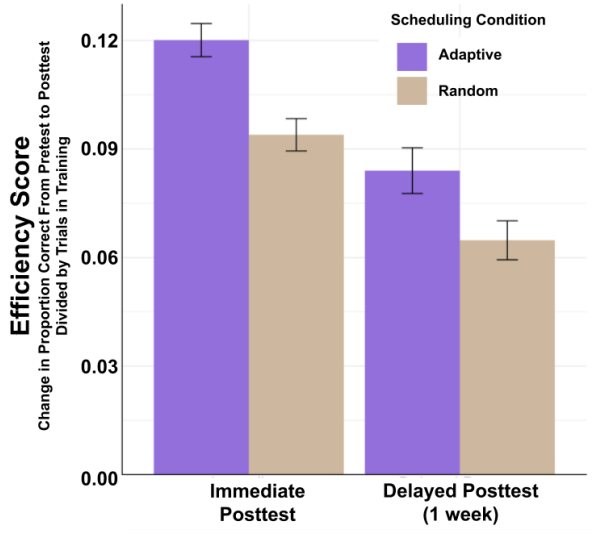

Figure 4: Efficiency at immediate and delayed posttests by scheduling condition in Experiment 2.

Trials To Criterion Despite identical dropout features, trials to criterion varied with condition (see Figure 5). The random condition took on average 231 trials to reach the end of the session (SD=48.3). The adaptive condition took on average 183 trials (SD=35.1). This difference was significant ($t(46)=3.88$, $p<.001$, $d=1.13$). Two participants in the random condition and one participant in the adaptive condition did not retire all items (23, 21 and 23 items were retired respectively).

Equivalent Learning Trials Analysis Accuracies were compared at equivalent points between Adaptive and Random conditions. A 2X3 mixed factor ANOVA on schedule condition and trial block was conducted (trial blocks 4 and 5 were not included because some participants did not have a complete 4th or 5th trial block). The ANOVA found main effects of scheduling condition ($F(1,46)=45.1$, $p<.001$) and trial block ($F(2,92)=285.7$, $p<.001$), and a significant condition by trial block interaction ($F(2,92)=24.9$, $p<.001$). Comparisons were conducted at each trial block. Accuracies were higher for the adaptive

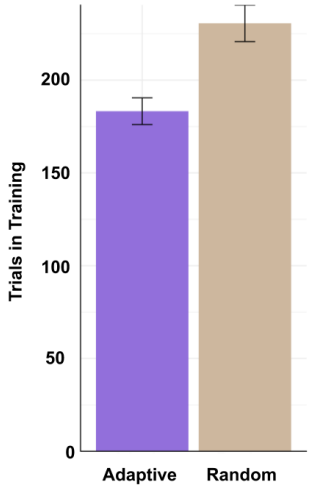

Figure 5: Number of trials in learning session for each scheduling condition in Experiment 2.

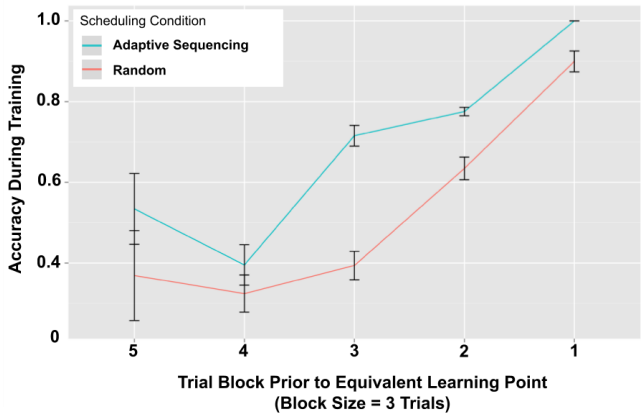

Figure 6: Learning session accuracy at equivalent points by blocks of 3 presentations of each item in Experiment 2.

condition than the random condition at blocks 1, 2, and 3 (ts(46)=3.91, 4.71, and 7.38 respectively, all p's<.001) but not at blocks 4 and 5 (block 4, t(45)=1.04, p=.30; block 5, t(26)=1.14, p=.26).

Pretest Accuracy Pretest accuracies were roughly equal across conditions (Adaptive: M=0.083, SD=0.12; Random: M=0.071, SD=0.08) and not significantly different (t(46)=0.48, p=.63; d=0.14).

Accuracy Change Accuracy change was computed between pretest and posttest. No difference between conditions was found (F(1,46)=0.106, p=.746, η_p^2 =0.002) and there was no interaction with test phase (F(1,46)=0.006, p=.937, η_p^2 < 0.001).

Experiment 2 Discussion

As in Experiment 1, efficiencies were higher for adaptive scheduling than random scheduling despite both schedules having identical mastery criteria (including both accuracy and reaction time criteria). The efficiencies in the Adaptive condition were 33 percent higher than the Random condition at immediate posttest, and 31 percent higher than the random condition at delayed posttest, with large and medium effect sizes respectively. An analysis of learning session accuracies at equivalent points in training found that learners were on average more accurate in the Adaptive condition than in the Random condition. These differences provide strong evidence that the superiority of adaptive schedules in learning derives from advantageous spacing above and beyond efficiency gains due to dropout during learning. Learning was enhanced by dropout as noted by accuracy increases in the random condition between Experiment 1 vs 2. However, the effect of dropout was not greater than the benefits of adaptive scheduling.

Conclusion

In two experiments we demonstrated learning advantages for adaptive schedules of practice under conditions of mastery learning. Learning efficiency improved when the schedule of presentation of items was determined by an adaptive method of spacing vs. a random schedule of practice. Adaptive spacing was generated using ARTS, Adaptive, Response-Time based Sequencing (Mettler, Massey, Kellman 2016) where learning items were individually spaced according to measures of ongoing learning strength estimated by response time. Experiment 1 demonstrated that adaptive sequencing with dropout produces greater learning than random presentations without dropout of items. In Experiment 2, random schedules included dropout with identical retirement criteria to the adaptive condition. Experiment 2 showed that adaptive schedules were more efficient than random schedules even when schedules were equated for learning criteria and dropout.

Mastery learning is thought to promote the best learning outcomes, and the present results indicate that adaptivity of spacing delays drives efficient learning when individual learners are tracked until reaching competence with each item. Dropout also showed positive effects on learning. As this outcome does not always occur (Vaughn, Rawson, & Pyc, 2013), the results suggest that dropout might be most advantageous when combined with efficient learning schedules and well-chosen mastery criteria.

Adaptive schedules performed better than random, fixed schedules. Among a variety of possible fixed schedule types, random schedules produce the longest absolute spacing delays and the most spacing variability. By some theories of spacing, these advantages should produce the best performance (Karpicke & Bauernschmidt, 2011; Glenberg, 1976; Maddox, 2016). Despite these advantages of random spacing, ARTS produces better learning because spacing is appropriate to the needs of individual learners and items during learning. Appropriate delays appear to be those that stress the learner's ability to remember across the spacing delay but do not result in forgetting during learning. Using both response times and accuracy data allows ARTS to dynamically adjust spacing intervals to meet these criteria for individual learners, items, and their interactions. Random schedules, which also include a robust set of spacing delays for each item, cannot match these specific and fluctuating needs of learners during learning. Similar results have been found when adaptive schedules are compared with fixed equal and fixed expanding schedules (Mettler, Massey & Kellman, 2016).

Adaptive schedules outperform fixed (predetermined) spacing schedules and random schedules of practice, both during the course of learning and when learning proceeds to objective criteria of mastery. The strength and generality of

these results has important implications for the design of learning interventions and learning technology.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grants No. 1109228 and No. 1644916. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Bjork, R. A. (1992). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge, MA: MIT Press..

Bloom, B. S. (1974). Time and learning. *The American Psychologist, 29*(9), 682–688.

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior*, *15*(1), 1–16.

Holland, J. G., & Skinner, B. F. (1961). *The analysis of behavior: A program for self-instruction*. New York: McGraw Hill.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 25*(4), 923–941.

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 37*(5), 1250–1257.

Karpicke, J. D., & Roediger, H. L., 3rd. (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), 704–719.

Keller, F. S. (1967). Engineering personalized instruction in the classroom. *Revista Interamericana de Psicologia/Interamerican Journal of Psychology, 1*(3).

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*(2), 219–224.

Kulik, C.-L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research*, *60*(2), 265–299.

Maddox, G. B. (2016). Understanding the underlying mechanism of the spacing effect in verbal learning: a case for encoding variability and study-phase retrieval. *Journal of Cognitive Psychology* , *28*(6), 684–706.

Mettler, E., & Kellman, P. J. (2014). Adaptive response-time-based category sequencing in perceptual learning. *Vision Research, 99*, 111–123.

Mettler, E., Massey, C. M., & Kellman, P. J. (2011). Improving adaptive learning technology through the use of response times. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2532-2537). Austin, TX: Cognitive Science Society.

Mettler, E., Massey, C. M., & Kellman, P. J. (2016). A comparison of adaptive and fixed schedules of practice. *Journal of Experimental Psychology. General, 145*(7), 897–917.

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing Instruction and Study to Improve Student Learning*. IES Practice Guide. NCER 2007-2004. National Center for Education Research.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*(4), 437–447.

Ritter, S., Yudelson, M., Fancsali, S. E., & Berman, S. R. (2016). How mastery learning works at scale. *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, 71–79.

Underwood, B. J. (1964). Degree of learning and the measurement of forgetting. *Journal of Verbal Learning and Verbal Behavior, 3*(2), 112–129.

Vaughn, K. E., Dunlosky, J., & Rawson, K. A. (2016). Effects of successive relearning on recall: Does relearning override the effects of initial learning criterion? *Memory & Cognition, 44*(6), 897–909.

Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion-level effects on memory: what aspects of memory are enhanced by repeated retrieval? *Psychological Science, 22*(9), 1127–1131.

Vaughn, K. E., Rawson, K. A., & Pyc, M. A. (2013). Repeated retrieval practice and item difficulty: Does criterion learning eliminate item difficulty effects? *Psychonomic Bulletin & Review, 20*, 1239-1245.