Maximizing Students' Retention Via Spaced Review: Practical Guidance From Computational Models Of Memory

Mohammad Khajah (mohammad.khajah@colorado.edu)
Robert Lindsey (robert.lindsey@colorado.edu)
Michael Mozer (mozer@colorado.edu)
Department of Computer Science
University of Colorado
Boulder, CO 80309-0430

Abstract

During the school semester, students face an onslaught of new material. Students work hard to achieve initial mastery of the material, but soon their skill degrades or they forget. Although students and educators both appreciate that review can help stabilize learning, time constraints result in a trade off between acquiring new knowledge and preserving old knowledge. To use time efficiently, when should review take place? Experimental studies have shown benefits to long-term retention with spaced study, but little practical advice is available to students and educators about the optimal spacing of study. The dearth of advice is due to the challenge of conducting experimental studies of learning in educational settings where material is introduced in blocks over the time frame of a semester. In this paper, we turn to two established models of memory—ACT-R and MCM—to conduct simulation studies exploring the impact of study schedule on long-term retention. Based on the premise of fixed time each week to review, converging evidence from the two models suggests that an optimal review schedule obtains significant benefits over haphazard (suboptimal) review schedules. Further, we identify two scheduling heuristics that obtain near optimal review performance: (1) review the material from μ-weeks back, and (2) review material whose predicted memory strength is closest to θ. The former has implications for classroom instruction and the latter for the design of electronic tutors.

Keywords: spacing effect; memory model; ACT-R, MCM, optimization, learning, review

Introduction

At every level of the educational system, from grade school through college through professional school, instructors and textbooks typically introduce students to new material in blocks. These blocks—sometimes called sections or units—represent conceptually coherent chunks of knowledge. For example, in a foreign language class, students may learn conversational skills concerning foods and restaurants one week, traveling the next week, and vacation activities the following week. In medical school, students may study vascular, pulmonary, and renal systems in consecutive months.

At the end of each block, teachers typically administer a quiz or assign a problem set to encourage students to master the material in the block. Because the students are rewarded for focusing on this task, they have little incentive at that moment to rehearse and practice material they have learned previously. As a result, forgetting is inevitable. Although anyone who has taught a class appreciates the need for review, the time demands of review of old material must be balanced against the need to introduce new material, explain concepts, and encourage students toward initial mastery.

Achieving this balance requires an understanding of when students will most benefit from review. Reviewing material when it is fresh provides minimal benefit; however, waiting until material has been forgotten is also costly because the earlier study provides little benefit. A long history of research in experimental psychology has shown that the temporal distribution or *spacing* of study has a substantive impact on long-term retention. Selecting the ideal spacing of study can lead to nearly doubling retention of material on an educationally relevant time scale of a year (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008). Evidence for the benefit of spaced study is found not only in the domain of declarative learning, but in conceptual understanding and cognitive skill acquisition (Carpenter et al., in press), and spacing manipulations have been shown to be effective in the classroom (e.g., Sobel, Cepeda, & Kapler, 2011).

The goal of this paper is to leverage computer simulations to offer educators practical guidance about the optimal spacing of review in the context of a semester- or quarter-long class. In such a context, we assume that the class is divided into blocks, new material is introduced in each block, and some time during each block is allotted for review of old material. The issue at hand is what material should be reviewed and when. To state the issue formally, suppose that a semester consists of B blocks, and in block b, $b = 1...B$, the opportunity exists to review material from N previous blocks, denoted $R_{b,n}$, $1 \leq R_{b,n} < b$ and $n = 1...N$. What review schedule, $\mathbf{R} \equiv \{R_{b,n}\}$, will maximize the students' memory for material following some retention interval RI weeks after the end of the semester?

Conducting controlled experimental studies to answer this question is not feasible. Even if the opportunity is afforded for the review of only *one* ($N = 1$) block, the number of review schedules is $1 \times 2 \times ... \times (B-1) = (B-1)!$, and the combinatorics get worse for larger N. A typical high-school semester or a typical college quarter may have $B = 10$ weeks of new material, for which $9! = 362880$ possible review schedules exist.

Although the number of candidate schedules could be greatly pruned, it would be a significant undertaking to conduct an experimental study comparing even *two* alternative schedules over a time window spanning ten study blocks and a subsequent final evaluation.

Because of the difficulty in conducting multi-session studies over extended time periods, nearly all prior research on spacing has either focused on the case of two study sessions or spanned such a compressed time scale that its educational relevance is questionable. (Kang, Lindsey, Mozer, & Pashler, submitted, offer a contrasting example.) Without recourse to controlled laboratory studies, one might conclude that cognitive science has little to offer educators beyond the qualitative advice to review material occasionally.

However, a trustworthy computational model can be used to optimize study, i.e., to search for the study schedule that will maximize student retention at some specified point or time window in the future. The cost of predicting performance with a computational model under a given study schedule is negligible relative to the cost of conducting a behavioral experiment. In past work, we've shown the potential benefits of optimizing study via a cognitive model (Lindsey, Mozer, Cepeda, & Pashler, 2009). In the present work, we use models to explore a range of scheduling algorithms in order to identify both optimal schedulers and heuristic schedulers that well approximate the optimum in an extended classroom setting.

Spaced Study And Memory Models

The spacing effect has been investigated for over a hundred years (Ebbinghaus, 1885/1964), and in additional to qualitative theories, many mathematical and computational models have been proposed to explain the phenomenon (e.g., Benjamin & Tullis, 2010; Raaijmakers, 2003). However, two recent efforts have been fairly comprehensive in obtaining quantitative fits to data and both have shown promise in predicting the outcome of experimental studies: an extension of the ACT-R model of memory (Pavlik & Anderson, 2005, 2008), and a model we developed called the *Multiscale Context Model* or *MCM* (Mozer, Pashler, Cepeda, Lindsey, & Vul, 2009). We summarize the two models and then turn to using the models as a proxy for human performance to predict the optimal spacing of study. Lindsey et al. (2009) compared qualitative predictions of ACT-R and MCM in hypothetical situations, and the models gave some contrasting results. However, these earlier simulation studies did not explore the predictions of the models in a practical educationally relevant setting.

ACT-R

ACT-R (Anderson et al., 2004) is an influential cognitive architecture whose declarative memory module is often used to account for explicit recall following study. ACT-R assumes that a separate trace is laid down each time an item is studied, and the trace decays according to a power law, t^{-d}, where t is the age of the memory and d is the power law decay for that trace. Following n study episodes, the activation for an item, m_n, combines the trace strengths of individual study episodes:

$$m_n = \ln\left(\sum_{k=1}^{n} b_k t_k^{-d_k}\right) + \beta,$$

where t_k and d_k refer to the age and decay associated with trace k, and β is a student- and/or item-specific parameter that influences memory strength. The variable b_k reflects the salience of the kth study session (Pavlik, 2007): larger values of b_k correspond to cases where, for example, the participant self-tested and therefore exerted more effort.

To explain spacing effects, Pavlik and Anderson (2005; 2008) made an additional assumption: the decay for the trace formed on study trial k depends on the item's activation at the point when study occurs:

$$d_k(m_{k-1}) = ce^{m_{k-1}} + \alpha,$$

where c and α are constants. If study trial k occurs shortly after the previous trial, the item's activation, m_{k-1}, is large, which will cause trace k to decay rapidly. Increasing spacing therefore benefits memory by slowing decay of trace k. However, this benefit is traded off against a cost incurred due to the aging of traces $1...k-1$ that causes them to decay further. The probability of recall is monotonically related to activation:

$$p(m) = 1/(1 + e^{\frac{\tau - m}{s}}),$$

where τ and s are additional parameters. In total, the variant of the model described here has six free parameters.

Pavlik and Anderson (2008) use ACT-R activation predictions in a heuristic algorithm for *within*-session scheduling of trial order and trial type (i.e., whether an item is merely studied, or whether it is first tested and then studied). They assume a fixed spacing between initial study and subsequent review. Thus, their algorithm reduces to determining how to best allocate a finite amount of time within a session. Although they show an effect of the algorithm used for within-session scheduling, we focus on the complementary issue of between-session scheduling. The between-session manipulation has a far greater impact on long-term retention (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006).

MCM

ACT-R is posited on the assumption that memory decay follows a power function. We developed an alternative model, the Multiscale Context Model or MCM (Mozer et al., 2009), which provides a mechanistic basis

for the power function. Adopting key ideas from previous models of the spacing effect (Kording, Tenenbaum, & Shadmehr, 2007; Raaijmakers, 2003; Staddon, Chelaru, & Higa, 2002) MCM proposes that each time an item is studied, it is stored in multiple item-specific memory traces that decay at different rates. Although each trace has an exponential decay, the sum of the traces decays approximately as a power function of time. Specifically, trace i, denoted x_i, decays over time according to:

$$x_i(t + \Delta t) = x_i(t) \exp(-\Delta t/\tau_i),$$

where τ_i is the decay time constant, ordered such that successive traces have slower decays, i.e., $\tau_i < \tau_{i+1}$. Traces $1 - k$ are combined to form a net trace strength, s_k, via a weighted average:

$$s_k = \frac{1}{\Gamma_k} \sum_{i=1}^{k} \gamma_i x_i, \text{ where } \Gamma_k = \sum_{i=1}^{k} \gamma_i$$

and γ_i is a factor representing the contribution of trace i. In a cascade of K traces, recall probability is simply the thresholded strength: $P(\text{recall}) = \min(1, s_K)$.

Spacing effects arise from the trace update rule, which is based on Staddon et al. (2002). A trace is updated only to the degree that it and faster decaying traces fail to encode the item at the time of study. This rule has the effect of storing information on a time scale that is appropriate given its frequency of occurrence in the environment. Formally, when an item is studied, the increment to trace i is negatively correlated with the net strength of the first i traces, i.e.,

$$\Delta x_i = \epsilon(1 - s_i),$$

where ϵ is a step size. We adopt the retrieval-dependent update assumption of Raaijmakers (2003): $\epsilon = 1$ for an item that is not recalled at the time of study, and $\epsilon = \epsilon_r$ ($\epsilon_r > 1$) for an item that is recalled.

The model has only 5 free parameters (ϵ_r, and 4 parameters that determine the contributions $\{\gamma_i\}$ and the time constants, $\{\tau_i\}$). MCM was designed such that its parameters could be fully constrained by data that are easy to collect—the function characterizing forgetting folllowing a single study session—which then allows the model to make predictions for data that are difficult to collect—the function characterizing forgetting following a study schedule consisting of two or more study sessions. MCM has been used to obtain parameter-free predictions for various results in the spacing literature.

Methodology

Model Parameterization

Different parameterizations of ACT-R and MCM are critical to accounting for a range of *learning scenarios*—scenarios that encode the ability and background knowledge of students, the difficulty of material, the manner of study, and the degree to which previously learned material interferes with or facilitates the learning of new material. Because our goal is to obtain results that have some generality across scenarios, we simulate a wide range of scenarios and base our results on the average over scenarios. We summarize the many factors that comprise a scenario in terms of a *forgetting curve*, which specifies the probability that material learned in a single study session will be available at some later point in time. Figure 1 shows a family of 105 forgetting curves, all of which decay according to a power function of time. This family expresses a diverse range of naturally occurring degrees of forgetting.

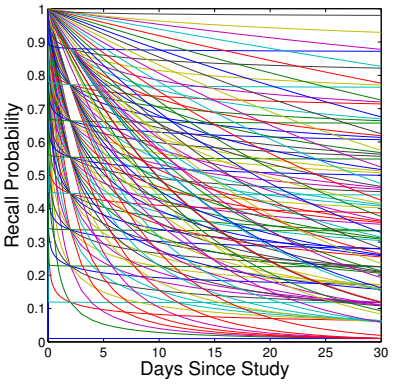

Figure 1: The 105 power-function forgetting curves used to represent a diversity of learning scenarios (i.e., learning tasks varying in material difficulty, student ability, manner of study, and potential interference.

For MCM, we search for model parameters that well approximate each forgetting curve. MCM has five free parameters, one of which (ϵ_r) was set based on previous simulations, and the other four of which directly determine and are fully constrained by the shape of the forgetting curve. For ACT-R, we fixed $b_k = 1$, but because its remaining free parameters are not fully constrained by the forgetting curve, we used the parameterized MCM to generate data which was then used to fit ACT-R parameters, ensuring that matched parameter sets had a loose correspondence. The generated data consisted of two study sessions with intersession intervals ranging from minutes to weeks, and a subsequent final test days to months later. This procedure yielded 105 matched instantiations of MCM and ACT-R, reflecting a wide range of scenarios.

Simulated Learning Experiment

We conducted separate simulations of MCM and ACT-R to model the performance of a student learning new material in each of $B = 10$ weekly blocks. We assumed homogeneity of material in a block, allowing the block's material to be distilled into a single item for the purpose of the simulation. Initial study was simulated as a single training trial to the model, though this training trial—and the corresponding memory trace—is intended to correspond to the net effect of concentrated study over multiple trials by a student learner.

Review was included in the curriculum starting after a

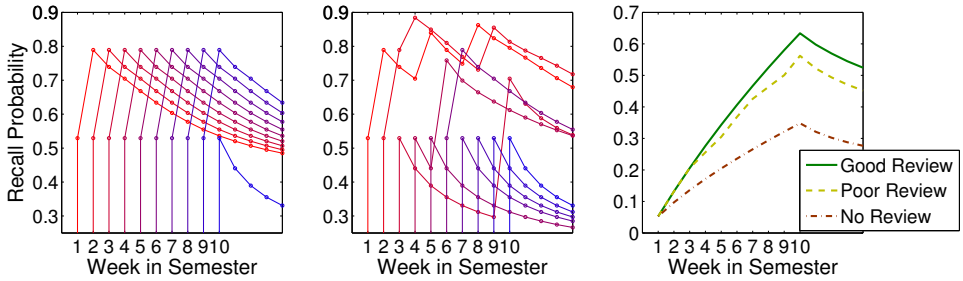

Figure 2: (left, middle panels) Activation trace from MCM for 10 blocks of material for good and poor review schedules. (right panel) Predicted performance on cumulative exam as a function of week in semester for alternative review schedules.

D-week delay. Review consists of selecting *one* previous block's material and presenting it as a training trial to the model. We simulated the $(B-1)!/(D-1)!$ distinct review schedules. We allowed D to vary because when review begins earlier in the semester, the number of sensible review schedules significantly shrinks. For example, with $D = 1$, the only option for week 2 review is week 1; this selection has consequences the next week because in week 3, review of week 1 again adds little benefit, so a sensible option is to review week 2; and so forth.

To evaluate the effectiveness of a review schedule, mean recall accuracy over the B blocks was assessed by querying the model with a final recall test following a retention interval of RI weeks past the end of the semester.

Alternative Review Schedulers

To summarize, we consider two models of human learning (ACT-R and MCM), 105 scenarios (model parameterizations), 3 retention intervals ($RI = 1, 4, 26$ weeks), and 3 review delays ($D = 1, 2, 3$ weeks), for a total of 1890 distinct combinations. For each combination, we conducted an exhaustive search through the set of distinct review schedules to determine the *optimal schedule*—the schedule that yields the highest average accuracy on the final test according to the model.

In addition, we considered various *heuristic schedulers*. Our goal is to identify heuristics that produce a close-to-optimal schedule. The two best heuristic schedulers were as follows. A *μ-back scheduler* follows a simple rule: in week i, review material from week $\max(1, i-\mu)$. A *θ-threshold scheduler* is motivated by Bjork's (1994) notion of *desireable difficulty*—that material should be restudied as it is on the verge of being forgotten. Using a memory model to determine the strength of each week's material, this scheduler selects the material whose recall probability closest to θ. Because we use the same model for scheduling as we use for modeling the student, this scheduler offers a best-case use of the θ-threshold. (We also explored several variants of the threshold scheduler which yielded poorer performance. One variant uses a scaled threshold rule whereby the threshold value is relative to the range of performance over all weeks' material. Another uses an asymmetric threshold where the selection is for material whose recall probability is close to the threshold on one side—either above or below.)

Results

Figure 2 provides an intuition about the operation of our model-based scheduling. The left panel of the Figure shows ten curves, each representing the memory strength predicted by MCM for one block of material as a function of weeks into the semester. The color coding from red to blue indicates blocks 1-10, respectively. In this example, block i is introduced in week i and is then reviewed in week $i+1$. As a result, the block gets a bump in strength in weeks i and $i+1$, and then decays from that point on. The curves in the Figure represent the average over the 105 learning scenarios, and the ordinate of the graph shows the expected recall probability over these scenarios. The absolute probability is immaterial and is a consequence of the specific scenarios we simulate. However, relative probabilities matter. To emphasize this point, the middle panel of the Figure shows an activation trace for an arbitrary and somewhat bad review schedule. The right panel shows the same time history of activation, but averaged over the individual blocks to obtain a prediction of cumulative-exam performance (weighting all blocks equally) at a given time. The superiority of the one-back schedule (left panel) over the arbitrary (middle panel) is reflected in a higher average recall probability. Four weeks following the end of the 10-week semester, the better review schedule achieves a 89.7% improvement in retention over no review, and a 16.1% improvement in retention over the poorer quality review schedule.

Exhaustive Search Of Alternative Schedules

Figure 3 shows a set of curves that reflect the expected perfomance of all possible review schedules for a given simulation, sorted from worst to best. The average is taken over learning scenarios. Each graph shows three simulations, one per retention interval ($RI = 1, 4, 26$ weeks). The top and bottom rows are simulations of MCM and ACT-R, respectively. The columns from left-to-right correspond to simulations in which review begins following weeks 1, 2, and 3 ($D = 1, 2, 3$). The colored squares on the left of each graph indicate the performance of a 'no review' condition for the retention interval of the corresponding color. Not surprisingly, all review schedules are superior to no review, and well-timed review is as much as 33% better than poorly-timed review.

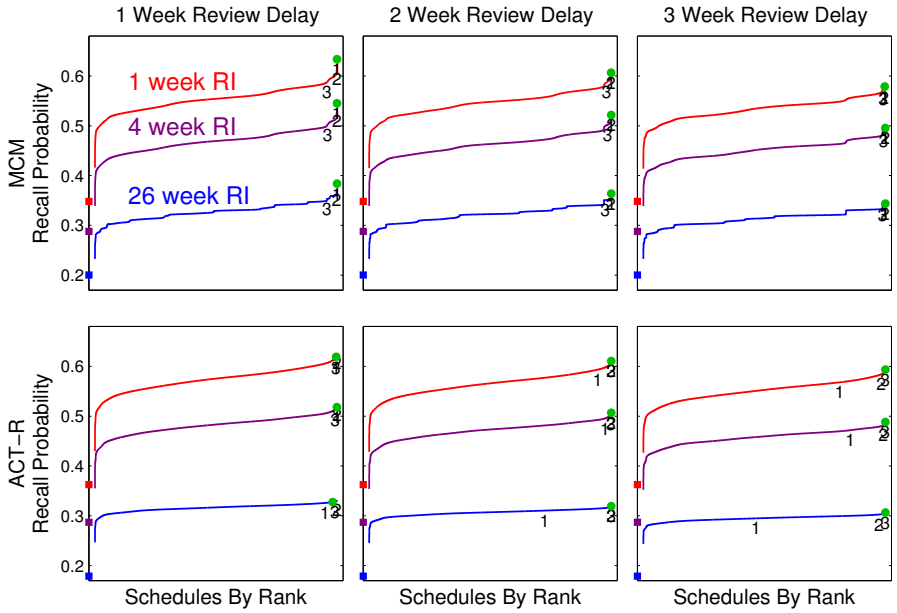

Figure 3: Each curve is the sorted performance of all possible review schedules (The x-axis is an index over distinct review schedules). Each curve corresponds to one retention interval (RI). Top and bottom rows are simulations of MCM and ACT-R, respectively. The columns correspond to simulations in which review onset begins following weeks 1, 2, and 3, respectively. Colored squares indicate the performance of a 'no review' schedule. Digits indicate the performance of the heuristic μ-back scheduler, for $\mu = 1, 2, 3$. The green disk (.) indicates the performance of the θ-threshold scheduler for optimal θ.

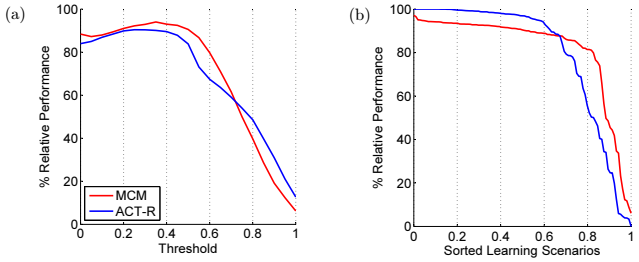

Figure 4: (a) Relative performance predicted by MCM and ACT-R for the θ-threshold heuristic as a function of θ (for $D = 1$, $RI = 1$). (b) Relative performance of the 2-back schedule over all learning scenarios, sorted from best to worst (for $D = 1$, $RI = 1$). In both graphs, performance is relative to the exhaustive space of schedules.

θ-Threshold Heuristic Scheduler

The key result in Figure 3 concerns the performance of heuristic schedulers relative to the optimum schedule discovered by exhaustive search. For each curve, the location of the green disk indicates the relative ranking of the θ-threshold schedule, for the best setting of θ. The further to the right along the x-axis, the higher the ranking. The two models are consistent in predicting that the θ-threshold scheduler is as good or nearly as good as the best schedule found by exhaustive search. Figure 4a shows how the predicted performance varies as a function of θ for the two models, for a delay of $D = 1$ week and a retention interval of $RI = 1$ week. The ordinate indicates the relative performance in the range defined by the complete space of schedules, where 100% and 0% correspond to the best and worst schedules found by exhaustive search, respectively. Notably, the two models yield very similar curves, and although the θ-threshold scheduler does not produce the very best schedule, it comes reasonably close. Notably, both MCM and ACT-R are consistent in indicating that a threshold in the neighborhood of $\theta = .4$ is best. We have shown the curve for $D = 1$ and $RI = 1$, but curves for the other values of D and RI are quite similar, and all have the same optimum for θ.

The limitation of a threshold scheduler is that it requires an accurate model to predict memory strength as a function of time given some history of study. In our simulation, we've assumed that the model we use for determining memory strength—either MCM or ACT-R—is a veridical model of our (simulated) student. An important question for future research concerns how the accuracy of the model used for scheduling affects the performance of the θ-threshold scheduler. However, it is clear that whatever model is used must take into account the history and spacing of past study, because the effect of distributed practice—as embodied in both MCM and ACT-R—is central to the difference in performance across review schedules.

μ-Back Heuristic Scheduler

Figure 3 also depicts the performance of the 1-, 2-, and 3-back schedules, all of which do reasonably well across models, delays, and retention intervals. However, because ACT-R predicts the 1-back schedule to be inferior for $D = 2, 3$, and because MCM predicts the 3-back schedule to be slightly worse for $D = 1, 2$, we suggest that the $\mu = 2$, or the 2-back schedule, might be adopted as a robust solution across conditions.

All results we've presented to this point are the average over the 105 learning scenarios. It's possible that

the μ-back schedules work well on average but not for specific scenarios. To examine the performance of the 2-back schedule across scenarios, Figure 4b shows the performance in each scenario, sorted from best to worst. The curves for MCM and ACT-R are remarkably similar, and indicate that the 2-back schedule performs well for the majority (60-80%) of scenarios we considered, further supporting our claim of its robustness.

Discussion

In a metaanalysis of the experimental literature, the optimal spacing of study was found to grow monotonically with the retention interval (Cepeda et al., 2006). Although in past work we've shown that MCM and ACT-R both predict this characteristic, neither model strongly predicts that the best μ in the μ-back scheduler should increase with the retention interval (Figure 3). Most likely, this inconsistency is due to the fact that as μ increases, the initial $\mu + 1$ weeks of study become focused on the *first* week's material, and there are diminishing returns of this focus. Consequently, the benefits of increased spacing must be outweighed by the cost of ill-spent review time. This result suggests to us the importance of moving beyond laboratory studies of spacing—typically with two study sessions and a single block of material to be learned—to situations more reflective of real-world educational constraints, i.e., semesters in which multiple blocks of material are presented staggered in time and initial study must be interlaced with review.

Our results provide practical guidance to educators: To preserve learning beyond the end of a semester, a 2-back review schedule should generally be appropriate. Although classroom teachers do not have access to mathematical models of human memory, and therefore cannot exploit the θ-threshold scheduler, we see great potential of incorporating model-based scheduling into electronic tutors used in synchronization with classroom instruction (Lindsey et al., in preparation). Indeed, such an approach opens the possibility to *personalized* review appropriate for a specific student rather than a one-size-fits-all approach. Our caveat in suggesting this approach is that it requires accurate psychological models of memory. Models based on intuition—as embodied in existing web-based flashcard apps—are unlikely to be adequate.

Acknowledgments

This research was supported by the UCSD Temporal Dynamics of Learning Center (NSF award SBE-0542013, Garrison Cottrell, PI), and by NSF award BCS-0720375.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of mind. *Psych. Rev.*, *111*, 1036–1060.

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cog. Psych.*, *61*, 228–247.

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

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H., & Pashler, H. (in press). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Ed. Res.*

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psych. Bull.*, *132*, 354–380.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psych. Sci.*, *19*, 1095–1102.

Ebbinghaus, H. (1885/1964). *Memory: A contribution to experimental psychology.* New York: Dover.

Kang, S. H. K., Lindsey, R. V., Mozer, M. C., & Pashler, H. (Submitted). Retrieval practice over the long term: Expanding or equal-interval spacing?

Kording, K. P., Tenenbaum, J. B., & Shadmehr, R. (2007). The dynamics of memory are a consequence of optimal adaptation to a changing body. *Nature Neuroscience*, *10*, 779-786.

Lindsey, R., Mozer, M. C., Cepeda, N. J., & Pashler, H. (2009). Optimizing memory retention with cognitive models. In A. Howes, D. Peebles, & R. Cooper (Eds.), *Proc. 9th intl. conf. on cog. modeling (ICCM).* Manchester, UK.

Mozer, M. C., Pashler, H., Cepeda, N., Lindsey, R., & Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. In *Advances in neural information processing systems 22* (pp. 1321–1329). La Jolla, CA: NIPS Foundation.

Pavlik, P. I. (2007). Understanding and applying the dynamics of test practice and study practice. *Instruc. Sci.*, *35*, 407–441.

Pavlik, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, *29*(4), 559-586.

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *J. Exptl. Psych.: Applied*, *14*, 101–117.

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: application of the SAM model. *Cognitive Science*, *27*, 431–452.

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Appl. Cog. Psych.*, *25*, 763-767.

Staddon, J. E. R., Chelaru, I. M., & Higa, J. J. (2002). Habituation, memory and the brain: The dynamics of interval timing. *Behav. Proc.*, *57*, 71-88.