Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. Psychological Science, 25, 639-647. doi: 10.1177/0956797613504302.

Improving students' long-term knowledge retention through personalized review

Robert V. Lindsey*
Jeff D. Shroyer*
Harold Pashler+
Michael C. Mozer*

*Institute of Cognitive Science and
Department of Computer Science
University of Colorado, Boulder

+Department of Psychology
University of California, San Diego

August 16, 2013

Corresponding Author:
Michael C. Mozer
Institute of Cognitive Science
University of Colorado
Boulder, CO 80309-0430
mozer@colorado.edu
(303) 517-2777

Keywords: long-term memory, declarative memory, spacing effect, adaptive scheduling, classroom education, Bayesian modeling

Abstract

Human memory is imperfect; thus, periodic review is required for the long-term preservation of knowledge and skills. However, students at every educational level are challenged by an evergrowing amount of material to review and an ongoing imperative to master new material. We developed a method for efficient, systematic, personalized review that combines statistical techniques for inferring individual differences with a psychological theory of memory. The method was integrated into a semester-long middle school foreign language course via retrieval-practice software. In a cumulative exam administered after the semester's end that compared time-matched review strategies, personalized review yielded a 16.5% boost in course retention over current educational practice (massed study) and a 10.0% improvement over a one-size-fits-all strategy for spaced study.

Forgetting is ubiquitous. Regardless of the nature of the skills or material being taught, regardless of the age or background of the learner, forgetting happens. Teachers rightfully focus their efforts on helping students acquire new knowledge and skills, but newly acquired information is vulnerable and easily slips away. Even highly motivated learners are not immune: medical students forget roughly 25–35% of basic science knowledge after one year, more than 50% by the next year (Custers, 2010), and 80–85% after 25 years (Custers & ten Cate, 2011).

Forgetting is influenced by the temporal distribution of study. For over a century, psychologists have noted that temporally spaced practice leads to more robust and durable learning than massed practice (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). Although spaced practice is beneficial in many tasks beyond rote memorization (Kerfoot et al., 2010) and shows promise in improving educational outcomes (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013), the reward structure of academic programs seldom provides an incentive to methodically revisit previously learned material. Teachers commonly introduce material in sections and evaluate students at the completion of each section; consequently, students' grades are well served by focusing study exclusively on the current section. Although optimal in terms of students' short-term goals, this strategy is costly for the long-term goal of maintaining accessibility of knowledge and skills. Other obstacles stand in the way of incorporating distributed practice into the curriculum. Students who are in principle willing to commit time to review can be overwhelmed by the amount of material, and their metacognitive judgments about what they should study may be unreliable (Nelson & Dunlosky, 1991). Moreover, though teachers recognize the need for review, the time demands of restudying old material compete against the imperative to regularly introduce new material.

We incorporated systematic, temporally distributed review into third-semester Spanish foreign language instruction using a web-based flaschard tutoring system, the *Colorado Optimized Language Tutor* or COLT. Throughout the semester, 179 students used COLT to drill on ten chapters of material. COLT presented vocabulary words and short sentences in English and required students to type the Spanish translation, after which corrective feedback was provided. The software was used both to practice newly introduced material and to review previously studied material.

For each chapter of course material, students engaged in three 20–30 minute sessions with COLT during class time. The first two sessions began with a study-to-proficiency phase for the current chapter and then proceeded to a review phase. On the third session, these activities were preceded by a quiz on the current chapter, which counted toward the course grade. During the review phase, study items from all chapters covered so far in the course were eligible for presentation. Selection of items was handled by three different

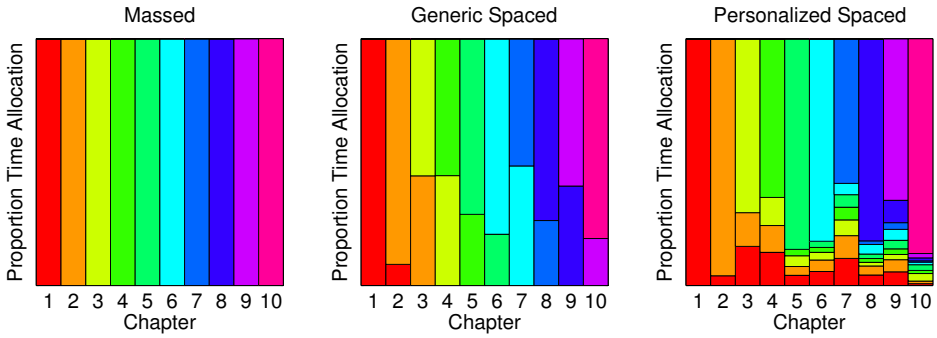

Figure 1: Time allocation of the three review schedulers. Course material was introduced one chapter at a time, generally at one-week intervals. Each vertical slice indicates the proportion of time spent in a week studying each of the chapters introduced so far. Each chapter is indicated by a unique color.

schedulers.

A *massed* scheduler continued to select material from the current chapter. It presented the item in the current chapter that students had least recently studied. This scheduler corresponds to recent educational practice: prior to the introduction of COLT, alternative software was used that allowed students to select the chapter they wished to study. Not surprisingly, given a choice, students focused their effort on preparing for the imminent end-of-chapter quiz, consistent with the preference for massed study found by Cohen, Yan, Halamish, and Bjork (2013).

A *generic-spaced* scheduler selected one previous chapter to review at a spacing deemed to be optimal for a range of students and a variety of material according to both empirical studies (Cepeda et al., 2006; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008) and computational models (Khajah, Lindsey, & Mozer, 2013; Mozer, Pashler, Cepeda, Lindsey, & Vul, 2009). On the time frame of a semester—where material must be retained for 1-3 months—a one-week lag between initial study and review obtains near-peak performance for a range of declarative materials. To achieve this lag, the generic-spaced scheduler selected review items from the previous chapter, giving priority to the least recently studied (Figure 1).

A *personalized-spaced* scheduler used a latent-state Bayesian model to predict what specific material a particular student would most benefit from reviewing. This model infers the instantaneous memory strength of each item the student has studied. The inference problem is difficult because past observations of a particular student studying a particular item provide only a weak source of evidence concerning memory strength. To illustrate, suppose that the student had practiced an item twice, having failed to translate it once 15 days ago but having succeeded 9 days ago. Based on these sparse observations, it would seem that one cannot reliably predict the student's current ability to translate the item. However, data from

Table 1: Presentation statistics of individual student-items over entire experiment

		Massed	Generic	Personalized
# study-to-criterion trials	mean	7.58	7.57	7.56
	std. dev.	6.70	6.49	6.47
# review trials	mean	8.03	8.05	8.03
	std. dev.	11.99	12.14	9.65
# days between review trials	mean	0.12	1.69	4.70
	std. dev.	1.43	3.29	6.39

the population of students studying the population of items over time can provide constraints helpful in characterizing the performance of a specific student for a specific item at a given moment. Our model-based approach is related to that used by e-commerce sites that leverage their entire database of past purchases to make individualized recommendations, even when customers have sparse purchase histories.

Our model defines memory strength as being jointly dependent on factors relating to (1) an item's latent difficulty, (2) a student's latent ability, and (3) the amount, timing, and outcome of past study. We refer to the model with the acronym DASH summarizing the three factors (difficulty, ability, and study history). By incorporating psychological theories of memory into a data-driven modeling approach, DASH characterizes both individual differences and the temporal dynamics of learning and forgetting. The Appendix describes DASH in detail.

The scheduler was varied within participant by randomly assigning one third of a chapter's items to each scheduler, counterbalanced across participants. During review, the schedulers alternated in selecting items for retrieval practice. Each selected from among the items assigned to it, ensuring that all items had equal opportunity and that all schedulers administered an equal number of review trials. Figure 1 and Table 1 present student-item statistics for each scheduler over the time course of the experiment.

Results

Two proctored cumulative exams were administered to assess retention, one at the semester's end and one 28 days later, at the beginning of the following semester. Each exam tested half of the course material, randomized for each student and balanced across chapters and schedulers; no corrective feedback was provided. On the first exam, the personalized spaced scheduler improved retention by 12.4% over the massed scheduler ($t(169) = 10.1$, $p < .0001$, Cohen's $d = 1.38$) and by 8.3% over the generic spaced scheduler ($t(169) = 8.2$, $p < .0001$, $d = 1.05$) (Figure 2a). Over the 28-day intersemester break, the forgetting rate was 18.1%, 17.1%, and 15.7% for the massed, generic, and personalized conditions, respectively, leading to

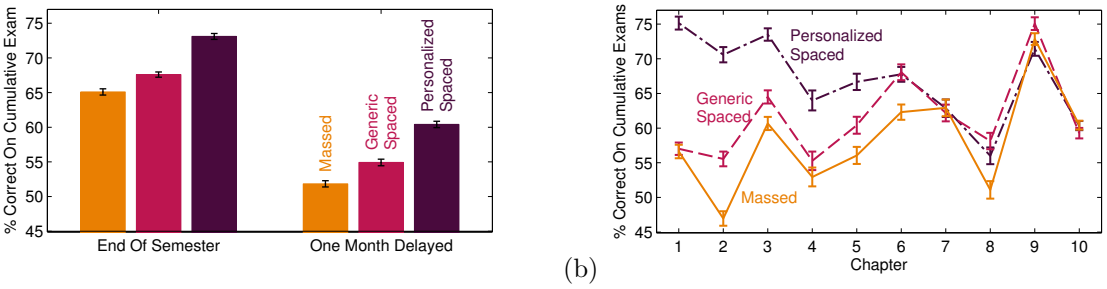

Figure 2: (a) Mean scores on the two cumulative end-of-semester exams, taken 28 days apart. (b) Mean score of the two exams as a function of the chapter in which the material was introduced. The personalized-spaced scheduler produced a large benefit for early chapters in the semester without sacrificing efficacy on later chapters. All error bars indicate ±1 within-student standard error (Masson & Loftus, 2003).

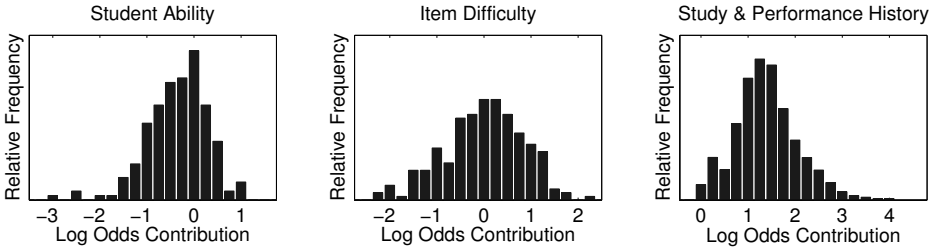

Figure 3: Histogram of three sets of inferred factors, expressed in their additive contribution to predicted log-odds of recall. Each factor varies over three log units, corresponding to a possible modulation of recall probability by 0.65.

an even larger advantage for personalized review. On the second exam, personalized review boosted retention by 16.5% over massed review ($t(175) = 11.1$, $p < .0001$, $d = 1.42$) and by 10.0% over generic review ($t(175) = 6.59$, $p < .0001$, $d = 0.88$). The primary impact of the schedulers was for material introduced earlier in the semester (Figure 2b), which is sensible because that material had the most opportunity for being manipulated via review. Among students who took both exams, only 22.3% and 13.5% scored better in the generic and massed conditions than in the personalized, respectively.

Note that "massed" review is spaced by usual laboratory standards, being spread out over at least seven days. This fact may explain both the small benefit of generic spaced over massed and the absence of a spacing effect for the final chapters.

DASH determines the contribution of a student's ability, an item's difficulty, and a student-item's specific study history to recall success. Histograms of these inferred contributions show substantial variability (Figure 3), yielding decisions about what to review that were markedly different across individual students and items.

DASH predicts a student's response accuracy to an item at a point in time given the response history

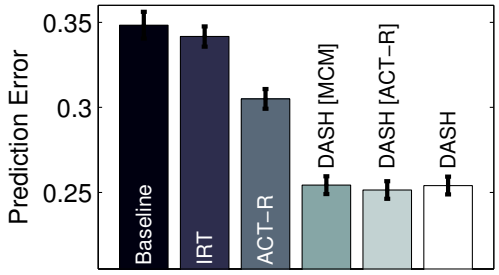

Figure 4: Accumulative prediction error of DASH and five alternative models using the data from the semester-long experiment. Error bars indicate ±1 standard error of the mean.

of all students and items to that point. To evaluate the quality of DASH's predictions, we compared DASH against alternative models by dividing the 597,990 retrieval practice trials recorded over the semester into 100 temporally contiguous disjoint sets, and the data for each set was predicted given the preceding sets. The *accumulative prediction error* (Wagenmakers, Grünwald, & Steyvers, 2006) was computed using the mean deviation between the model's predicted recall probability and the actual binary outcome, normalized such that each student is weighted equally. Figure 4 compares DASH against five alternatives: a *baseline* model that predicts a student's future performance to be the proportion of correct responses the student has made in the past, a Bayesian form of *item-response theory* (IRT) (De Boeck & Wilson, 2004), a model of spacing effects based on the memory component of ACT-R (Pavlik & Anderson, 2005), and two variants of DASH that incorporate alternative representations of study history motivated by models of spacing effects (ACT-R, MCM). Details of the alternatives and the evaluation are described in the Supplemental Online Material.

The three variants of DASH perform better than the alternatives. Each variant has two key components: (1) a dynamical representation of study history that can characterize learning and forgetting, and (2) a Bayesian approach to inferring latent difficulty and ability factors. Models that omit the first component (baseline and IRT) or the second (baseline and ACT-R) do not fare as well. The DASH variants all perform similarly. Because these variants differ only in the manner in which the temporal distribution of study and recall outcomes is represented, this distinction does not appear to be critical.

Discussion

Our work builds on the rich history of applied human-learning research by integrating two distinct threads: classroom-based studies that compare massed versus spaced presentation of material (Carpenter, Pashler, & Cepeda, 2009; Seabrook, Brown, & Solity, 2005; Sobel, Cepeda, & Kapler, 2011), and laboratory-based

investigations of techniques that select material for an individual to study based on that individual's past study history and performance, known as *adaptive scheduling* (e.g., Atkinson, 1972).

Previous explorations of temporally distributed study in real-world educational settings have targeted a relatively narrow body of course material that was chosen such that exposure to the material outside of the experimental context was unlikely. Further, these studies compared just a few spacing conditions and the spacing was the same for all participants and materials, like our generic-spaced condition.

Previous evaluations of adaptive scheduling have demonstrated the advantage of one algorithm over another or over nonadaptive algorithms (Metzler-Baddeley & Baddeley, 2009; Pavlik & Anderson, 2008; van Rijn, van Maanen, & van Woudenberg, 2009), but these evaluations have been confined to the laboratory and have spanned a relatively short time scale. The most ambitious previous experiment (Pavlik & Anderson, 2008) involved three study sessions in one week and a test the following week. This compressed time scale limits the opportunity to manipulate spacing in a manner that would influence long-term retention (Cepeda et al., 2008). Further, brief laboratory studies do not deal with the complex issues that arise in a classroom, such as the staggered introduction of material and the certainty of exposure to the material outside of the experimental context.

Whereas previous studies offer in-principle evidence that human learning can be improved by the timing of review, our results demonstrate in practice that integrating personalized-review software into the classroom yields appreciable improvements in long-term educational outcomes. Our experiment goes beyond past efforts in its scope: it spans the time frame of a semester, covers the content of an entire course, and introduces material in a staggered fashion and in coordination with other course activities. We find it remarkable that the review manipulation had as large an effect as it did, considering that the duration of roughly 30 minutes a week was only about 10% of the time students were engaged with the course. The additional, uncontrolled exposure to material from classroom instruction, homework, and the textbook might well have washed out the effect of the experimental manipulation.

Personalization

Consistent with the adaptive-scheduling literature, our experiment shows that a one-size-fits-all variety of review is significantly less effective than personalized review. The traditional means of encouraging systematic review in classroom settings—cumulative exams and assignments—is therefore unlikely to be ideal.

We acknowledge that our design confounds personalization and the coarse temporal distribution of review (Figure 1, Table 1). However, the limited time for review and the evergrowing collection of material to review

would seem to demand deliberate selection.

Any form of personalization requires estimates of an individual's memory strength for specific knowledge. Previously proposed adaptive-scheduling algorithms base their estimates on observations from only that individual, whereas the approach taken here is fundamentally data driven, leveraging the large volume of quantitative data that can be collected in a digital learning environment to perform statistical inference on the knowledge states of individuals at an atomic level. This leverage is critical to obtaining accurate predictions (Figure 4).

Apart from the academic literature, two traditional adaptive-scheduling techniques have attracted a degree of popular interest: the Leitner (1972) system and SuperMemo (Wozniak & Gorzelanczyk, 1994). Both aim to review material when it is on the verge of being forgotten. As long as each retrieval attempt succeeds, both techniques yield a schedule in which the interpresentation interval expands with each successive presentation. These techniques underlie many flashcard-type web sites and mobile applications, which are marketed with the claim of optimizing retention. Though one might expect that any form of review would show some benefit, the claims have not yet undergone formal evaluation in actual usage, and based on our comparison of techniques for modeling memory strength, we suspect that there is room for improving these two traditional techniques.

Beyond fact learning

Our approach to personalization depends only on the notion that understanding and skill can be cast in terms of collections of primitive *knowledge components* or *KCs* (van Lehn, Jordan, & Litman, 2007) and that observed student behavior permits inferences about the state of these KCs. The approach is flexible, allowing for any problem posed to a student to depend on arbitrary combinations of KCs. The approach is also general, having application beyond declarative learning to domains focused on conceptual, procedural, and skill learning.

Educational failure at all levels often involves knowledge and skills that were once mastered but cease to be accessible due to lack of appropriately timed rehearsal. While it is common to pay lip service to the benefits of review, providing comprehensive and appropriately timed review is beyond what any teacher or student can reasonably arrange. Our results suggest that a digital tool which solves this problem in a practical, time-efficient manner will yield major payoffs for formal education at all levels.

Appendix: Modeling Students' Knowledge State

To personalize review, we must infer a student's *knowledge state*—the dynamically varying strength of each atomic component of knowledge (KC) as the student learns and forgets. Knowledge-state inference is a central concern in fields as diverse as educational assessment, intelligent tutoring systems, and long-term memory research. We describe two contrasting approaches taken in the literature, *data driven* and *theory driven*, and propose a synthesis used by our personalized-spaced scheduler.

A traditional psychometric approach to inferring student knowledge is item-response theory (IRT) (De Boeck & Wilson, 2004). Given a population of students answering a set of questions (e.g., SAT tests), IRT decomposes response accuracies into student- and question-specific parameters. The simplest form of IRT (Rasch, 1961) parameterizes the log-odds that a particular student will correctly answer a particular question through a student-specific ability factor α_s and a question-specific difficulty factor δ_i. Formally, the probability of recall success or failure R_{si} on question i by student s is given by

$$\Pr(R_{si} = 1 \mid \alpha_s, \delta_i) = \text{logistic}\left(\alpha_s - \delta_i\right),$$

where $\text{logistic}(z) = [1 + e^{-z}]^{-1}$.

IRT has been extended to incorporate additional factors into the prediction, including the amount of practice, the success of past practice, and the types of instructional intervention (Cen, Koedinger, & Junker, 2006, 2008; Pavlik, Cen, & Koedinger, 2009; Chi, Koedinger, Gordon, Jordan, & van Lehn, 2011). This class of models, known as *additive factors models*, has the form:

$$\Pr(R_{si} = 1 \mid \alpha_s, \delta_i, \boldsymbol{\gamma}, \boldsymbol{m}_{si}) = \text{logistic}\big(\alpha_s - \delta_i + \sum_j \gamma_j m_{sij}\big),$$

where j is an index over factors, γ_j is the skill level associated with factor j, and m_{sij} is the jth factor associated with student s and question i.

Although this class of model personalizes predictions based on student ability and experience, it does not consider the temporal distribution of practice. In contrast, psychological theories of long-term memory are designed to characterize the strength of stored information as a function of time. We focus on two recent models, MCM (Mozer et al., 2009) and a theory based on the ACT-R declarative memory module (Pavlik & Anderson, 2005). These models both assume that a distinct memory trace is laid down each time an item is studied, and this trace decays at a rate that depends on the temporal distribution of past study.

The psychological plausibility of MCM and ACT-R is demonstrated through fits of the models to behavioral data from laboratory studies of spaced review. Because minimizing the number of free parameters is key to a compelling account, cognitive models are typically fit to aggregate data—data from a population of students studying a body of material. They face a serious challenge in being useful for modeling the state of a particular KC for a particular student: a proliferation of parameters is needed to provide the flexibility to characterize different students and different types of material, but flexibility is an impediment to making strong predictions.

Our model, DASH, is a synthesis of data- and theory-driven approaches that inherits the strengths of each: the ability of data-driven approaches to exploit population data to make inferences about individuals, and the ability of theory-driven approaches to characterize the temporal dynamics of learning and forgetting based on study history and past performance. The synthesis begins with the data-driven additive factors model, and, through the choice of factors, embodies a theory of memory dynamics inspired by ACT-R and MCM. The factors are sensitive to the number of past study episodes and their outcomes. Motivated by the multiple traces of MCM, we include factors that span increasing windows of time, which allows the model to modulate its predictions based on the temporal distribution of study. Formally, DASH posits that

$$\Pr(R_{si} = 1 \mid \alpha_s, \delta_i, \boldsymbol{\phi}, \boldsymbol{\psi}) = \text{logistic}\Big[\alpha_s - \delta_i + \sum_w \phi_w \log(1 + c_{siw}) - \psi_w \log(1 + n_{siw})\Big], \tag{1}$$

where w is an index over time windows, c_{siw} is the number of times student s correctly recalled KC i in window w out of n_{siw} attempts, and ϕ_w and ψ_w are window-specific factor weights. The counts c_{siw} and n_{siw} are regularized by add-one smoothing, which ensures that the logarithm terms are finite.

We will explain the selection of time windows shortly, but we first provide an intuition for the specific form of the factors. The difference of factors inside the summation of Equation 1 determines a power law of practice. Odds of correct recall improve as a power function of: the number of correct trials with $\phi_w > 0$ and $\psi_w = 0$, the number of study trials with $\psi_w < 0$ and $\phi_w = 0$, and the proportion of correct trials with $\phi_w = \psi_w$. The power law of practice is a ubiquitous property of human learning incorporated into ACT-R. Our two-parameter formulation allows for a wide variety of power function relationships, from the three just mentioned to combinations thereof. The formulation builds a bias into DASH that additional study in a given time window helps, but has logarithmically diminishing returns. To validate the form of DASH in Equation 1, we fit a single-window model to data from the first week of our experiment, predicting performance on the end-of-chapter quiz for held-out data. We verified that Equation 1 outperformed variations of the formula

which omitted one term or the other or which expressed log-odds of recall directly in terms of the counts instead of the logarithmic form.

To model effects of temporally distributed study and forgetting, DASH includes multiple time windows. Window-specific parameters (ψ_w, ϕ_w) encode the dependence between recall at the present moment and the amount and outcome of study within the window. Motivated by theories of memory, we anchored all time windows at the present moment and varied their spans such that the temporal span of window w, denoted s_w, increased with w. We chose the distribution of spans such that there was finer temporal resolution for shorter spans, i.e., $s_{w+2} - s_{w+1} > s_{w+1} - s_w$. This distribution allows the model to efficiently represent rapid initial forgetting followed by a more gradual memory decay, which is a hallmark of the ACT-R power-function forgetting. This distribution is also motivated by the overlapping time scales of memory in MCM. ACT-R and MCM both suggest the elegant approach of exponentially expanding time windows, i.e., $s_w \propto \mathrm{e}^{\rho w}$. We roughly followed this suggestion, with three caveats. First, we did not try to encode the distribution of study on a very fine scale—less than an hour—because the fine-scale distribution is irrelevant for retention intervals on the order of months (Cepeda et al., 2008) and because the fine-scale distribution typically could not be exploited by DASH due to the cycle time of retraining. Second, we wished to limit the number of time scales so as to minimize the number of free parameters in the model to prevent overfitting and to allow for sensible generalization early in the semester when little data existed for long-term study. Third, we synchronized the time scales to the natural periodicities of student life. Taking these considerations into account, we chose five time scales: $\mathbf{s} = \{1/24, 1, 7, 30, \infty\}$. The Supplemental Online Material describes inference in the model.

Personalized Review Scheduling

DASH predicts the probability of successful recall for each student-KC pair. Although these predictions are required to schedule review optimally, optimal scheduling is computationally intractable because it requires planning over all possible futures. Consequently, COLT uses a heuristic policy for selecting review material, motivated by two distinct arguments, summarized here.

Using simulation studies, Khajah et al. (2013) examined policies that approximate the optimal policy found by exhaustive combinatorial search. To serve as a proxy for the student, they used a range of parameterizations of MCM and ACT-R. Their simulations were based on a set of assumptions approximately true for COLT, including a 10-week experiment in which new material is introduced each week, and a limited, fixed time allotted for review each week. With a few additional assumptions, exact optimization could be performed for a student who behaved according to a particular parameterization of either MCM or ACT-R.

Comparing long-term retention under alternative policies, the optimal policy obtained performance only slightly better than a simple heuristic policy that prioritizes for review the item whose expected recall probability is closest to a threshold θ, with the threshold $\theta = 0.33$ being best over a range of conditions. Note that with $\theta > 0$, DASH's student-ability parameter, α_s, influences the *relative* prioritization of items.

A threshold-based scheduler is also justified by Bjork's (1994) notion of *desirable difficulty*, which suggests that material should be restudied as it is on the verge of being forgotten. This qualitative prescription for study maps maps naturally into a threshold-based policy, assuming one has a model like DASH that can accurately estimate retrieval probability.

Acknowledgments

The research was supported by an NSF Graduate Research Fellowship, NSF grants SBE-0542013 and SMA-1041755, and a collaborative activity award from the McDonnell Foundation. We thank F. Craik, A. Glass, J.L. McClelland, H.L. Roediger III, and P. Wozniak for valuable feedback on the manuscript.

References

Atkinson, R. C. (1972). Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology*, *96*, 124–129.

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

Carpenter, S., Pashler, H., & Cepeda, N. (2009). Using tests to enhance 8th grade students' retention of U. S. history facts. *Applied Cognitive Psychology*, *23*, 760-771.

Cen, H., Koedinger, K., & Junker, B. (2006). Learning factors analysis—a general method for cognitive model evaluation and improvement. In *Proceedings of the Eighth International Conference on Intelligent Tutoring Systems.*

Cen, H., Koedinger, K., & Junker, B. (2008). Comparing two IRT models for conjunctive skills. In B. W. et al. (Ed.), *Proceedings of the Ninth International Conference on Intelligent Tutoring Systems.*

Cepeda, N., Pashler, H., Vul, E., Wixted, J., & Rohrer, D. (2006, May). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychonomic Bulletin & Review*, *132*(3), 354–380.

Cepeda, N., Vul, E., Rohrer, D., Wixted, J., & Pashler, H. (2008, Nov). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095-1102.

Chi, M., Koedinger, K., Gordon, G., Jordan, P., & van Lehn, K. (2011). Instructional factors analysis: A cognitive model for multiple instructional interventions. In C. Conati & S. Ventura (Eds.), *Proceedings of the Fourth International Conference on Educational Data Mining* (p. 61-70).

Cohen, M. S., Yan, V. X., Halamish, V., & Bjork, R. A. (2013). Do students think that difficult or valuable materials should be restudied sooner rather than later? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Advance online publication. doi:10.1037/a0032425.

Custers, E. (2010). Long-term retention of basic science knowledge: a review study. *Advances in Health Science Education: Theory & Practice*, *15*(1), 109-128.

Custers, E., & ten Cate, O. (2011). Very long-term retention of basic science knowledge in doctors after graduation. *Medical Education*, *45*(4), 422-430.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Dunlosky, J., Rawson, K., Marsh, E., Nathan, M., & Willingham, D. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4-58.

Kerfoot, B., Fu, Y., Baker, H., Connelly, D., Ritchey, M., & Genega, E. (2010, Sep). Online spaced education generates transfer and improves long-term retention of diagnostic skills: A randomized controlled trial. *Journal of the American College of Surgeons*, *211*(3), 331-337.

Khajah, M., Lindsey, R., & Mozer, M. (2013). Maximizing students' retention via spaced review: Practical guidance from computational models of memory. In *Proceedings of the Thirty-Fifth Annual Conference of the Cognitive Science Society.*

Leitner, S. (1972). So lernt man lernen. *Angewandte Lernpsychologie – ein Weg zum Erfolg*.

Masson, M., & Loftus, G. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, *57*, 203-220.

Metzler-Baddeley, C., & Baddeley, R. (2009). Does adaptive training work? *Applied Cognitive Psychology*, *23*, 254-266.

Mozer, M., Pashler, H., Cepeda, N., Lindsey, R., & Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems* (Vol. 22, p. 1321-1329).

Nelson, T., & Dunlosky, J. (1991). When people's judgments of learning (JOL) are extremely accurate at predicting subsequent recall: The delayed-JOL effect. *Psychological Science*, *2*, 267-270.

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

Pavlik, P., & Anderson, J. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*, 101-117.

Pavlik, P., Cen, H., & Koedinger, K. (2009). Performance factors analysis—a new alternative to knowledge tracing. In V. Dimitrova & R. Mizoguchi (Eds.), *Proceeding of the Fourteenth International Conference on Artificial Intelligence in Education.* Brighton, England.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proc. 4th berkeley*

symp on math. stat. & prob. (p. 321-333).

Seabrook, R., Brown, G., & Solity, J. (2005). Distributed and massed practice: from laboratory to classroom. *Applied Cognitive Psychology*, *19*, 107-122.

Sobel, H., Cepeda, N., & Kapler, I. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, *25*, 763-767.

van Lehn, K., Jordan, P., & Litman, D. (2007). Developing pedagogically effective tutorial dialogue tactics: Experiments and a testbed. In *Proceedings of the SLaTE Workshop on Speech and Language* (p. 17-20).

van Rijn, D. H., van Maanen, L., & van Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. In *Proceedings of the Ninth International Conference on Cognitive Modeling.*

Wagenmakers, E.-J., Grünwald, P., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, *50*, 149–166.

Wozniak, P., & Gorzelanczyk, E. (1994). Optimization of repetition spacing in the practice of learning. *Acta Neurobiologiae Experimentalis*, *54*, 59-62.

Supplementary Online Materials (SOM-R) Improving students' long-term knowledge retention through personalized review

Robert Lindsey, Jeff Shroyer, Harold Pashler, Michael Mozer

Materials

The instructor provided 409 Spanish-English words and phrases, covering 10 chapters of material. The material came from the textbook *¡Ven Conmigo! Adelante, Level 1a*, of which every student had a copy. Rather than treating minor variants of words and phrases as distinct and learned independently, we formed clusters of highly related words and phrases which were assumed to roughly form an equivalence class; i.e., any one is representative of the cluster. Included in the clustering were (1) all conjugations of a verb, whether regular or irregular; (2) masculine, feminine, and plural forms of a noun, e.g., *la prima* and *el primo* and *los primos* for cousin; and (3) thematic temporal relations, e.g., *el martes* and *los martes* for Wednesday (or on Wednesday) and on Wednesdays, respectively.

The 409 words and phrases were reassembled into 221 clusters. Following terminology of the intelligent tutoring community, we refer to a cluster as a *knowledge component* or *KC*. However, in the main article we used the term *item* as a synonym to avoid introducing unnecessary jargon. The course organization was such that all variants of a KC were introduced in a single chapter. During practice trials, COLT randomly drew one variant of a KC.

For each chapter, KCs were assigned to the three scheduling conditions for each student in order to satisfy three criteria: (1) each KC occurred equally often in each condition across students, (2) each condition was assigned the same number of KCs for each student, and (3) the assignments of each pair of KCs were independent across students. Although these three counterbalancing criteria could not be satisfied exactly because the total number of items in a chapter and the total number of students were outside our control, the first two were satisfied ± 1, and the third served as the objective of an assignment-optimization procedure that we ran.

Procedure

In each COLT session, students began with a study-to-proficiency stage with material from only the current chapter. This phase involved a drop-out procedure which began by sequentially presenting items from the current chapter in randomly ordered retrieval-practice trials. After the set of items from the current chapter had been presented, items that the student translated correctly were dropped from the set, trial order was re-randomized, and students began another pass through the reduced set. Once all items from the current chapter had been correctly translated, students proceeded to a review stage where material from any chapter that had been introduced so far could be presented for study.

The review stage lasted until the end of the session. During the review stage, items from any of the chapters covered so far in the course were eligible for study. Review was handled by one of three schedulers, each of which was responsible for a random one-third of the items from each chapter, assigned on a per-student basis. During review, the three schedulers alternated in selecting items for practice. Each selected from among the items assigned to it, ensuring that all items had equal opportunity and that all schedulers were matched for number of review trials offered to them.

Quizzes were administered through COLT using retrieval-practice trials. From a student's perspective, the only difference between a quiz trial and a typical study trial was that quiz trials displayed the phrase "quiz question" above them. From an experimental perspective, the quiz questions are trials selected by neither the review schedulers nor the study-to-proficiency procedure. The motivation for administering the quizzes on COLT was to provide more data to constrain the predictions of our statistical model.

The two cumulative exams followed the same procedure as the end-of-chapter quizzes, except that no corrective feedback was given after each question. Each exam tested half of the KCs from each chapter in each condition, and KCs appeared in only one exam or the other. KCs were assigned randomly to exams per student. Each exam was administered over the Wednesday-Thursday split of class periods, allowing the students up to 90 minutes per exam. The semester calendar is presented in detail in the Supporting Information, along with the distribution of KCs by chapters.

Participants

Participants were eighth graders (median age 13) at a suburban Denver middle school. A total of 179 students—82 males and 97 females—were divided among six class periods of a third-semester Spanish course taught by a single instructor. Every class period met on Mondays, Tuesdays, and Fridays for 50 minutes. Half of the class periods met on Wednesdays and the other half on Thursdays for 90 minues. The end-of-semester cumulative exam was taken by 172 students; the followup exam four weeks later was taken by 176 students. Two students were caught cheating on the end-of-semester exam and were not included in our analyses.

In seventh grade Spanish 1 and 2, these same students had used commercial flashcard software for optional at-home vocabulary practice. Like COLT, that software was preloaded with the chapter-by-chapter vocabulary for the course. Unlike COLT, that software required students to select the chapter that they wished to study. Because review was scheduled by the students themselves and because students had weekly quizzes, students used the software almost exclusively to learn the current chapter's material.

From the students' perspective, COLT was simply a replacement for the software they had been using and a substitute for pencil-and-paper quizzes. Students were not aware of the details of our experimental manipulation, beyond the notion that the software would spend some portion of study time reviewing older vocabulary items.

Students occasionally missed COLT sessions due to illness or other absences from class. They were permitted to make up practice sessions (but not weekly graded quizzes) at home if they chose to. They were also permitted to use COLT at home for supplemental practice (see SOM-U for details). As a result, there was significant variability in total usage of COLT from one student to the next. All students are included in our analyses as long as they took either of the cumulative exams.

The instructor who participated in our experiment is a veteran of 22 years of teaching Spanish as a foreign language and has a Master's degree in education. To prevent bias, the instructor was aware only of the experiment's general goal. In previous years, the instructor had given students pencil-and-paper quizzes at the end of each chapter and had also dedicated some class time to the use of paper-based flashcards. COLT replaced both those activities.

Supplementary Online Materials (SOM-U) Improving students' long-term knowledge retention through personalized review

Robert Lindsey, Jeff Shroyer, Harold Pashler, Michael Mozer

Abstract

These supplementary online materials provides additional details concerning the experiment and modeling reported in the main article. The materials are divided into three parts. In part 1, we give additional details about the experimental design and methods. In part 2, we present additional analyses of the experiment results. In part 3, we describe the statistical modeling methodology used throughout the experiment in the personalized review condition.

Experimental Methods

Software

For the experiment, we developed a web-based flashcard tutoring system, the *Colorado Optimized Language Tutor* or *COLT*. Students participating in the study were given anonymous user names and passwords with which they could log in to COLT. Upon logging in, students are taken to a web page showing how many flashcards they have completed on the website, how many flashcards they have correctly answered, and a *Begin Studying* button.

When students click the *Begin Studying* button, they are taken to another web page which presents English-Spanish flashcards through *retrieval-practice trials*. At the start of a retrieval-practice trial, students are prompted with a *cue*—an English word or phrase. Students then attempt to type the corresponding *target*—the Spanish translation—after which they receive feedback (Fig. S1). The feedback consists of the correct translation and a change to the screen's background color: the tint shifts to green when a response is correct and to red when it is incorrect. This form of study exploits the *testing effect*: when students are tested on material and can successfully recall it, they will remember it better than if they had not been tested (Roediger & Karpicke, 2006). Translation was practiced only from English to Spanish because of approximate associative symmetry and the benefit to students from their translating in the direction of the less familiar orthography (Kahana & Caplan, 2002; Schneider, Healy, & Bourne, 2002).

Trials were self-paced. Students were allowed as much time as they needed to type in a response and view feedback. However, students were prevented from advancing past the feedback screen in less than three seconds to encourage them to attend to the feedback. Except on the final exams, students had the option of clicking a button labeled *I don't know* when they could not formulate a response. If they clicked it, the trial was recorded as an incorrect response and the student received corrective feedback as usual. The instructor encouraged students to guess instead of using the button.

COLT provided a simple means of entering diacritical marks through a button labeled *Add Accent*. When a student clicked this button, the appropriate diacritical mark was added to the letter next to the text cursor.

Many stimuli had multiple acceptable translations. If a student produced any of one them, his or her response was judged correct. A response had to have exactly the correct spelling and have the appropriate diacritical marks to be scored as correct, per the instructor's request. Capitalization and punctuation were ignored in scoring a response.

Figure S1: Interface to COLT. Left figure shows the start of a retrieval-practice trial. Right figure shows consequence of an incorrect response.

Implementation

COLT consisted of a front end and a back end. The front end was the website students used to study, which we programmed specifically for this experiment. It was written in a combination of HTML, PHP, and Javascript. Whenever a student submitted an answer in a retrieval practice trial on the website, the response was immediately sent via AJAX to a MySQL database where it was recorded. Database queries were then executed to determine the next item to present to the student, and the chosen item was transmitted back to the student's web browser. Because responses were saved after every trial, students could simply close their browser when they were finished studying and would not lose their progress.

A separate back-end server continually communicated with the front-end server's database. It continually downloaded all data recorded on the website, ran our statistical model to compute posterior expectations of recall probability on each student-KC conditioned on the data recorded until then, and then uploaded the predictions to the front-end database via Python scripts. Thus, whenever an item needed to be chosen by the personalized-spaced scheduler, the scheduler queried the database and selected the item with the appropriate current predicted mean recall probability.

The amount of time it took to run the model's inference algorithm increased steadily as the amount of data recorded increased. It ranged from a few seconds early in the experiment to half an hour late in the semester, by which point we had recorded nearly 600,000 trials. In the future, the inference method could easily be changed to a sequential Monte Carlo technique in order for it to scale to larger applications. The posterior inference algorithm was written in C++. In the event of a back-end server failure, the front-end was programmed to use the most recently computed predictions in a round-robin fashion, cycling through material in an order prioritized by the last available model predictions. On at least three occasions, the back-end server crashed and was temporarily offline.

The front-end server was rented from a private web-hosting company, and the back-end server was a dedicated quad-core machine located in our private laboratory space on the campus of the University of Colorado at Boulder. We used two servers in order to separate the computationally demanding inference algorithm from the task of supplying content to the students' web browsers. This division of labor ensured that the students' interactions with the website were not sluggish.

	Textbook Section	Day of Study	# Words & Phrases	# KCs	# KCs on Quiz
Chapter 1 Introduced	4-1	1	99	25	24
Chapter 2 Introduced	4-1	8	46	22	22
Chapter 3 Introduced	4-2	15	26	26	25
Chapter 4 Introduced	4-3	21	30	16	16
Chapter 5 Introduced	5-1	42	28	18	18
Chapter 6 Introduced	5-2	49	62	17	15
Chapter 7 Introduced	5-2	56	31	16	16
Chapter 8 Introduced	5-3	63	14	14	12
Chapter 9 Introduced	5-3	74	24	24	21
Chapter 10 Introduced	6-1	84	49	43	-
Cumulative Exam 1	-	89-90	-	112	-
Cumulative Exam 2	-	117-118	-	109	-

Table S 1: Calendar of events throughout the semester.

Semester Calendar

The course proceeded according to the calendar in Table S1. The table shows the timeline of presentation of 10 chapters of material and the cumulative end-of-semester exams, along with the amount of material associated with each chapter. The amount of material is characterized in terms of both the number of unique words or phrases (column 4) and the number of KCs (column 5).

The course was organized such that in-class introduction of a chapter's material was coordinated with practice of the same material using COLT. Typically, students used COLT during class time for three 20-30 minute sessions each week, with exceptions due to holiday schedules or special classroom activities. New material was typically introduced in COLT on a Friday, followed by additional practice the following Tuesday, followed by an end-of-chapter quiz on either Wednesday or Thursday. In addition to the classroom sessions, students were allowed to use COLT at their discretion from home. Each session at home followed the same sequence as the in-class sessions. Figure S2 presents pseudocode outlining the selection of items for presentation within each session.

The quizzes were administered on chapters 1-9 and counted toward the students' course grade. On each quiz, the instructor chose the variants of a KC that would be tested. For all but the chapter 8 quiz, the instructor selected material only from the current chapter. The chapter 8 quiz had material from chapters 7 and 8. Quizzes typically tested most of the KCs in a chapter (column 6 of Table S1).

Two cumulative final exams were administered following introduction of all 10 chapters. Cumulative exam 1 occurred around the end of the semester; cumulative exam 2 occurred four weeks later, following an intersemester break. Students were not allowed to use COLT between semesters.

Experimental Results: Additional Analyses

The amount of use of COLT varied by chapter due to competing classroom activities, the amount of material introduced in each chapter, the number of class days devoted to each chapter, and the amount of at-home use of COLT. Fig. S3 presents the median number of retrieval practice trials undergone by students, broken down by chapter and response type (correct, incorrect, and "I don't know") and by in-class versus at-home use of COLT.

Fig. S4 graphs the proportion correct recall on the two final exams by class section and review scheduler. The class sections are arranged in order from best to worst performing. An Analysis of Variance (ANOVA) was conducted on each exam with the dependent variable being proportion recalled on the exam and with three factors: class period, scheduler (massed, generic spaced, personalized spaced), and chapter of course

% Study to Proficiency Phase

Let $c \leftarrow$ the current chapter

Let $x \leftarrow$ the set of KCs in chapter c

While x **is not empty and the student has not quit**

    Let $y \leftarrow$ a random permutation of x

    For each KC i **in** y

        Execute a retrieval practice trial on i

        If the student answered correctly

            Remove i from x

% Review Phase

Let $m \leftarrow$ {MASSED, GENERIC, PERSONALIZED}

Let $z \leftarrow$ a random permutation of m

Let $k \leftarrow 0$

Until the student quits

    Let $w \leftarrow$ the set of all items assigned to scheduler z_k for the student

    If $z_k =$ MASSED

        Let $i \leftarrow$ the KC in w and in chapter c that has been least recently studied by the student

    Else If $z_k =$ GENERIC

        If $c > 0$

            Let $i \leftarrow$ the KC in w and in chapter $c - 1$ that has been least recently studied by the student

        Else

            Let $i \leftarrow$ the KC in w and in chapter c that has been least recently studied by the student

    Else $z_k =$ PERSONALIZED

        Let $i \leftarrow$ the KC in w and in any of chapters $1 \ldots c$ whose current posterior mean recall probability for the student is closest to the desirable difficulty level d

    Execute a retrieval practice trial on i

    Set $k = (k + 1)$ modulo 3

Figure S2: Pseudocode showing the sequence of steps that each student undergoes in a study session in the experiment. Students begin in a study-to-proficiency phase on material from the chapter currently being covered in class. If students complete the study-to-proficiency phase, they proceed to a review phase. During the review phase, trials alternate between schedulers so that each scheduler receives an equal number of review trials. The graded end-of-chapter quizzes did not follow this pseudocode and instead presented the same sequence of instructor-chosen retrieval practice trials to all students, ensuring that all students saw the same questions and had them in the same order.

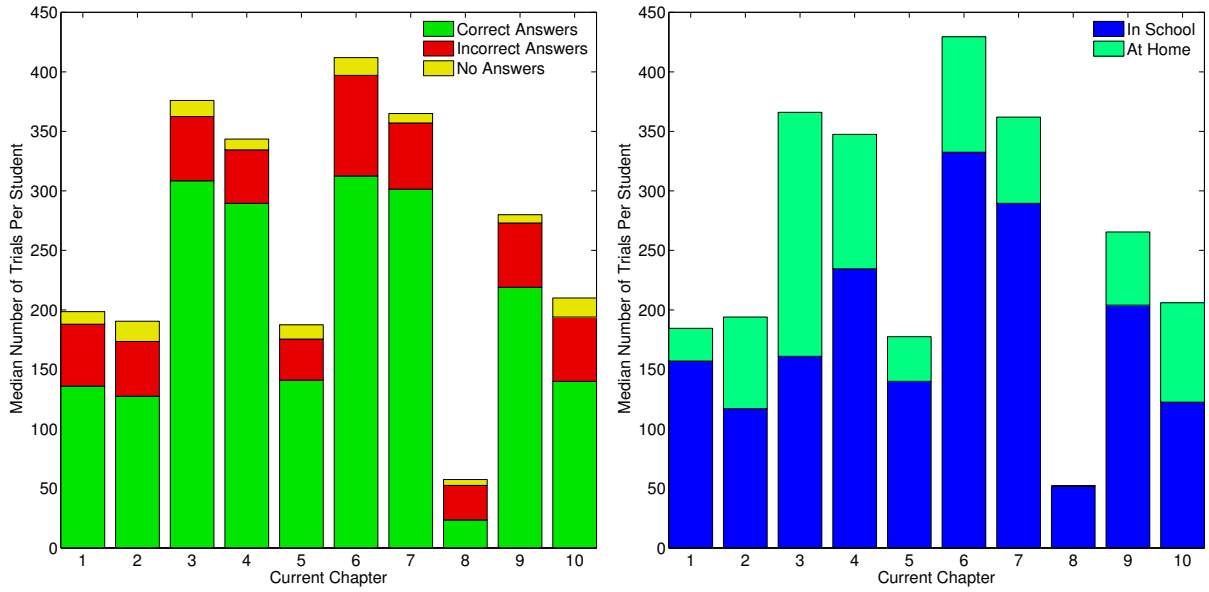

Figure S3: Median number of study trials undergone while each chapter was being covered in class. In the left panel, the number is broken down by whether the student responded correctly, responded incorrectly, or clicked "I don't know." In the right panel, the number is broken down by whether the trial happened on a weekday during school hours or not. Chapter 8 has few trials because it was covered in class only the day before a holiday break and the day after it.

(1-10). The main effect of scheduler is highly reliable in both exams (exam 1: $F(2, 328) = 52.3$, $p < .001$; exam 2: $F(2, 340) = 55.1$, $p < .001$); as reported in the primary article, the personalized-spaced scheduler outperforms the two control schedulers. The main effect of class period is significant in both exams (exam 1: $F(5, 164) = 6.77$, $p < .001$; exam 2: $F(5, 170) = 9.72$, $p < .001$): some sections perform better than others. A scheduler $\times$ chapter interaction is observed (exam 1: $F(18, 2952) = 8.90$, $p < .001$; $F(9, 1530) = 29.67$, $p < .001$), as one would expect from Fig. 4: the scheduler has a larger influence on retention for the early chapters in the semester. The scheduler $\times$ period interaction is not reliable (exam 1: $F(10, 328) = 1.44$, $p = .16$; exam 2: $F(10, 340) = 1.36$, $p = .20$), nor is the three-way scheduler $\times$ period $\times$ chapter interaction (exam 1: $F(90, 2952) < 1$; exam 2: $F(90, 3060) < 1$).

Figure S6 splits Figure 2b from the main article into performance separately on the end-of-semester exam and the exam administered 28 days later. As the ANOVAs in the previous paragraph suggest, the qualitative pattern of results is similar across the two exams. Note that Figure 2b includes only students who took both exams, whereas Figure S6 shows students who took either exam. Only a few students missed each exam.

Fig. S5 shows the mean quiz scores on each chapter for the three conditions. Except for the chapter 8 quiz, all quizzes were on only the current chapter. Ignore chapter 8 for the moment, and also ignore chapter 1 because the three conditions were indistinguishable the first week of the semester. An ANOVA was conducted with the dependent variable being proportion correct on a quiz and with the chapter number (2-7, 9) as a factor. Only the 156 students who took all seven of these quizzes were included. The main effect of review scheduler is significant ($F(2, 310) = 11.8$, $p < .001$): the massed scheduler does best on the quizzes—89.4% versus 87.2% and 88.1% for the generic and personalized spaced schedulers—because it provided the largest number of study trials on the quizzed chapter. The main effect of the chapter is significant ($F(6, 930) = 49.0$, $p < .001$), and the scheduler $\times$ chapter interaction is not reliable ($F(12, 1860) = 1.56$, $p = .096$). The simultaneous advantage of the massed condition on immediate tests (the chapter quizzes) and the spaced conditions on delayed tests (the final exams) is consistent with the experimental literature on the distributed-practice effect.

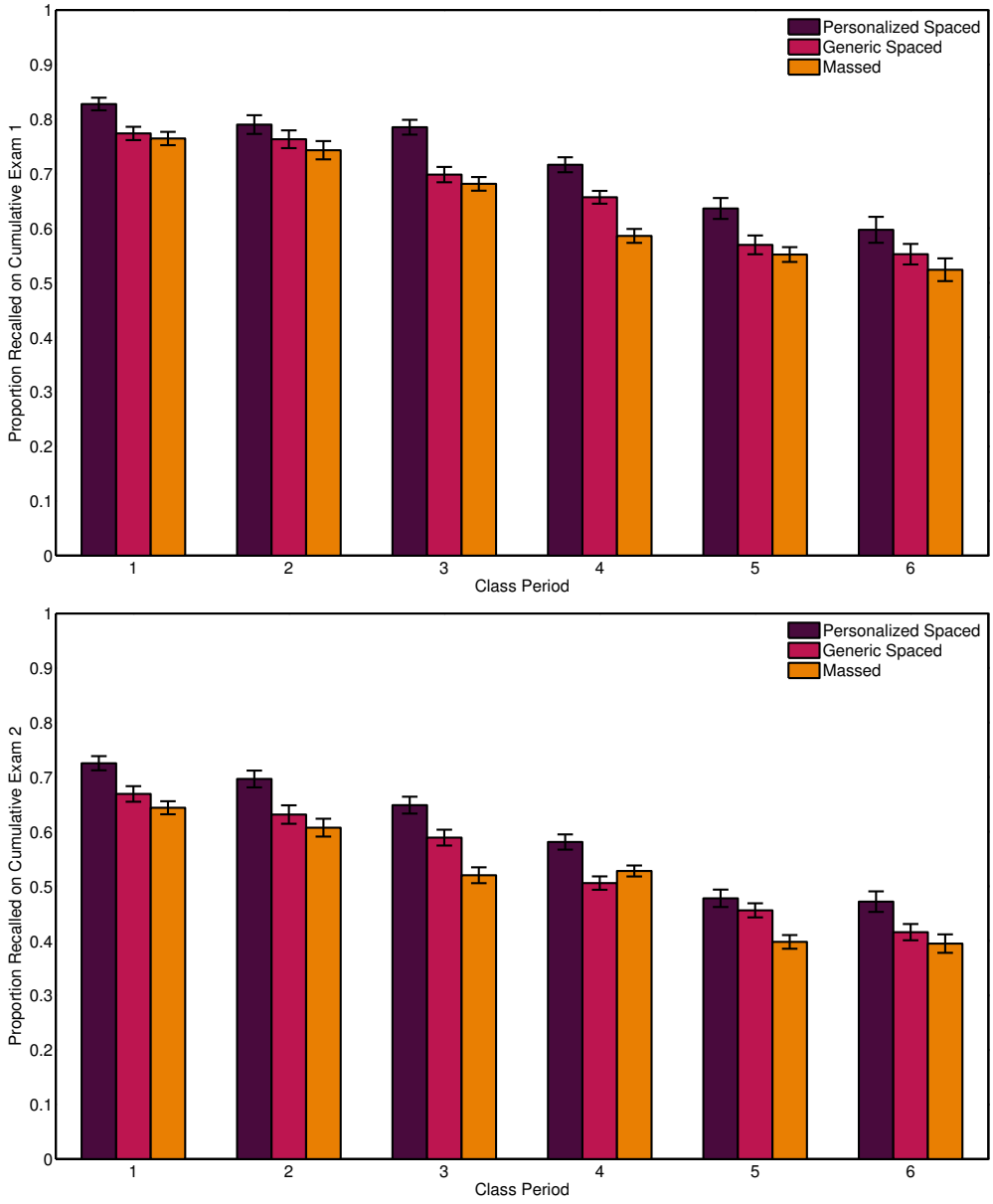

Figure S4: Scores on cumulative exams 1 and 2 for each class period. Each group of bars is a class period. The class periods are presented in rank order by their mean Exam 1 score.

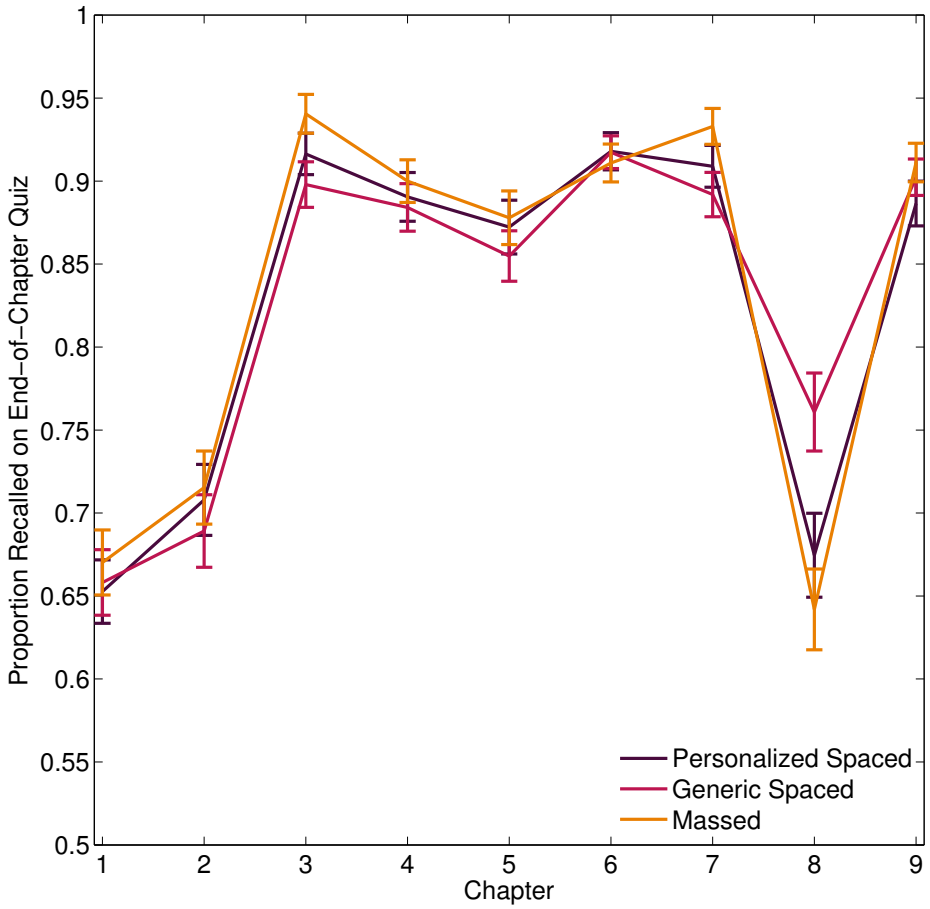

Figure S5: End-of-chapter quiz scores by chapter. Note that the chapter 8 quiz included material from chapter 7, but all the other quizzes had material only from the current chapter. There was no chapter 10 quiz.

Returning to the chapter 8 quiz, which we omitted from the previous analysis, it had the peculiarity that the instructor chose to include material mostly from chapter 7. Because the generic-spaced condition focused review on chapter 7 during chapter 8, it fared the best on the week 8 quiz (generic spaced 76.1%, personalized spaced 67.5%, massed 64.2%; $F(2, 336) = 14.4$, $p < .001$).

Modeling

Other models that consider time

A popular methodology that does consider history of study is Bayesian knowledge tracing (Corbett & Anderson, 1995). Although originally used for modeling procedural knowledge acquisition, it could just as well be used for other forms of knowledge. However, it is based on a simple two-state model of learning which makes the strong assumptions that forgetting curves are exponential and decay rates are independent of the past history of study. The former is inconsistent with current beliefs about long-term memory (Wixted & Carpenter, 2007), and the latter is inconsistent with empirical observations concerning spacing effects (Pavlik & Anderson, 2005). Knowledge tracing's success is likely due to its use in modeling massed practice, and therefore it has not had to deal with variability in the temporal distribution of practice or the long-term

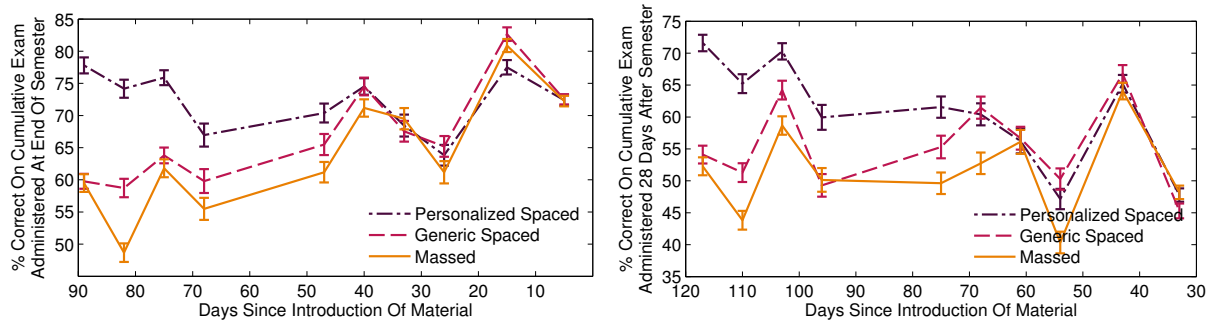

Figure S6: Mean score on each of the two exams as a function of the number of days that had passed since the material was introduced. The two exams show similar results by scheduler and chapter.

retention of skills.

Hierarchical Distributional Assumptions

Bayesian models have a long history in the intelligent tutoring community (Corbett & Anderson, 1995; Koedinger & MacLaren, 1997; Martin & van Lehn, 1995). In virtually all such work, parameters of these models are fit by maximum likelihood estimation, meaning that parameters are found that make the observations have high probability under a model. However, if the model has free parameters that are specific to the student and/or KC, fitting the parameters independently of one another can lead to overfitting. An alternative estimation procedure, hierarchical Bayesian inference, is advocated by statisticians and machine learning researchers to mitigate overfitting. In this approach, parameters are treated as random variables with hierarchical priors. We adopt this approach in DASH, using the following distributional assumptions:

$$\begin{aligned} \alpha_s &\sim \text{Normal}(\mu_\alpha, \sigma_\alpha^2) \\ (\mu_\alpha, \sigma_\alpha^{-2}) &\sim \text{Normal-Gamma}(\mu_0^{(\alpha)}, \kappa_0^{(\alpha)}, a_0^{(\alpha)}, b_0^{(\alpha)}) \\ \delta_i &\sim \text{Normal}(\mu_\delta, \sigma_\delta^2) \\ (\mu_\delta, \sigma_\delta^{-2}) &\sim \text{Normal-Gamma}(\mu_0^{(\delta)}, \kappa_0^{(\delta)}, a_0^{(\delta)}, b_0^{(\delta)}) \end{aligned} \tag{1}$$

where the Normal-Gamma distribution has parameters $\mu_0, \kappa_0, a_0, b_0$. Individual ability parameters α_s are drawn independently from a normal distribution with unknown population-wide mean μ_α and variance σ_α^2. Similarly, individual difficulty parameters δ_i are drawn independently from a normal distribution with unknown population-wide mean μ_δ and variance σ_δ^2. When the unknown means and variances are marginalized via the conjugacy of the Normal distribution with a Normal-Gamma prior, the parameters of one individual student or item become tied to the parameters of other students or items (i.e., are no longer independent). This lends statistical strength to the predictions of individuals with little data associated with them, which would otherwise be underconstrained. The weights ϕ_w and ψ_w are independently distributed with improper priors: $p(\phi_w) \propto$ constant, $p(\psi_w) \propto$ constant.

Gibbs-EM Inference Algorithm

Inference in DASH consists of calculating the posterior distribution over recall probability for all student-KC pairs at the current time given all data observed up until then. In this section, we present a flexible algorithm for inference in DASH models that is readily applicable to variants of the model (e.g., DASH[MCM] and DASH[ACT-R]). For generality, we write the probability of a correct response in the kth trial of a KC i for a student s in the form

$$P(R_{sik} = 1 \mid \alpha_s, \delta_i, \mathbf{t}_{1:k}, \mathbf{r}_{1:k-1}, \boldsymbol{\theta}) = \sigma(\alpha_s - \delta_i + h_{\boldsymbol{\theta}}(\mathbf{t}_{s,i,1:k}, \mathbf{r}_{s,i,1:k-1})) \tag{2}$$

where $\sigma(x) \equiv [1+\exp(-x)]^{-1}$ is the logistic function, $\mathbf{t}_{s,i,1:k}$ are the times at which trials 1 through k occurred, $\mathbf{r}_{s,i,1:k-1}$ are the binary response accuracies on trials 1 through $k-1$. $h_{\boldsymbol{\theta}}$ is a model-specific function that summarizes the effect of study history on recall probability; it is governed by parameters $\boldsymbol{\theta} \equiv \{\theta_1, \theta_2, \ldots, \theta_M\}$ where M is the number of parameters. The DASH model described in the main text is defined as

$$h_{\boldsymbol{\theta}} = \sum_{w=0}^{W-1} \theta_{2w+1} \log(1 + c_{si,w+1}) + \theta_{2w+2} \log(1 + n_{si,w+1}) \tag{3}$$

where the summation is over W time windows.

Given an uninformative prior over $\boldsymbol{\theta}$, the optimal hyperparameters $\boldsymbol{\theta}^*$ are the ones that maximizes the marginal likelihood of the data

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \iint P(\mathbf{r}|\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\theta}) p(\boldsymbol{\alpha}) p(\boldsymbol{\delta}) \; d\boldsymbol{\alpha} \; d\boldsymbol{\delta} \tag{4}$$

Though this is intractable to compute, we can use an EM algorithm to search for θ^*. An outline of the inference algorithm is as follows

1. Initialize $\boldsymbol{\theta}^{(0)}$ and set $i = 1$
2. Iteration i
   - **E-step:** Draw N samples $\left\{\boldsymbol{\alpha}^{(\ell)}, \boldsymbol{\delta}^{(\ell)}\right\}_{\ell=1}^{N}$ from $p(\boldsymbol{\alpha}, \boldsymbol{\delta} \mid \mathbf{r}, \boldsymbol{\theta}^{(i-1)})$ using a Gibbs sampler
   - **M-step:** Find
   $$\boldsymbol{\theta}^{(i)} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{\ell=1}^{N} \log P(\mathbf{r}, \boldsymbol{\alpha}^{(\ell)}, \boldsymbol{\delta}^{(\ell)} | \boldsymbol{\theta}) \tag{5}$$
3. $i \leftarrow i + 1$, go to 2 if not converged.

Following these steps, $\theta^{(i)}$ will reach a local optimum to the marginal likelihood. Each $\theta^{(i)}$ is guaranteed to be a better set of hyperparameters than $\theta^{(i-1)}$.

E-Step. The E-step involves drawing samples from $p(\boldsymbol{\alpha}, \boldsymbol{\delta} \mid \mathbf{r}, \boldsymbol{\theta}^{(i-1)})$ via Markov chain Monte Carlo (MCMC). We performed inference via *Metropolis within Gibbs* sampling. This MCMC algorithm is appropriate because drawing directly from the conditional distributions of the model parameters is not feasible. The algorithm requires iteratively taking a Metropolis-Hastings step from each of the conditional distributions of the model. These are

$$\begin{aligned} p(\alpha_s \mid \boldsymbol{\alpha}_{\neg s}, \boldsymbol{\delta}, \boldsymbol{\theta}, \mathbf{r}) &\propto p(\alpha_s \mid \boldsymbol{\alpha}_{\neg s}) \prod_{i,k} P(r_{sik} \mid \alpha_s, \delta_i, \boldsymbol{\theta}) \\ p(\delta_i \mid \boldsymbol{\delta}_{\neg i}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \mathbf{r}) &\propto p(\delta_i \mid \boldsymbol{\delta}_{\neg i}) \prod_{s,k} P(r_{sik} \mid \alpha_s, \delta_i, \boldsymbol{\theta}) \end{aligned} \tag{6}$$

where $\boldsymbol{\alpha}_{\neg s}$ denotes all ability parameters excluding student s's and $\boldsymbol{\delta}_{\neg i}$ denotes all difficulty parameters excluding item i's. Both $p(\alpha_s \mid \boldsymbol{\alpha}_{\neg s})$ and $p(\delta_i \mid \boldsymbol{\delta}_{\neg i})$ are non-standard t-distributions. We have left the dependence of these distributions on the model's hyperparameters implicit. The products are over the data likelihood of student-item-trials affected by a change in the parameter in question (e.g., a change in α_s affects the likelihood of all trials undergone by s).

M-Step. Let S be the number of students, I be the number of items, and n_{si} be the number of trials undergone by student s on item i. By assumption, the hyperparameters of the normal-gamma distributions are not part of $\boldsymbol{\theta}$. Thus, the M-step is equivalent to finding the hyperparameters which maximize the expectation of the data log-likelihood,

$$\boldsymbol{\theta}^{(i)} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{\ell=1}^{N} \log P(\mathbf{r} \mid \boldsymbol{\alpha}^{(\ell)}, \boldsymbol{\delta}^{(\ell)}, \boldsymbol{\theta}) \tag{7}$$

For convenience, denote $\mathcal{L}^{(\ell)} \equiv \log P(\mathbf{r}|\boldsymbol{\alpha}^{(\ell)}, \boldsymbol{\delta}^{(\ell)}, \boldsymbol{\theta})$, $\gamma^{(\ell)} = a_s^{(\ell)} - d_i^{(\ell)} + h$, and use the shorthand $h \equiv h_{\boldsymbol{\theta}}(\mathbf{t}_{s,i,1:k}, \mathbf{r}_{s,i,1:k-1})$. We have

$$\mathcal{L}^{(\ell)} = \sum_{s=1}^{S} \sum_{i=1}^{I} \sum_{k=1}^{n_{si}} r_{sik} \gamma^{(\ell)} - \log\left(1 + e^{\gamma^{(\ell)}}\right) \tag{8}$$

We can solve for $\boldsymbol{\theta}^{(i)}$ by function optimization techniques. We used Matlab's *fminunc* function which exploits the gradient and hessian of $\mathcal{L}^{(\ell)}$. The gradient is given by

$$\frac{\partial \mathcal{L}^{(\ell)}}{\partial \theta_j} = \sum_{s=1}^{S} \sum_{i=1}^{I} \sum_{k=1}^{n_{si}} (r_{sik} - \sigma(\gamma^{(\ell)})) \frac{\partial h}{\partial \theta_j} \tag{9}$$

for all $j \in 1 \ldots M$. The hessian is given by

$$\frac{\partial^2 \mathcal{L}^{(\ell)}}{\partial \theta_z \theta_j} = \sum_{s=1}^{S} \sum_{i=1}^{I} \sum_{k=1}^{n_{si}} (r_{sik} - \sigma(\gamma^{(\ell)})) \frac{\partial^2 h}{\partial \theta_z \partial \theta_j} - \sigma(\gamma^{(\ell)})(1 - \sigma(\gamma^{(\ell)})) \frac{\partial h}{\partial \theta_z} \frac{\partial h}{\partial \theta_j} \tag{10}$$

for all $z \in 1 \ldots M, j \in 1 \ldots M$.

Model Comparison And Evaluation

The models were trained on all data up to a given point in timetask on the 597,990 retrieval practice trials COLT recorded across the semester-long experiment. (These trials include the quizzes and material assigned to all three scheduling conditions.)

We divided these time-ordered trials into contiguous segments with each segment containing 1% of the trials. We then tested each model's ability to predict a segment n given segments $1 \ldots n-1$ as training data, for $n \in \{2 \ldots 100\}$. We scored each model's across-segment average prediction quality using cross entropy[1] and mean per-trial prediction error[2]. The former method more strongly penalizes heldout trials for which the model assigned low probability to the observed recall event.

Because the amount of at-home COLT usage was largely self-determined, the number of trials undergone throughout the semester varied greatly from student to student. Because students who study much more than their peers will tend to be over-represented in the training and test data, they are generally the easiest to predict. However, models should provide good predictions regardless of how much a student studies. Therefore, we report results for a *normalized* version of the two error metrics in which each student contributes equally to the reported value. We calculated the mean error metric across heldout trials for each student in the test segment, then averaged across students. Thus, each student's mean contributed equally to the overall error metric.

- *Baseline Model.* As a baseline, we created a model which predicts that recall probability in a heldout trial for a student is the proportion of correct responses that student has made in the training data.

- ACT-R. Pavlik and Anderson (Pavlik & Anderson, 2005, 2008) extended the ACT-R memory model to account for the effects of temporally distributed study; we will refer to their model as ACT-R. The model includes parameters similar to the ability and difficulty factors in IRT that characterize individual differences among students and among KCs. Further, the model allows for parameters that characterize each student-KC pair. Whereas DASH is fully specified by eight parameters,[3] the number of free parameters in the ACT-R model increases multiplicatively with the size of the student pool and amount of study material. To fit the data recorded in this experiment, the model requires over forty thousand

[1] Cross entropy is calculated as the negative of the mean per-trial $\log_2$-likelihood.

[2] Letting $\hat{p}$ be the expected recall probability and $r \in \{0, 1\}$ be the recall event, we define prediction error of a trial as $(1 - \hat{p})^r \, \hat{p}^{1-r}$

[3] The eight model parameters are the parameters of the two normal-gamma priors, which we set to the reference prior.

free parameters, and there are few data points per parameter. Fitting such a high-dimensional and weakly constrained model is an extremely challenging problem. Pavlik and Anderson had the sensible idea of inventing simple heuristics to adapt the parameters as the model is used. We found that these heuristics did not fare well for our experiment. Therefore, in our simulation of ACT-R, we eliminated the student-KC specific parameters and used Monte Carlo maximum likelihood estimation, which is a search method that repeatedly iterates through all the model parameters, stochastically adjusting their values so as to increase the data log-likelihood.[4]

- IRT. We created a hierarchical Bayesian version of the Rasch Item-Response Theory model with the same distributional assumptions over $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ as made in DASH. We will refer to this model as IRT. It corresponds to the assumption that $h_{\boldsymbol{\theta}} = 0$ in Equation 2.

- DASH[ACT-R]. We experimented with a version of DASH which does not have a fixed number of time windows, but instead—like ACT-R—allows for the influence of past trials to continuously decay according to a power-law. Using the DASH likelihood equation in Equation 2, the model is formalized as

$$h_{\boldsymbol{\theta}} = c \log(1 + \sum_{k'<k} m_{r_{k'}} t_{k'}^{-d}) \tag{11}$$

  where the four hyperparameters are $c \equiv \theta_1$, $m_0 \equiv \theta_2$, $m_1 \equiv \theta_3$, $d \equiv \theta_4$. We will refer to this model as DASH[ACT-R] because of its similarity to ACT-R. Like DASH, it is a synthesis of data-driven and theory-based models for predicting student recall over time. This formalism ensures that recall probability is non-zero on the first trial of a student-KC, which is necessary in our application because students are expected to have prior experience with the material before using COLT. The parameter h is split in two: a value h_1 for when the student responded correctly in a trial, $r(k') = 1$, and a value h_0 for when the student responded incorrectly, $r(k') = 0$. This gives each trace a different initial strength depending on response accuracy.

- DASH[MCM]. Motivated by the Multiscale Context Model (MCM), a model of the spacing effect we developed which has a fixed set of continuously, exponentially decaying memory traces (Mozer, Pashler, Cepeda, Lindsey, & Vul, 2009), we experimented with a version of DASH which has a fixed number of continuously decaying windows. The model assumes that the counts n_{siw} and c_{siw} are incremented at each trial and then decay over time at a window-specific exponential rate τ_w. Formally,

$$h_{\boldsymbol{\theta}} = \sum_{w=0}^{W-1} \theta_{2w+1} \log(1 + \tilde{c}_{si,w+1}(t)) + \theta_{2w+2} \log(1 + \tilde{n}_{si,w+1}(t)) \tag{12}$$

  where

$$\tilde{n}_{siw}^{(k)} = 1 + \tilde{n}_{siw}^{(k-1)} \exp(-\tfrac{t_k - t_{k-1}}{\tau_w}) \qquad \tilde{c}_{siw}^{(k)} = r_{sik} + \tilde{c}_{siw}^{(k-1)} \exp(-\tfrac{t_k - t_{k-1}}{\tau_w}) \tag{13}$$

  We determined the decay rates by deduction. Three desired qualitative properties of the exponential half-half of each window are

  - The smallest half-life should be about 30 minutes, roughly the time between COLT prediction updates. Thus, $t_1^{(1/2)} = .0208$ and so $\tau_1 = .0301$.
  - The largest half-life should be about the length of the experiment. Thus, $t_W^{(1/2)} = 90$ and so $\tau_W = 129.8426$.

[4]Note that the ACT-R model assumes that the base level activation b is given by $b \equiv \alpha_s - \delta_i + \beta_{si}$, where the student abiliity α_s and KC difficulty δ_i combine with a student-KC parameter β_{si}. Because having one parameter per student-KC leads to extreme overfitting, we set all $\beta_{si} = 0$. We estimated missing δ_i values by averaging across the difficulty parameter of all KCs with training data. We bounded the model predictions to lie on $[.001, .999]$ to keep the cross-entropy well-defined. The model ordinarily can assign zero probability to recall events, hence does not always have a finite log-likelihood.

– The half-lives should be exponentially increasing. It is important to be able to differentiate between, for example, whether a trial is 1 or 2 days old. Differentiating between, for example, trials that are 60 vs. 61 days old is less important. Thus, we want $t_w^{(1/2)} = ct_{w-1}^{(1/2)}$ where c is a constant.

Given these constraints and because we want to have $W = 5$ windows as in DASH, we can solve for the decay rates of each window as $\boldsymbol{\tau}_{1:W} = \{0.0301, 0.2434, 1.9739, 16.0090, 129.8426\}$. Like DASH and DASH[ACT-R], DASH[MCM] is a synthesis of data-driven and theory-based models for predicting student recall over time.

For the Bayesian models—IRT, DASH, DASH[ACT-R], and DASH[MCM]—we collected 200 posterior samples during each E-step after a 100 iteration burn-in. The MCMC sampler generally mixed quickly, which allowed us to have such a small burn-in. To reduce autocorrelation, we used every other sample. The Gibbs-EM algorithm generally converged to a solution within 3-6 iterations. For ACT-R, we ran 1500 iterations of the stochastic hill-climbing algorithm and kept the maximum likelihood solution.

References

Corbett, A., & Anderson, J. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling & User-Adapted Interaction*, *4*, 253-278.

Kahana, M., & Caplan, J. (2002). Associative asymmetry in probed recall of serial lists. *Memory & Cognition*, *30*(6), 841-849.

Koedinger, K., & MacLaren, B. (1997). Implicit strategies and errors in an improved model of early algebra problem solving. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (p. 382-387). Hillsdale, NJ: Erlbaum.

Martin, J., & van Lehn, K. (1995). Student assessment using bayesian nets. *International Journal of Human-Computer Studies*, *42*, 575-591.

Mozer, M., Pashler, H., Cepeda, N., Lindsey, R., & Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems* (Vol. 22, p. 1321-1329).

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

Pavlik, P., & Anderson, J. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*, 101-117.

Roediger, H., & Karpicke, J. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249-255.

Schneider, V., Healy, A., & Bourne, L. (2002). What is learned under difficult conditions is hard to forget: Contextual inference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, *46*, 419-440.

Wixted, J., & Carpenter, S. (2007). The wickelgren power law and the ebbinghaus savings function. *Psychological Science*, *18*, 133-134.