Improving students’ long-term knowledge retention through personalized review

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

* Department of Computer Science, University of Colorado, Boulder
† Department of Psychology, University of California, San Diego

Longer version accepted for publication in Psychological Science.

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. 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.

1 Introduction

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. The curse of forgetting occurs over many time scales. It happens from one week to the next as, for example, new skills are introduced in a math class, and it happens from one semester to the next as, for example, physics students advance from a mechanics course to an electricity and magnetism course. 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 [1], and 80–85% after 25 years [2].

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 [3]. Although spaced practice is beneficial in many tasks beyond rote memorization [4] and shows promise in improving educational outcomes [5], 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 are likely to be unreliable [6, 7]. 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 flashcard tutoring system, the Colorado Optimized Language
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. (left) The massed scheduler had students spend all their time only on the current chapter. (middle) The generic-spaced scheduler had students spend their review time studying the previous chapter. (right) The personalized-spaced scheduler made granular decisions about what each student should study.

*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 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 [8].

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 [3, 9] and computational models [10, 11]. 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, as reflected in the probability of correct recall. 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 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, we can make predictions about the performance of a specific student for a specific item at a given moment. Our 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.

---

1 Indeed, at the end of our experiment, an informal survey of students indicated a widespread concern that mandatory review interfered with learning new material. Students requested a means of opting out of review.
Figure 2: (left) Mean scores on the two cumulative end-of-semester exams, taken 28 days apart. (right) 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 and did so without sacrificing efficacy on later chapters. All error bars indicate ±1 within-student standard error [12].

approach, DASH characterizes both individual differences and the temporal dynamics of learning and forgetting.

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.

Table 1: Presentation statistics of individual student-items over entire experiment. Mean and std. dev. shown.

<table>
<thead>
<tr>
<th></th>
<th>Massed</th>
<th>Generic</th>
<th>Personalized</th>
</tr>
</thead>
<tbody>
<tr>
<td># study-to-criterion trials</td>
<td>7.58 (6.70)</td>
<td>7.57 (6.49)</td>
<td>7.56 (6.47)</td>
</tr>
<tr>
<td># review trials</td>
<td>8.03 (11.99)</td>
<td>8.05 (12.14)</td>
<td>8.03 (9.65)</td>
</tr>
<tr>
<td># days between review trials</td>
<td>0.12 (1.43)</td>
<td>1.69 (3.29)</td>
<td>4.70 (6.39)</td>
</tr>
</tbody>
</table>

2 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 ($t(169) = 10.1, p < .0001, \text{Cohen's } d = 1.38$) 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 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 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 [13] 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. The rightmost plot in Figure 3 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) [14], a model of spacing effects based on the memory component of ACT-R [15], 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 section 4.

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.

3 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 [16, 17, 18], 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 [19, 20, 21].

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. (One exception is a study in [22] that examines the effect of compressing the time scale of an entire course by a factor of three.)

Previous evaluations of adaptive scheduling have demonstrated the advantage of one algorithm over another or over nonadaptive algorithms [23, 24, 25], but these evaluations have been confined to the laboratory and have spanned a relatively short time scale. The most ambitious previous experiment [24] 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 [9]. 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.

3.1 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 3).

Apart from the academic literature, two traditional adaptive-scheduling techniques have attracted a degree of popular interest: the Leitner [20] system and SuperMemo [21]. Both aim to review material at the point of desirable difficulty [26]—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. Empirical and theoretical analyses provide qualitative support for such an expanding-spacing schedule [27, 28]. 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. Software vendors tend to be protective of their intellectual property, but the few scheduling algorithms we have been able to investigate can be claimed to optimize long-term retention only through tenuous verbal arguments.

Traditionally, students are motivated to review when their grade is affected. Although frequent cumulative exams or homework assignments might impel students to undergo spaced review, we have shown that this one-size-fits-all solution is significantly less effective than personalized review.

3.2 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 [29] 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.

4 Modeling Student Knowledge States

To personalize review, we needed to 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) [14]. 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 [30] parameterizes the log-odds that a particular student will correctly answer a particular question through a student-specific ability factor \( \alpha_s \) and a question-specific difficulty factor \( \delta_i \). Formally, the probability of recall success or failure \( R_{si} \) on question \( i \) by student \( s \) is given by

\[
Pr(R_{si} = 1 | \alpha_s, \delta_i) = \left[ 1 + e^{-(\alpha_s - \delta_i)} \right]^{-1}
\]

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 [31, 32, 33, 34]. This class of models, known as *additive factors models*, has the form: \( \Pr(R_{si} = 1 | \alpha_s, \delta_i, \gamma, \mathbf{m}_{si}) = [1 + e^{-\left(\alpha_s - \delta_i + \sum_j \gamma_j m_{sij}\right)}]^{-1} \), where \( j \) is an index over factors, \( \gamma_j \) is the skill level associated with factor \( j \), and \( m_{sij} \) is the \( j \)th 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 [11] and a theory based on the ACT-R declarative memory module [15]. 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 \( \Pr(R_{si} = 1 | \alpha_s, \delta_i, \phi, \psi) = \text{logistic} \left[ \alpha_s - \delta_i + \sum_w \phi_w \log(1+c_{siw}) - \psi_w \log(1+n_{siw}) \right] \), 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 \( \phi_w \) and \( \psi_w \) 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.

The difference of factors inside the summation of DASH determines a power law of practice. 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. The formulation builds a bias into DASH that additional study in a given time window helps, but has logarithmically diminishing returns.

To model effects of temporally distributed study and forgetting, DASH includes multiple time windows. Window-specific parameters (\( \psi_w, \phi_w \)) 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. 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., \( s_w \propto e^{\beta w} \).

Bayesian models have a long history in the intelligent tutoring community [35]. In virtually all such work, parameters of these models are fit by maximum likelihood. 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. We instead used hierarchical Bayesian inference to mitigate overfitting. In our model, individual ability and difficulty parameters are drawn independently from normal distributions with unknown population-wide mean and variance. When the unknown means and variances are marginalized, the parameters of one individual student or item become tied to the parameters of other students or items. This lends statistical strength to the predictions of individuals with little data associated with them, which would otherwise be underconstrained.
References


