Research Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
1–9

DOI: 10.1177/0956797613504302
pss.sagepub.com
SAGE

Improving Students' Long-Term Knowledge Retention Through Personalized Review

Robert V. Lindsey[1,2], Jeffery D. Shroyer[1], Harold Pashler[3], and Michael C. Mozer[1,2]

[1]Institute of Cognitive Science, University of Colorado, Boulder; [2]Department of Computer Science, University of Colorado, Boulder; and [3]Department of Psychology, University of California, San Diego

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 ever-growing 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. Using a cumulative exam administered after the semester's end, we compared time-matched review strategies and found that 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.

Keywords

long-term memory, declarative memory, spacing effect, adaptive scheduling, classroom education, Bayesian modeling, educational psychology, individual differences

Received 6/18/13; Revision accepted 8/15/13

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% to 35% of basic science knowledge after 1 year, more than 50% by the next year (Custers, 2010), and 80% to 85% after 25 years (Custers & ten Cate, 2011).

Forgetting is influenced by the temporal distribution of study. For more than 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 also 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 with the imperative to regularly introduce new material.

Method

We incorporated systematic, temporally distributed review into third-semester, eighth-grade Spanish foreign-language

Corresponding Author:
Michael C. Mozer, Institute of Cognitive Science, University of Colorado, Boulder, CO 80309-0430
E-mail: mozer@colorado.edu

instruction using a Web-based flash-card tutoring system, the Colorado Optimized Language Tutor (COLT). Throughout the semester, 179 students used COLT to drill on 10 chapters of material, which were introduced at approximately 1-week intervals. COLT presented vocabulary words and short sentences in English and required students to type the Spanish translations, after which corrective feedback was provided. The software was used both to practice newly introduced material and to review previously studied material. More information about the software and semester schedule can be found in the Experimental Methods section of Additional Methods and Results in the Supplemental Material available online.

For each chapter of course material, students engaged in three 20- to 30-min 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. In the third session, these activities were preceded by a quiz on the current chapter, which counted toward the course grade. During the review phase of each session, study items from all chapters covered so far in the course were eligible for presentation. Selection of items for the review phase was handled by three different schedulers.

The *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, the educational software used by these students allowed them 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, which is consistent with the preference for massed study found by Cohen, Yan, Halamish, and Bjork (2013).

The *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). Given the time frame of a semester—during which material must be retained for 1 to 3 months—a 1-week lag between initial study and review results in 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 items (Fig. 1).

The *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. For example, suppose that a student has practiced an item twice, failing to get the correct answer 15 days ago but succeeding 9 days ago. Given these sparse observations, it would seem that one cannot reliably predict the student's current ability regarding the item. However, data from 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.

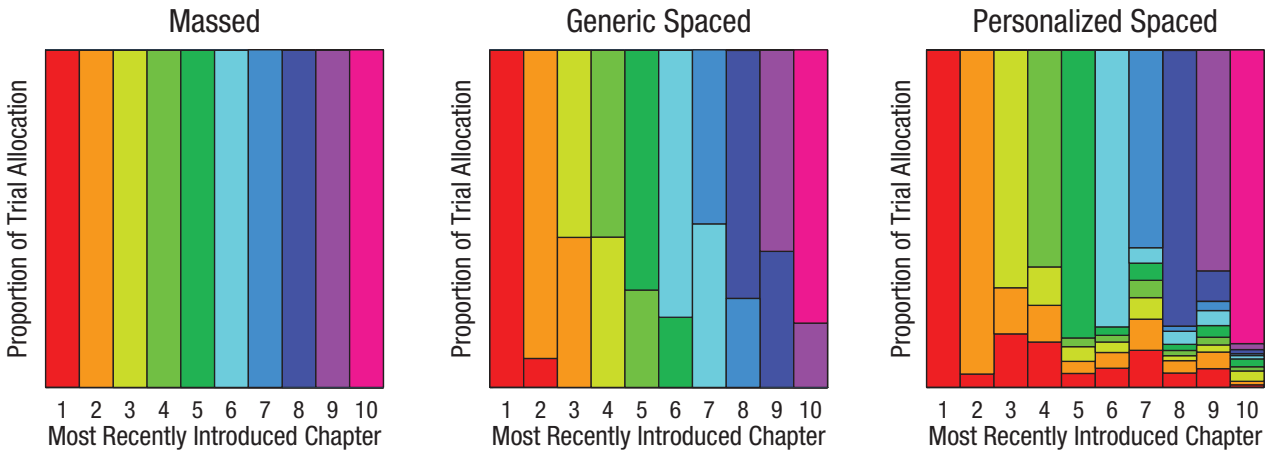

Fig. 1. Trial allocation of the three review schedulers. Course material was introduced one chapter at a time, generally at 1-week intervals. Each vertical slice indicates the across-student average proportion of trials spent in a given week studying each of the chapters introduced up to that point. (Each slice includes trials from both the study-to-proficiency and the review phases.) Each chapter is indicated by a unique color.

The model we used defines memory strength as being jointly dependent on factors relating to (a) an item's latent difficulty, (b) a student's latent ability, and (c) the amount, timing, and outcome of past study. We refer to the model with the acronym DASH (i.e., 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 participants by randomly assigning one third of a chapter's items to each scheduler, with assignment counterbalanced across participants. During review, the schedulers alternated in selecting items for retrieval practice. Each scheduler selected from among the items assigned to it, ensuring that all items had equal opportunity. All schedulers administered an equal number of review trials. Figure 1 and Table 1 present statistics of how often and when individual items were studied by individual students for each scheduler over the time course of the experiment. More information about the experimental procedure, subject pool, and study materials can be found in Materials, Procedure, and Participants in the Supplemental Material available online.

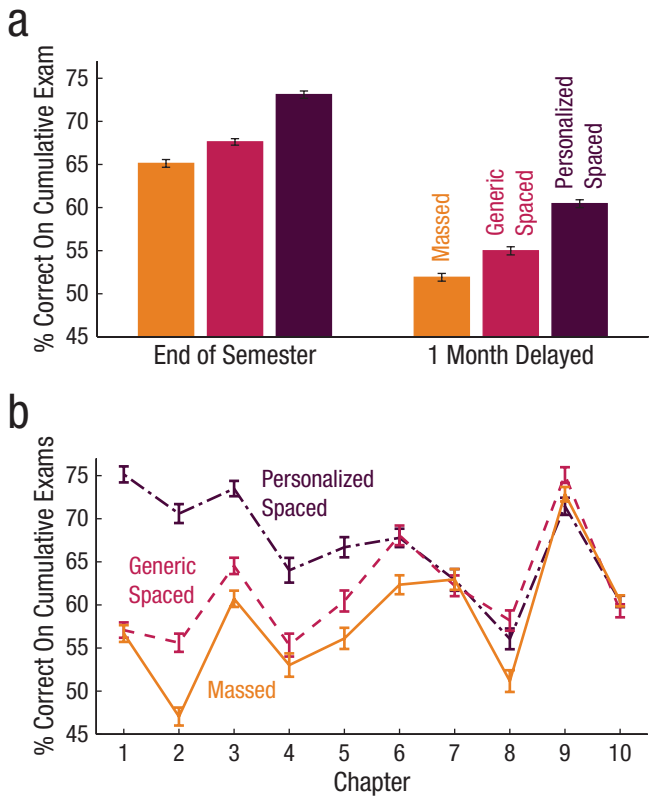

Fig. 2. Scores on the cumulative end-of-semester exams. The bar graph (a) presents mean score as a function of condition for each of the two exams separately. The line graph (b) presents mean score across the two exams as a function of the chapter in which the material was introduced, separately for each condition. Error bars indicate ±1 *SE*, calculated within subjects (Masson & Loftus, 2003).

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, with items randomly selected for each student and balanced across chapters and schedulers; no corrective feedback was provided. On the first exam, retention for items assigned to the personalized spaced scheduler was 12.4% higher than retention for items assigned to the massed scheduler, $t(169) = 1.01$, $p < .001$, Cohen's $d = 1.38$, and 8.3% better than retention for items assigned to the generic spaced scheduler, $t(169) = 8.2$, $p < .001$, Cohen's $d = 1.05$ (Fig. 2a). Over the 28-day intersemester break, the forgetting rate was 18.1%, 17.1%, and 15.7% for the massed, generic spaced, and personalized spaced conditions, respectively, so that the advantage of personalized review became even larger. On the second exam, personalized review boosted retention by 16.5% over massed review, $t(175) = 1.11$, $p < .001$, Cohen's $d = 1.42$, and by 10.0% over generic review, $t(175) = 6.59$, $p < .001$, Cohen's $d = 0.88$ (Fig. 2a).

The schedulers had their primary impact for material introduced earlier in the semester (Fig. 2b), which makes sense because memory for that material had the most opportunity to be manipulated via review. The personalized spaced scheduler produced a large benefit for early chapters in the semester without sacrificing efficacy for later chapters. Among students who took both exams,

Table 1. Presentation Statistics of the Three Schedulers for Individual Students on Individual Items

	Massed scheduler		Generic spaced scheduler		Personalized spaced scheduler	
Presentation statistic	Mean	*SD*	Mean	*SD*	Mean	*SD*
Number of study-to-proficiency trials	7.58	6.70	7.57	6.49	7.56	6.47
Number of review trials	8.03	11.99	8.05	12.14	8.03	9.65
Number of days between review trials	0.12	1.43	1.69	3.29	4.70	6.39

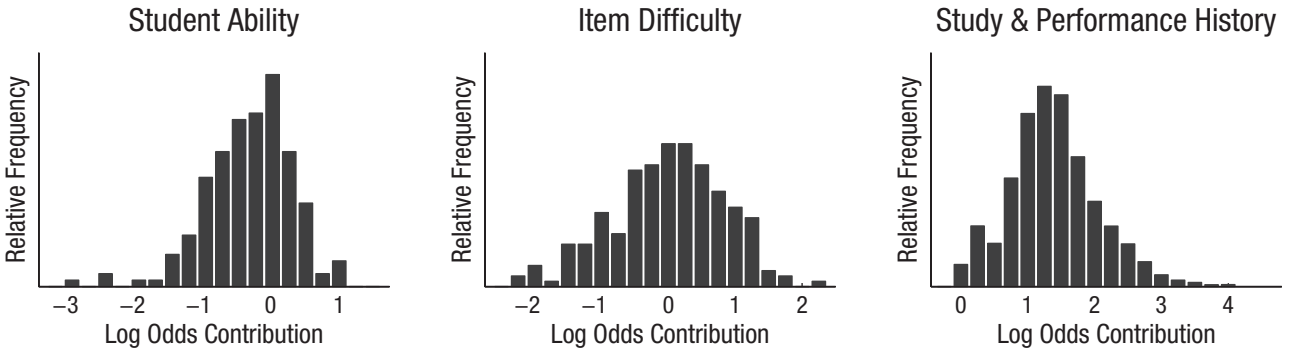

Fig. 3. Histograms of three inferred factors, expressed in terms of their additive contribution to predicted log odds of recall. Each factor varies over 3 log units, which corresponds to a possible modulation of .65 in recall probability.

only 22.3% and 13.5% scored better in the generic spaced and massed conditions, respectively, than in the personalized spaced condition.

Note that massed review was spaced by usual laboratory standards, being spread out over at least 6 days (new material was introduced on a Friday and practiced until Wednesday or Thursday the following week). This fact may explain both the small benefit of the generic spaced over the massed scheduler and the absence of a spacing effect (generic and personalized spaced schedulers outperforming the massed scheduler) for the final chapters (see Fig. 2).

DASH infers three factors contributing to recall success: an item's difficulty, a student's ability, and the study history of the specific student on the specific item. Histograms of these inferred contributions showed substantial variability (Fig. 3), so decisions about what items to review were markedly different across individual students and items.

DASH predicts a student's response accuracy for an item at a point in time given the response history 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; we then used the models to predict the data for each set 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 was weighted equally. Figure 4 compares DASH against five alternatives: a *baseline* model that predicted a student's future performance to be the proportion of correct responses the student had 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, multiscale context model). Details of the alternative models, model evaluations, and additional analyses of the experimental results are available in Additional Methods and Results in the Supplemental Material.

The three variants of DASH performed better than the alternatives. Each variant had two key components: (a) a dynamic representation of study history that characterized learning and forgetting and (b) a Bayesian approach to inferring latent difficulty and ability factors. Models that omitted the first component (baseline and IRT) or the second component (baseline and ACT-R) did not fare as well. The DASH variants all performed similarly. Because these variants differed only in the manner in which the temporal distribution of study and recall

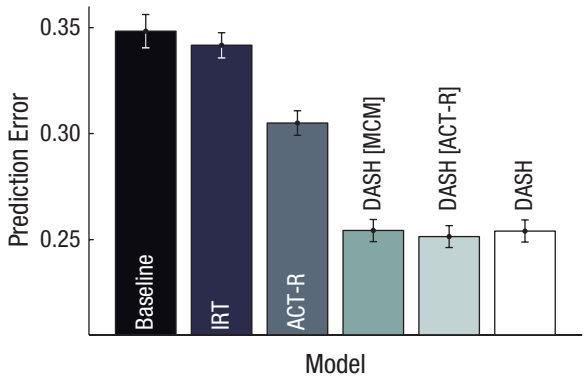

Fig. 4. Accumulative prediction error of six models using the data from the semester-long experiment. The models are as follows: a baseline model that predicts performance from the proportion of correct responses made by each student, a model based on item-response theory (IRT), a model based on Pavlik and Anderson's (2005) ACT-R model, DASH, and variants of DASH including components of ACT-R and the multiscale context model (MCM). Error bars indicate ±1 *SEM*.

outcomes was 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 with spaced presentation of material (Carpenter, Pashler, & Cepeda, 2009; Seabrook, Brown, & Solity, 2005; Sobel, Cepeda, & Kapler, 2011) and laboratory-based investigations of *adaptive scheduling* techniques, which are used to select material for an individual to study on the basis of that individual's past study history and performance (e.g., Atkinson, 1972).

Previous explorations of temporally distributed study in real-world educational settings targeted a relatively narrow body of course material to which participants were unlikely to be exposed outside the experimental context. Further, these studies compared just a few spacing conditions, and the spacing was the same for all participants and materials, as in 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 1 week and a test the following week. This compressed time scale limited 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 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 went beyond past efforts in its scope: It spanned the time frame of a semester, covered the content of an entire course, and introduced 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 min 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 confounded personalization and the coarse temporal distribution of review (Fig. 1, Table 1). However, indiscriminate review of older material is unlikely to be beneficial because it comes at the expense of newer material, and because time limitations permit the selection of only a small fraction of the ever-growing collection of candidate material.

Any form of personalization requires estimates of an individual's memory strength for specific knowledge. Previously proposed adaptive-scheduling algorithms based their estimates on observations from only the given 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 (Fig. 4).

Outside 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 present material for review 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 given 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 (VanLehn, 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 because of lack of appropriately timed rehearsal. Although it is common to pay lip service to the benefits of review, comprehensive and appropriately timed review is beyond what any teacher or student can reasonably arrange. Our results suggest that a digital tool that 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, one must infer a student's *knowledge state*—the dynamically varying strength of each atomic knowledge component (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. Here, 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., on SAT tests), IRT decomposes response accuracies into student- and question-specific parameters. The simplest form of IRT (Rasch, 1961) models the probability 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 on question i by student s, R_{si}, is given by

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

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; Chi, Koedinger, Gordon, Jordan, & VanLehn, 2011; Pavlik, Cen, & Koedinger, 2009). This class of models, known as *additive-factors models*, has the following form:

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

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

Although this class of models personalizes predictions on the basis of a student's 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, the multiscale context model (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 that 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 on the basis of 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 on the basis of the temporal distribution of study. Formally, DASH posits that

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

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 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 into DASH a bias 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 1st 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 that omitted one term or the other or that 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., $s_w \propto e^{\rho w}$).

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 as a result of 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: $s = \{1/24, 1, 7, 30, \infty\}$. Additional Methods and Results in the Supplemental Material available online describes inference in the model.

Personalized review scheduling

DASH predicts the probability of successful recall for each student on each KC. Although these predictions are necessary for optimal scheduling of review, optimal scheduling is computationally intractable because it requires planning over all possible futures. Consequently, the Colorado Optimized Language Tutor (COLT) uses a heuristic policy for selecting review material. This policy is 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, Khajah et al. found that 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 θ of .33 being best over a range of conditions. Note that with θ greater than 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 naturally into a threshold-based policy, assuming one has a model like DASH that can accurately estimate retrieval probability.

Author Contributions

R. V. Lindsey, M. C. Mozer, and H. Pashler developed the study concept. All authors contributed to the study design. R. V. Lindsey designed and implemented the computer software. J. D. Shroyer conducted the study. R. V. Lindsey performed the data analysis and interpretation under the supervision of M. C. Mozer and H. Pashler. R. V. Lindsey and M. C. Mozer were primarily responsible for writing the manuscript, and J. D. Shroyer and H. Pashler provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments

We thank F. Craik, A. Glass, J. L. McClelland, H. L. Roediger, III, and P. Wozniak for valuable feedback on the manuscript.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This research was supported by a National Science Foundation (NSF) Graduate Research Fellowship, by NSF Grants

SBE-0542013 and SMA-1041755, and by a collaborative-activity award from the McDonnell Foundation.

Supplemental Material

Additional supporting information may be found at http://pss.sagepub.com/content/by/supplemental-data

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* (pp. 185–205). Cambridge, MA: 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 M. Ikeda, K. D. Ashley, & T.-W. Chan (Eds.), *Intelligent Tutoring Systems: 8th International Conference* (pp. 164–175). Berlin, Germany: Springer.

Cen, H., Koedinger, K., & Junker, B. (2008). Comparing two IRT models for conjunctive skills. In B. P. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Intelligent Tutoring Systems: 9th International Conference* (pp. 796–798). Berlin, Germany: Springer.

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

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

Chi, M., Koedinger, K., Gordon, G., Jordan, P., & VanLehn, K. (2011). Instructional factors analysis: A cognitive model for multiple instructional interventions. In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, & J. Stamper (Eds.), *Proceedings of the 4th International Conference on Educational Data Mining* (pp. 61–70). Retrieved from http://educationaldatamining.org/EDM2011/proceedings-2

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*, *39*, 1682–1696.

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

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

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

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*, 4–58.

Kerfoot, B., Fu, Y., Baker, H., Connelly, D., Ritchey, M., & Genega, E. (2010). 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*, 331–337.

Khajah, M., Lindsey, R., & Mozer, M. C. (2013). Maximizing students' retention via spaced review: Practical guidance from computational models of memory. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 758–763). Austin, TX: Cognitive Science Society.

Leitner, S. (1972). *So lernt man lernen* [How to learn]. Freiburg im Breisgau, Germany: Herder.

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 22*. Retrieved from http://books.nips.cc/nips22.html

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* (pp. 531–538). Brighton, England: IOS Press.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Math, Statistics, and Probability* (pp. 321–333). Berkeley: University of California Press.

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.

VanLehn, 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 Technology in Education* (pp. 17–20). Retrieved from http://www.isca-speech.org/archive/slate_2007

van Rijn, D. H., van Maanen, L., & van Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. In A. Howes, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the Ninth International Conference on Cognitive Modeling*. Retrieved from http://sideshow.psyc.bbk.ac.uk/rcooper/iccm2009/proceedings/

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.