Predicting and Improving Memory Retention: Psychological Theory Matters in the Big Data Era

Michael C. Mozer[*†]
Robert V. Lindsey[*]
[*]Department of Computer Science and
[†]Institute of Cognitive Science
University of Colorado, Boulder

February 11, 2016

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

word count: 8441 (text) + 586 (captions, etc.)

Abstract

Cognitive psychology has long had the aim of understanding mechanisms of human memory, with the expectation that such an understanding will yield practical techniques that support learning and retention. Although research insights have given rise to qualitative advice for students and educators, we present a complementary approach that offers quantitative, individualized guidance. Our approach synthesizes theory-driven and data-driven methodologies. Psychological theory characterizes basic mechanisms of human memory shared among members of a population, whereas machine-learning techniques use observations from a population to make inferences about individuals. We argue that despite the power of big data, psychological theory provides essential constraints on models. We present models of forgetting and spaced practice that predict the dynamic time-varying knowledge state of an individual student for specific material. We incorporate these models into retrieval-practice software to assist students in reviewing previously mastered material. In an ambitious year-long intervention in a middle-school foreign language course, we demonstrate the value of systematic review on long-term educational outcomes, but more specifically, the value of adaptive review that leverages data from a population of learners to personalize recommendations based on an individual's study history and past performance.

1 Introduction

Human memory is fragile. The initial acquisition of knowledge is slow and effortful. And once mastery is achieved, the knowledge must be exercised periodically to mitigate forgetting. Understanding the cognitive mechanisms of memory has been a longstanding goal of modern experimental psychology, with the hope that such an understanding will lead to practical techniques that support learning and retention. Our specific aim is to go beyond the traditional qualitative forms of guidance provided by psychology and express our understand in terms of computational models that characterize the temporal dynamics of a learner's *knowledge state*. This knowledge state specifies what material the individual already grasps well, what material can be easily learned, and what material is on the verge of slipping away. Given a knowledge-state model, individualized teaching strategies can be constructed that select material to maximize instructional effectiveness.

In this chapter we describe a hybrid approach to modeling knowledge state that combines the complementary strengths of psychological theory and a big-data methodology. Psychological theory characterizes basic mechanisms of human memory shared among members of a population, whereas the big-data methodology allows us to use observations from a population to make inferences about individuals. We argue that despite the power of big data, psychological theory provides essential constraints on models, and that despite the success of psychological theory in providing a qualitative understanding of phenomena, big data enables quantitative, individualized predictions of learning and performance.

This chapter is organized as follows. First, we discuss the notion of knowledge state and the challenges involved in inferring knowledge state from behavior. Second, we turn to traditional psychological theory, describing key human-memory phenomena and computational models that have been developed to explain these phenomena. Third, we explain the data-mining technique known as *collaborative filtering*, which involves extracting patterns from large data sets for the purpose of making personalized recommendations. Traditionally, collaborative filtering has been used by e-commerce merchants to recommend products to buy and movies to watch, but in our context, we use the technique to recommend material to study. Fourth, we illustrate how a synthesis of psychological theory and collaborative filtering improves predictive models. And finally, we incorporate our predictive models into software that provides personalized review to students, and show the benefit of this type of modeling in two semester-long experiments with middle-school students.

2 Knowledge State

In traditional electronic tutors (e.g., Anderson, Conrad, & Corbett, 1989; Koedinger & Corbett, 2006; Martin & VanLehn, 1995), the modeling of a student's knowledge state has depended on extensive handcrafted analysis of the teaching domain and a process of iterative evaluation and refinement. We present a complementary approach to inferring knowledge state that is fully automatic and independent of the content domain. We hope to apply this approach in any domain whose mastery can be decomposed into distinct, separable *components* of knowledge or *items* to be learned (van Lehn, Jordan, & Litman, 2007). Applicable domains range from the concrete to the abstract, and from the perceptual to the cognitive, and span qualitatively different forms of knowledge from declarative to procedural to conceptual.

What does it mean to infer a student's knowledge state, especially in a domain-independent manner? The knowledge state consists of latent attributes of the mind such as the strength of a specific declarative memory or a stimulus-response association, or the psychological representations of interrelated concepts. Because such attributes cannot be observed directly, a theory of knowledge state must be be validated through its ability to *predict* a student's future abilities and performance.

Inferring knowledge state is a daunting challenge for three distinct reasons.

1. *Observations of human behavior provide only weak clues about the knowledge state.* Consider fact learning, the domain which will be a focus of this chapter. If a student performs cued recall trials, as when flashcards are used for drilling, each retrieval attempt provides one bit of information: whether it is successful or not. From this meager signal, we hope to infer quantitative properties of the memory trace, such as its strength, which we can then use to predict whether the memory will be accessible in an hour, a week, or a month. Other behavioral indicators can be diagnostic, including response latency (Lindsey, Lewis, Pashler, & Mozer, 2010; Mettler & Kellman, 2014; Mettler, Massey, & Kellman, 2011) and confidence (Metcalfe & Finn, 2011), but they are also weak predictors.

2. *Knowledge state is a consequence of the entire study history*, i.e., when in the past the specific item and related items were studied, the manner and duration of study, and previous performance indicators. Study history is particularly relevant because all forms of learning show forgetting over time, and unfamiliar and newly acquired information is particularly vulnerable (Rohrer & Taylor, 2006; Wixted, 2004). Further, the temporal distribution of practice has an impact on the durability of learning for various types of material (Cepeda, Pashler, Vul, & Wixted, 2006; Rickard, Lau, & Pashler, 2008).

3. *Individual differences are ubiquitous in every form of learning.* Taking an example from fact learning

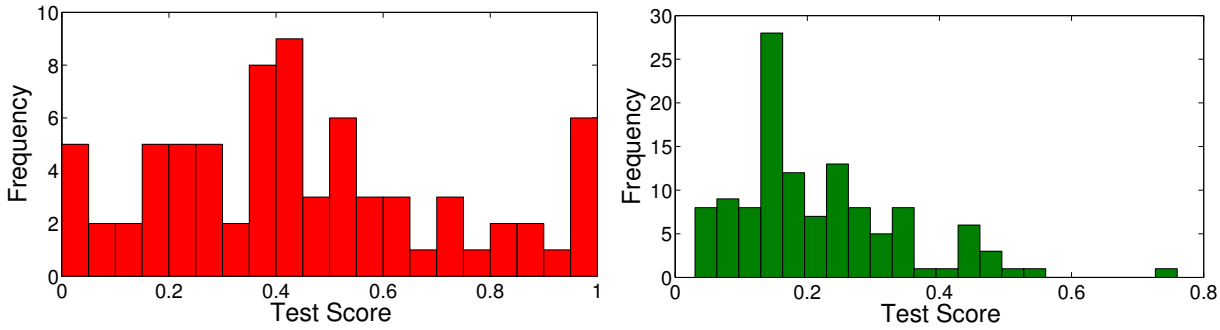

Figure 1: (left) Histogram of proportion of items reported correctly on a cued recall task for a population of 60 students learning 32 Japanese-English vocabulary pairs (Kang et al., 2014); (right) Histogram of proportion of subjects correctly reporting an item on a cued recall task for a population of 120 Lithuanian-English vocabulary pairs being learned by roughly 80 students (Grimaldi et al., 2010)

(Kang, Lindsey, Mozer, & Pashler, 2014), Figure 1a shows extreme variability in a population of 60 participants. Foreign-language vocabulary was studied at four precisely scheduled times over a four week period. A cued-recall exam was administered after an eight week retention period. The the exam scores are highly dispersed despite the uniformity in materials and training schedules. In addition to inter-student variability, inter-item variability is a consideration. Learning a foreign vocabulary word may be easy if it is similar to its English equivalent, but hard if it is similar to a different English word. Figure 1b shows the distribution of recall accuracy for 120 Lithuanian-English vocabulary items averaged over a set of students (Grimaldi, Pyc, & Rawson, 2010). With a single round of study, an exam administered several minutes later suggests that items show a tremendous range in difficulty (*krantas*→*shore* was learned by only 3% of students; *lova*→*bed* was learned by 76% of students).

3 Psychological Theories of Long-Term Memory Processes

The most distressing feature of memory is the inevitability of forgetting. Forgetting occurs regardless of the skills or material being taught, and regardless of the age or background of the learner 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 often assessed by teaching participants some material in a single session and then assessing cued-recall accuracy following some lag t. The probability of recalling the studied material decays according

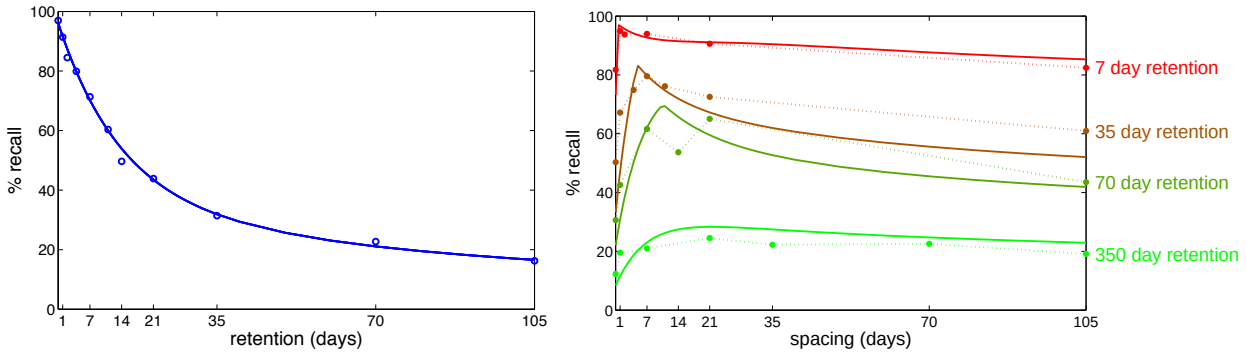

Figure 2: (left) Recall accuracy as a function of lag between study and test for a set of obscure facts; circles represent data provided by Cepeda et al. (2008) and solid line is the best power-law fit. (right) Recall accuracy as a function of the temporal spacing between two study sessions (on the ordinate) and the retention period between the second study session and a final test. Circles represent data provided by Cepeda et al. (2008), and solid lines are fits of the model MCM, as described in the text.

to a generalized power-law as a function of t (Wixted & Carpenter, 2007),

$$Pr(\text{recall}) = m(1 + ht)^{-f},$$

where m, h, and f are constants interpreted as the degree of initial learning ($0 \leq m \leq 1$), a scaling factor on time ($h > 0$), and the memory decay exponent ($f > 0$), respectively. The left panel of Figure 2 shows recall accuracy at increasing study-test lags from an experiment by Cepeda, Vul, Rohrer, Wixted, and Pashler (2008) in which participants were taught a set of obscure facts. The solid line in the Figure is the best fitting power-law forgetting curve.

When material is studied over several sessions, the temporal distribution of study influences the durability of memory. This phenomenon, known as the *spacing effect*, is observed for a variety materials—skills and concepts as well as facts (Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012)—and has been identified as showing great promise for improving educational outcomes (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013).

The spacing effect is typically studied via a controlled experimental paradigm in which participants are asked to study unfamiliar paired associates in two sessions. The time between sessions, known as the *intersession interval* or *ISI*, is manipulated across participants. Some time after the second study session, a cued-recall test is administered to the participants. The lag between second session and the test is known as the *retention interval* or *RI*. Cepeda et al. (2008) conducted a study in which RIs were varied from 7 to 350 days and ISIs were varied from minutes to 105 days. Their results are depicted as circles connected

with dashed lines in the right panel of Figure 2. (The solid lines are model fits, which we discuss shortly.) For each RI, Cepeda et al. find an inverted-U relationship between ISI and retention. The left edge of the graph corresponds to massed practice, the situation in which session two immediately follows session one. Recall accuracy rises dramatically as the ISI increases, reaching a peak and then falling off gradually. The optimal ISI—the peak of each curve—increases with the RI. Note that for educationally relevant RIs on the order of weeks and months, the Cepeda et al. (2009) result indicates that the effect of spacing can be tremendous: optimal spacing can double retention over massed practice. Cepeda, Pashler, Vul, Wixted, and Rohrer (2006) conducted a metaanalysis of the literature to determine the functional relationship between RI and optimal ISI. We augmented their data set with the more recent results of Cepeda et al. (2008) and observed an approximately power-function relationship between RI and optimal ISI (boh in days):

$$Optimal\ ISI = 0.097 RI^{0.812}.$$

This relationship suggests that as material becomes more durable with practice, ISIs should increase, supporting even longer ISIs in the future, consistent with an expanding-spacing schedule as qualitatively embodied in the Leitner method (Leitner, 1972) and SuperMemo (Woźniak, 1990).

Many models have been proposed to explain the mechanisms of the spacing effect (e.g., Benjamin & Tullis, 2010; Kording, Tenenbaum, & Shadmehr, 2007; Mozer, Pashler, Cepeda, Lindsey, & Vul, 2009; Pavlik & Anderson, 2005a; Raaijmakers, 2003; Staddon, Chelaru, & Higa, 2002). These models have been validated through their ability to account for experimental results, such as those in Figure 2, which represent mean performance of a population of individuals studying a set of items. Although the models can readily be fit to an individual's performance for a set of items (e.g., Figure 1a) or a population's performance for a specific item (e.g., Figure 1b), it is a serious challenge in practice to use these models to predict an individual's memory retention for a specific item.

We will shortly describe an approach to making such individualized predictions. Our approach incorporates key insights from two computational models, ACT-R (Pavlik & Anderson, 2005a) and MCM (Mozer et al., 2009), into big-data technique that leverages population data to make individualized predictions. First, we present a brief overview of the two models.

3.1 ACT-R

ACT-R (Anderson et al., 2004) is an influential cognitive architecture whose declarative memory module is

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

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

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

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

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

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

The probability of recall is monotonically related to activation, m:

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

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

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

3.2 MCM

ACT-R is posited on the assumption that memory decay follows a power function. We developed an alternative model, the *Multiscale Context Model* or MCM (Mozer et al., 2009), which provides a mechanistic basis for the power function. Adopting key ideas from previous models of the spacing effect (Kording et al., 2007; Raaijmakers, 2003; Staddon et al., 2002) MCM proposes that each time an item is studied, it is stored in multiple item-specific memory traces that decay at different rates. Although each trace has an exponential decay, the sum of the traces decays approximately as a power function of time. Specifically, trace i, denoted x_i, decays over time according to:

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

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

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

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

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

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

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

The model has 5 free parameters (ϵ_r, and 4 parameters that determine the contributions $\{\gamma_i\}$ and the time constants, $\{\tau_i\}$). MCM was designed such that all of its parameters, with the exception of ϵ_r, could be fully constrained by data that are easy to collect—the function characterizing forgetting folllowing a single study session—which then allows the model to make predictions for data that are difficult to collect—the

function characterizing forgetting following a study schedule consisting of two or more study sessions. MCM has been used to obtain parameter-free predictions for a variety of results in the spacing literature. The solid lines in the right panel of Figure 2 show parameter-free predictions of MCM for the Cepeda et al. (2008) study.

4 Collaborative Filtering

In the last several years, an alternative approach to predicting learners' performance has emerged from the machine-learning community. This approach essentially sets psychological theory aside in favor of mining large data sets collected as students solve problems. To give a sense of the size of these data sets, we note that Khan Academy had over 10 million unique users per month and delivered over 300 million lessons at the end of 2013 (Mullany, 2013). Figure 3a visualizes a data set in which students solve problems over time. Each cell in the tensor corresponds to a specific student solving a particular problem at a given moment in time. The contents of a cell indicate whether an attempt was made and if so whether it was successful. Most of the cells in the tensor are empty. A *collaborative filtering* approach involves filling in those missing cells. While the tensor may have no data about student S solving problem P given a particular study history, the tensor will have data about other similar students solving P, or about S solving problems similar to P. Filling in the tensor also serves to make predictions about future points in time.

Collaborative filtering has a long history in e-commerce recommender systems, for example, Amazon wishes to recommend products to customers and Netflix wishes to recommend movies to its subscribers. The problems are all formally equivalent, simply replace 'student' in Figure 3a with 'customer' or 'subscriber,' and replace 'problem' with 'product' or 'movie.' The twist that distinguishes memory prediction from product or movie prediction is our understanding of the temporal dynamics of human memory. These dynamics are not fully exploited in generic machine-learning approaches. We shortly describe initial efforts in this regard that leverage computational models like ACT-R and MCM to characterize memory dynamics.

Collaborative filtering involves inferring a relatively compact set of latent variables that can predict or explain the observed data. In the case of product recommendations, the latent variables may refer to features of a product (e.g., suitable for children) or of the customer (e.g., has children). In the case of student modeling, the latent variables describe skills required to solve a problem or a student's knowledge state. Using these latent variable representations of problems and student knowledge states, Figure 3b presents an extremely general data-driven framework that has been fruitfully instantiated to predict and

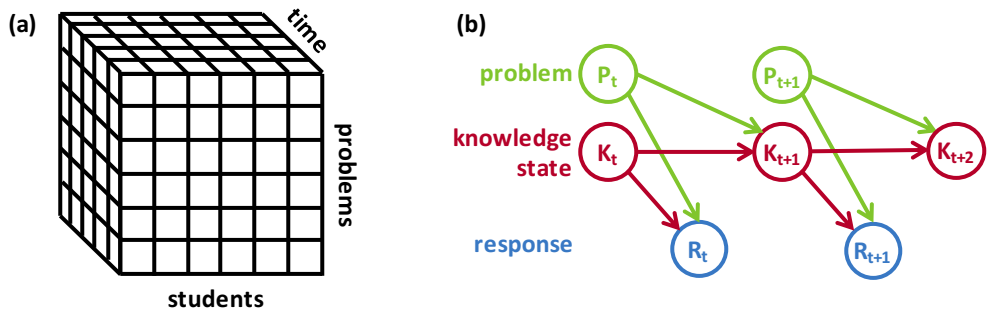

Figure 3: (a) A tensor representing students × problems × time. Each cell describes a student's attempt to solve a problem at a particular moment in time. (b) A naive graphical model representing a teaching paradigm. The nodes represent random variables and the arrows indicate conditional dependencies among the variables. Given a student with knowledge state K_t at time t, and a problem P_t posed to that student, R_t denotes the response the student will produce. The evolution of the student's knowledge state will depend on the problem that was just posed. This framework can be used to predict student responses or to determine an optimal sequence of problems for a particular student given a specific learning objective.

guide learning (e.g., Lan, Studer, & Baraniuk, 2014; Sohl-Dickstein, 2013).

A simple collaborative-filtering method that may be familiar to readers is *item-response theory* or *IRT*, the classic psychometric approach to inducing latent traits of students and items based on exam scores (DeBoek & Wilson, 2004). IRT is used to analyze and interpret results from standardized tests such as the SAT and GRE, which consist of multiple-choice questions and are administered to large populations of students. Suppose that n_S students take a test consisting of n_I items, and the results are coded in the binary matrix $R \equiv \{r_{si}\}$, where s is an index over students, i is an index over items, and r_{si} is the binary (correct or incorrect) score for student s's response to item i. IRT aims to predict R from latent traits of the students and the items. Each student s is assumed to have an unobserved *ability*, represented by the scalar a_s. Each item i is assumed to have an unobserved *difficulty* level, represented by the scalar d_i.

IRT specifies the probabilistic relationship between the predicted response, R_{si} and a_s and d_i. The simplest instantiation of IRT, called the one-parameter logistic (1PL) model because it has one item-associated parameter, is:

$$Pr(R_{si} = 1) = \frac{1}{1 + \exp(d_i - a_s)}. \tag{2}$$

A more elaborate version of IRT, called the 3PL model, includes an item-associated parameter for guessing, but that is mostly useful for multiple-choice questions where the probability of correctly guessing is nonnegligible. Another variant, called the 2PL model, includes parameters that allow for student ability to have a nonuniform influence across items. (In simulations we shortly describe, we explored the 2PL model but found that it provided no benefit over the 1PL model.) Finally, there are more sophisticated latent-trait models

that characterize each student and item not as a scalar but as a feature vector (Koren, Bell, & Volinsky, 2009).

5 Integrating Psychological Theory With Big-Data Methods: A Case Study of Forgetting

IRT is typically applied post hoc to evaluate the static skill level of students (Roussos, Templin, & Henson, 2007). Extensions have been proposed to model a time varying skill level (e.g., Andrade & Tavares, 2005), allowing the technique to predict future performance. However, these extensions are fairly neutral with regard to their treatment of time: skill levels at various points in time are treated as unrelated or as following a random walk. Thus, the opportunity remains to explore dynamic variants of latent-trait models that integrate the longitudinal history of study and properties of learning and forgetting to predict future performance of students. In this section, we take an initial step in this direction by incorporating the latent traits of IRT into a theory of forgetting. Instead of using IRT to directly predict behavioral outcomes, we use latent-trait models to infer variables such as initial memory strength and memory decay rate, and then use the theory of forgetting to predict knowledge state and behavioral outcomes.

5.1 Candidate Models

The forgetting curve we described earlier, based on the generalized power law, is supported by data from populations of students and/or populations of items. The forgetting curve cannot be measured for an individual item *and* a particular student—which we'll refer to as a *student-item*—due to the observer effect and the all-or-none nature of forgetting. Regardless, we will assume the functional form of the curve for a student-item is the same, yielding:

$$Pr(R_{si} = 1) = m(1 + ht_{si})^{-f}, \tag{3}$$

where R_{si} is the response of student s to item i following retention interval t_{si}. This model has free parameters m, h, and f, as described earlier.

We would like to incorporate the notion that forgetting depends on latent IRT-like traits that characterize student ability and item difficulty. Because the critical parameter of forgetting is the memory decay exponent, f, and because f changes as a function of skill and practice (Pavlik & Anderson, 2005a), we can individuate

forgetting for each student-item by determining the decay exponent in Equation 3 from latent IRT-like traits:

$$f = e^{\tilde{a}_s - \tilde{d}_i} \tag{4}$$

We add the tilde to $\tilde{a}_s$ and $\tilde{d}_i$ to indicate that these ability and difficulty parameters are not the same as those in Equation 2. Using the exponential function ensures that f is nonnegative.

Another alternative we consider is individuating the degree-of-learning parameter in Equation 3 as follows:

$$m = \frac{1}{1 + \exp(d_i - a_s)}. \tag{5}$$

With this definition of m, Equation 3 simplifies to 1PL IRT (Equation 2) at $t = 0$. For $t > 0$, recall probability decays as a power-law function of time.

We explored five models that predict recall accuracy of specific student-items: (1) IRT, the 1PL IRT model (Equation 2); (2) MEMORY, a power-law forgetting model with population-wide parameters (Equation 3); (3) HYBRID DECAY, a power-law forgetting model with decay rates based on latent student and item traits (Equations 3 and 4); (4) HYBRID SCALE, a power-law forgetting model with the degree-of-learning based on latent student and item traits (Equations 3 and 5); and (5) HYBRID BOTH, a power-law forgetting model that individuates both the decay rate and degree-of-learning (Equations 3, 4, and 5). The Appendix describes a hierarchical Bayesian inference method for parameter estimation and obtaining model predictions.

5.2 Simulation results

We present simulations of our models using data from two previously published psychological experiments exploring how people learn and forget facts, summarized in Table 1. In both experiments, students were trained on a set of items (cue-response pairs) over multiple rounds of practice. In the first round, the cue and response were both shown. On subsequent rounds, retrieval practice was given: students were asked

Study name	$\mathcal{S}_1$	$\mathcal{S}_2$
Source	Kang et al. (2014)	Cepeda et al. (2008)
Materials	Japanese-English vocabulary	Interesting but obscure facts
# Students	32	1354
# Items	60	32
Rounds of Practice	3	1
Retention Intervals	3 min–27 days	7 sec–53 min

Table 1: Experimental data used for simulations

to produce the appropriate response to each cue. Whether successful or not, the correct response was then displayed. Following this training procedure was a retention interval t_{si} specific to each student and each item, after which an exam was administered. The exam obtained the r_{si} binary value for that student-item.

To evaluate the models, we performed fifty-fold validation. In each fold, a random 80% of elements of R were used for training and the remaining 20% were used for evaluation. Each model generates a prediction, conditioned on the training data, of recall probability at the exam time t_{si}, which can be compared against the observed recall accuracy in the held-out data.

Each model's capability of discriminating successful from unsuccessful recall trials was assessed with a signal-detection analysis (Green & Swets, 1966). For each model, we compute the mean area under the ROC curve (hereafter, AUC) across validation folds as a measure of the model's predictive ability. The measure ranges from 0.5 for random guesses to 1.0 for perfect predictions. The greater the AUC, the better the model is at predicting a particular student's recall success on a specific item after a given lag.

Figure 4a and 4b summarize the AUC values for Studies $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. The baseline MEMORY model performs poorly ($p < .01$ for all pairwise comparisons by a two-tailed t test unless otherwise noted), suggesting that the other models have succeeded in recovering latent student and item traits that facilitate inference about the knowledge state of a particular student-item. The baseline IRT model, which ignores the lag between study and test, does not perform as well as the latent-state models that incorporate forgetting. The HYBRID BOTH model does best in $\mathcal{S}_1$ and ties for best in $\mathcal{S}_2$, suggesting that allowing for individual differences both in degree of learning and rate of forgetting is appropriate. The consistency of results between the two studies is not entirely trivial considering the vastly different retention intervals examined in the two studies (see Table 1).

Generalization to new material

The simulation we described holds out individual student-item pairs for validation. This approach was convenient for evaluating models but does not correspond to the manner in which predictions might ordinarily be used. Typically, we may have some background information about the material being learned, and we wish to use this information to predict how well a new set of students will fare on the material. Or we might have some background information about a group of students, and we wish to use this information to predict how well they will fare on new material. For example, suppose we collect data from students enrolled in Spanish 1 in the fall semester. At the onset of the spring semester, when our former Spanish 1 students begin Spanish 2, can we benefit from the data acquired in the fall to predict their performance on new material?

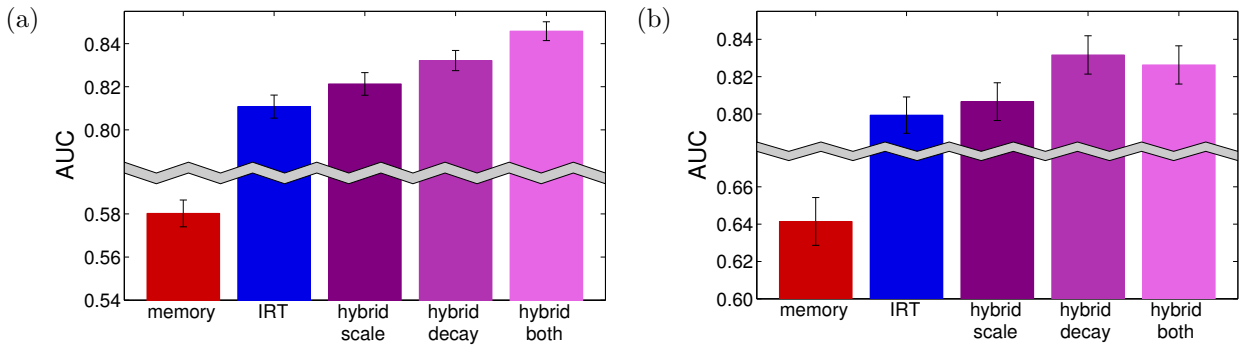

Figure 4: Mean AUC values on the five models trained and evaluated on (a) Study $\mathcal{S}_1$ and (b) Study $\mathcal{S}_2$. The error bars indicate a 95% confidence interval on the AUC value over multiple validation folds. Note that the error bars are not useful for comparing statistical significance of the differences across models, because the validation folds are matched across models, and the variability due to the fold must be removed from the error bars.

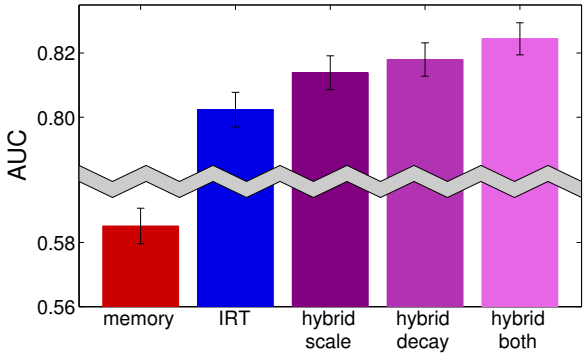

Figure 5: Mean AUC values when random items are held out during validation folds, Study $\mathcal{S}_1$

To model this situation, we conducted a further validation test in which, instead of holding out random student-item pairs, we held out random items for all students. Figure 5 shows mean AUC values for Study $\mathcal{S}_1$ data for the various models. Performance in this item-generalization task is slightly worse than performance when the model has familiarity with both the students and the items. Nonetheless, it appears that the models can make predictions with high accuracy for new material based on inferences about latent student traits and about other items.[1]

To summarize, in this section we demonstrated that systematic individual (student and item) differences can be discovered and exploited to better predict a particular student's retention of a specific item. A model that combines a psychological theory of forgetting with a collaborative filtering approach to latent-

[1]Note that making predictions for new items or new students is principled within the hierarchical Bayesian modeling framework. From training data, the models infer not only student- or item-specific parameters, but also hyperparameters that characterize the population distributions. These population distributions are used to make predictions for new items and new students.

trait inference yields better predictions than models based purely on psychological theory or purely on collaborative filtering. However, the data sets we explored are relatively small—1920 and 43328 exam questions. Ridgeway, Mozer, and Bowles (2016) explore a much larger data set consisting of 46.3 million observations collected from 125k students learning foreign language skills with online training software. Even in this much larger data set, memory retention is better predicted using a hybrid model over a purely data-driven approach.[2] Furthermore, in naturalistic learning scenarios, students are exposed to material multiple times, in various contexts, and over arbitrary temporal distributions of study. The necessity for mining a large data set becomes clear in such a situation, but so does the role of psychological theory, as we hope to convince the reader in the next section.

6 Integrating Psychological Theory With Big-Data Methods: A Case Study of Personalized Review

We turn now to an ambitious project in which we embedded knowledge-state models into software that offers personalized recommendations to students about specific material to study. The motivation for this project is the observation that, at all levels of the educational system, instructors and textbooks typically introduce students to course material in blocks, often termed chapters or lessons or units. At the end of each block, instructors often administer a quiz or problem set to encourage students to master material in the block. Because students are rewarded for focusing on this task, they have little incentive at that moment to rehearse previously learned material. Although instructors appreciate the need for review, the time demands of reviewing old material must be balanced against the need to introduce new material, explain concepts, and encourage students toward initial mastery. Achieving this balance requires understanding when students will most benefit from review. Controlled classroom studies have demonstrated the importance of spaced over massed study (Carpenter, Pashler, & Cepeda, 2009; Seabrook, Brown, & Solity, 2005; Sobel, Cepeda, & Kapler, 2011), but these studies have been a one-size-fits-all type of approach in which all students reviewed all material in synchrony. We hypothesized that personalized review might yield greater benefits, given individual differences such as those noted in the previous section of this chapter.

We developed software that was integrated into middle-school Spanish foreign language courses to guide students in the systematic review of course material. We conducted two semester-long experiments with this software. In each experiment, we compared several alternative strategies for selecting material to review.

[2] In contrast to the present results, Ridgeway and colleagues found no improvement with the HYBRID BOTH over the HYBRID SCALE model.

Our goal was to evaluate a big-data strategy for personalized review that infers the dynamic knowledge state of each student as the course progressed, taking advantage of both population data and psychological theory. Just as we leveraged theories of forgetting to model retention following a single study session, we leverage theories of spaced practice—in particular, the two models we described earlier, ACT-R and MCM—to model retention following a complex multi-episode study schedule.

6.1 Representing Study History

Before turning to our experiments, we extend our approach to modeling knowledge state. Previously, we were concerned with modeling forgetting after a student was exposed to material one time. Consequently, we were able to make the strong assumption that all student-items have an identical study history. To model knowledge state in a more naturalistic setting, we must relax this assumption and allow for an arbitrary *study history*, defined as zero or more previous exposures at particular points in time.

Extending our modeling approach, we posit that knowledge state is 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).

DASH predicts the likelihood of student s making a correct response on the kth trial for item i, conditioned on that student-item's specific study history:

$$P(R_{sik} = 1 \mid a_s, d_i, \mathbf{t}_{1:k}, \mathbf{r}_{1:k-1}, \boldsymbol{\theta}) = \sigma(a_s - d_i + h_{\boldsymbol{\theta}}(\mathbf{t}_{s,i,1:k}, \mathbf{r}_{s,i,1:k-1})), \tag{6}$$

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 function that summarizes the effect of study history on recall probability, and $\boldsymbol{\theta}$ is a parameter vector that governs $h_{\boldsymbol{\theta}}$. As before, a_s and d_i denote the latent ability of student s and difficulty of item i, respectively. This framework is an extension of additive-factors models used in educational data mining (Cen, Koedinger, & Junker, 2006, 2008; Pavlik, Cen, & Koedinger, 2009).

DASH draws on key insights from the psychological models MCM and ACT-R via a representation of study history that is based on log counts of practice and success with an item over multiple expanding windows of time, formalized as:

$$h_{\boldsymbol{\theta}} = \sum_{w=1}^{W} \theta_{2w-1} \log(1 + c_{siw}) + \theta_{2w} \log(1 + n_{siw}) \tag{7}$$

where $w \in \{1, ..., W\}$ is an index over time windows, c_{siw} is the number of times student s correctly recalled item i in window w out of n_{siw} attempts, and $\boldsymbol{\theta}$ are window-specific weightings. Motivated by the multiple traces of MCM, we include statistics of study history that span increasing windows of time. These windows allow the model to modulate its predictions based on the temporal distribution of study. Motivated by the diminishing benefit of additional study in ACT-R (Equation 1), we include a similar log transform in Equation 7.[3] Both MCM and ACT-R modulate the effect of past study based on response outcomes, i.e., whether the student performed correctly or not on a given trial. This property is incorporated into Equation 7 via the separation of parameters for counts of total and correct attempts.

Being concerned that the memory dynamics of MCM and ACT-R provided only loose inspiration to DASH, we designed two additional variants of DASH that more strictly adopted the dynamics of MCM and ACT-R. The variant we call DASH[MCM] replaces expanding time windows with expanding time constants which determine the rate of exponential decay of memory traces. The model assumes that the counts n_{siw} and c_{siw} are incremented at each trial and then decay over time at a timescale specific exponential rate τ_w. Formally, we use Equation 7 with the counts redefined as:

$$n_{siw} = \sum_{\kappa=1}^{k-1} e^{-(t_{sik} - t_{si\kappa})/\tau_w} \qquad c_{siw} = \sum_{\kappa=1}^{k-1} r_{si\kappa} e^{-(t_{sik} - t_{si\kappa})/\tau_w} \tag{8}$$

The variant we call DASH[ACT-R] 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. DASH[ACT-R] formalizes the effect of study history to be identical to the memory trace strength of ACT-R (Equation 1):

$$h_{\boldsymbol{\theta}} = \theta_1 \log(1 + \sum_{\kappa=1}^{k-1} \theta_{3+r_{si\kappa}} (t_{sik} - t_{si\kappa})^{-\theta_2}) \tag{9}$$

Further details of the modeling and a hierarchical Bayesian scheme for inferring model parameters is given in Lindsey (2014).

6.2 Classroom Studies of Personalized Review

We incorporated systematic, temporally distributed review into Spanish foreign language instruction at a Denver area middle school using an electronic flaschard tutoring system. Each week of the semester, students engaged during class in three 20–30 minute sessions with the system, called COLT. COLT presented vocabulary words and short sentences in English and required students to type the Spanish translation, after

[3]The counts c_{siw} and n_{siw} are regularized by add-one smoothing, which ensures that the logarithm terms are finite.

which corrective feedback was provided. The first two sessions of each week began with a study-to-proficiency phase for new material that was introduced in that week's lesson, and then proceeded to a phase during which previously introduced material was reviewed. In the third session, these activities were preceded by a quiz on the current lesson, which counted toward the course grade.

We conducted two semester-long experiments with COLT, the first of which is described in detail in Lindsey, Shroyer, Pashler, and Mozer (2014) and the second of which appears only in the Ph.D. thesis of Lindsey (2014). We summarize the two experiments here.

Experiment 1

Experiment 1 involved 179 third semester Spanish students, split over six class periods. The semester covered 10 lessons of material. COLT incorporated three different *schedulers* to select material from these lessons for review. The goal of each scheduler was to make selections that maximize long-term knowledge preservation given the limited time available for review. The scheduler was varied within participant by randomly assigning one third of a lesson's items to each scheduler, 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 and that all schedulers administered an equal number of review trials.

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

A *generic-spaced* scheduler selected one previous lesson 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, Pashler, Vul, Wixted, & Rohrer, 2006; Cepeda et al., 2008) and computational models (Khajah, Lindsey, & Mozer, 2013; Mozer et al., 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 lesson, giving priority to the least recently studied.

A *personalized-spaced* scheduler used our knowledge-state model, DASH, to determine the specific item

a particular student would most benefit from reviewing. DASH infers the instantaneous memory strength of each item the student has studied. Although a knowledge-state model is required to schedule review optimally, optimal scheduling is computationally intractable because it requires planning over all possible futures (when and how much a student studies, including learning that takes place outside the context of COLT, and within the context of COLT, whether or not retrieval attempts are successful, etc.). Consequently, a heuristic policy is required for selecting review material. We chose a threshold-based policy that prioritizes items whose recall probability is closest to a threshold θ. This heuristic policy is justified by simulation studies as being close to optimal under a variety of circumstances (Khajah et al., 2013) and by Bjork's (1994) notion of *desirable difficulty*, which suggests that memory is best served by reviewing material as it is on the verge of being forgotten.

As the semester progressed, COLT continually collected data and DASH was retrained with the complete data set at regular intervals. The retraining was sufficiently quick and automatic that the model could use data from students in the first class period of the day to improve predictions for students in the second class period. This updating was particularly useful when new material was introduced and DASH needed to estimate item difficulty. By the semester's end, COLT had amassed data from about 600,000 retrieval-practice trials.

To assess student retention, two proctored cumulative exams were administered, 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 6a). 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 ($t(175) = 6.59$, $p < .0001$, $d = 0.88$). Note that "massed" review is spaced by usual laboratory standards, being spread out over at least seven days. This fact may explain the small benefit of generic spaced over massed.

In Lindsey et al. (2014), we showed that personalized review has its greatest effect on the early lessons of the semester, which is sensible because that material had the most opportunity for being manipulated via review. We also analyzed parameters of DASH to show that its predictions depend roughly in equal part on

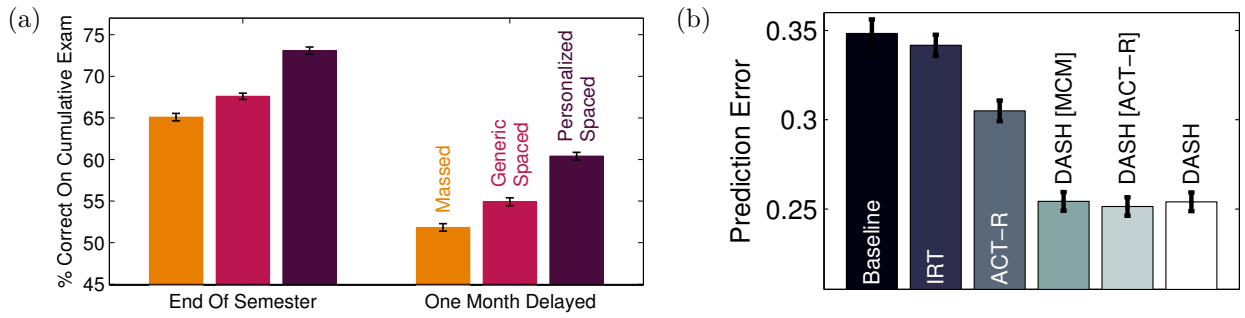

Figure 6: COLT Experiment 1. (a) Mean scores on the two cumulative end-of-semester exams, taken 28 days apart. All error bars indicate ±1 within-student standard error (Masson & Loftus, 2003). (b) Accumulative prediction error of six models using the data from the experiment. The models are as follows: 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 two variants of DASH that adhere more strictly to the tenets of MCM and ACT-R. Error bars indicate ±1 SEM.

student abilities, item difficulties, and study history.

To evaluate the quality of DASH's predictions, we compared DASH against alternative models by dividing the 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 likelihood and the actual binary outcome, normalized such that each student is weighted equally. Figure 6b 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 IRT, Pavlik and Anderson's (2005b) ACT-R model of spacing effects, and the two variants of DASH we described earlier that incorporate alternative representations of study history motivated by models of spacing effects. DASH and its two variants perform better than the alternatives. The DASH models each have two key components: (1) a dynamical representation of study history that can characterize learning and forgetting, and (2) a collaborative filtering 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.

Experiment 2

Experiment 1 took place in the fall semester with third-semester Spanish students. We conducted a follow-up experiment in the next (spring) semester with the same students, then in their fourth semester of Spanish. (One student of the 179 in Experiment 1 did not participate in Experiment 2 because of a transfer.) The semester was organized around 8 lessons, followed by two cumulative exams administered 28 days apart. The two cumulative exams each tested half the course material, with a randomized split by student.

The key motivations for Experiment 2 are as follows.

- In Experiment 1, the personalized-review scheduler differed from the other two schedulers both in its personalization and in its ability to select material from early in the semester. Because personalized review and long-term review were conflated, we wished to include a condition in Experiment 2 that involved long-term review but without personalization. We thus incorporated a *random* scheduler that drew items uniformly from the set of items that had been introduced in the course to date. Because the massed scheduler of Experiment 1 performed so poorly, we replaced it with the random scheduler.

- Because the same students participated in Experiments 1 and 2, we had the opportunity to initialize students models based on all the data from Experiment 1. The old data provided DASH with fairly strong evidence from the beginning of the semester about individual student abilities and about the relationship of study schedule to retention. Given that Experiment 2 covered only 8 lessons, versus the 10 in Experiment 1, this bootstrapping helped DASH to perform well out of the gate.

- Using the data from Experiment 1, DASH[ACT-R] obtains a slighly lower accumulative prediction error than DASH (Figure 6b). Consequently, we substituted DASH[ACT-R] as the model used to select items for review in the personalized condition.

Figure 7 summarizes the experiment outcome. The bars represent scores in the three review conditions on the initial and delayed exams. The differences among conditions are not as stark as we observed in Experiment 1, in part because we eliminated the weak massed condition and in part due to an unanticipated issue which we address shortly. Nonetheless, on the first exam, the personalized-spaced scheduler improved retention by 4.8% over the generic-spaced scheduler ($t(167) = 3.04$, $p < .01$, Cohen's $d = 0.23$) and by 3.4% over the random scheduler ($t(167) = 2.29$, $p = .02$, $d = 0.18$). Between the two exams, the forgetting rate is roughly the same in all conditions: 16.7%, 16.5%, and 16.5% for the generic, random, and personalized conditions, respectively. On the second exam, personalized review boosted retention by 4.6% over generic

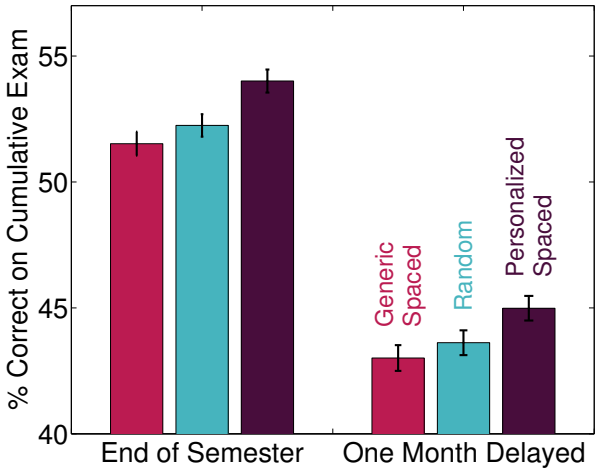

Figure 7: COLT Experiment 2. Mean scores on the cumulative end-of-semester exam. All error bars indicate ± 1 within-student standard error (Masson & Loftus, 2003).

review ($t(166) = 2.27$, $p = .024$, $d = 0.18$) and by 3.1% over random review, although this difference was not statistically reliable ($t(166) = 1.64$, $p = .10$, $d = 0.13$).

At about the time we obtained these results, we discovered a significant problem with the experimental software. Students did not like to review. In fact, at the end of Experiment 1, an informal survey indicated concern among students that mandatory review interfered with their weekly quiz performance because they were not able to spend all their time practicing the new lesson that was the subject of their weekly quiz. Students *wished* to mass their study due to the incentive structure of the course, and they requested a means of opting out of review. We did not accede to their request; instead, the teacher explained the value of review to long-term retention. Nonetheless, devious students found a way to avoid review: upon logging in, COLT began each session with material from the new lesson. Students realized that if they regularly closed and reopened their browser windows, they could avoid review. Word spread throughout the student population, and most students took advantage of this unintended feature of COLT. The total number of review trials performed in Experiment 2 was a small fraction of the number of review trials in Experiment 1. Consequently, our failure to find large and reliable differences among the schedulers is mostly due to the fact that students simply did not review.

One solution might be to analyze the data from only those students who engaged in a significant number of review trials during the semester. We opted instead to use data from all students and to examine the relative benefit of the different review schedulers as a function of the amount of review performed. The amount of review is quantified as the total number of review trials performed by a student divided by the

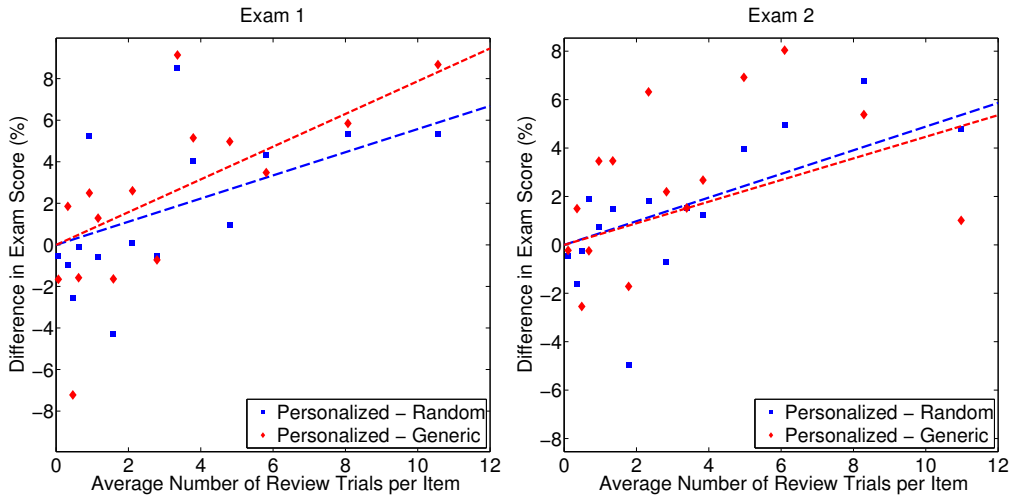

Figure 8: COLT Experiment 2. Scatterplot for exams 1 and 2 (left and right, respectively) showing the advantage of personalized-spaced review over random and generic-spaced review, as a function of the amount of review that a student performed. The amount of review is summarized in terms of the the total number of review trials during the semester divided by the number of items.

total number of items, i.e., the mean number of review trials. Note, however, that this statistic does not imply that each item was reviewed the same number of times. For each student, we computed the difference of exam scores between personalized and generic conditions, and between personalized and random conditions. We performed a regression on these two measures given the amount of review. Figure 8 shows the regression curves that represent the exam score differences as a function of mean review trials per item. The regressions were constrained to have an intercept at 0.0 because the conditions are identical when no review is included. The data points plotted in Figure 8 are averages based on groups of about 10 students who performed similar amounts of review. These groupings make it easier to interpret the scatterplot, but the raw data were used for the regression analysis.

Figure 8 shows a positive slope for all four regression lines (all reliable by t tests with $p < 0.01$), indicating that with more time devoted to review, the personalized-review scheduler increasingly outperforms the random and generic-review schedulers. If, for example, students had studied on COLT for an average of one more review trial per item for each of the 13 weeks in the semester leading up to exam 1, Figure 8 predicts an (absolute) improvement on exam 1 scores of 10.2% with personalized-spaced review over generic-spaced review and 7.2% with personalized-spaced review over random review. We wish to emphasize that we are not simply describing an advantage of review over no review. Our result suggests that students will score a letter grade higher (7-10 points out of 100) with time-matched personalized review over the other forms of

review.

In contrast to Experiment 1, the effect of personalized review was not amplified on exam 2 relative to exam 1. Our explanation is that in Experiment 1, the 28-day retention interval between exams 1 and 2 was the holiday break, a time during which students were unlikely to have much contact with course material. In Experiment 2, the intervening 28-day period occurred during the semester when students were still in class but spent their time on enrichment activities that were not part of our experiment (e.g., class projects, a trip to the zoo). Consequently, students had significant exposure to course content, and this exposure could only have served to inject noise in our assessment exam.

6.3 Discussion

Whereas previous studies offer in-principle evidence that human learning can be improved by the inclusion and 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. Our experiments go beyond showing that spaced practice is superior to massed practice: taken together, Experiments 1 and 2 provide strong evidence that personalization of review is superior to other forms of spaced practice.

Although the outcome of Experiment 2 was less impressive than the outcome of Experiment 1, the mere fact that students went out of their way to avoid a review activity that would promote long-term retention indicates the great need for encouraging review of previously learned material. One can hardly fault the students for wishing to avoid an activity they intuited to be detrimental to their grades. The solution is to better align the student's goals with the goal of long-term learning. One method of alignment is to administer only cumulative quizzes. In principle, there's no reason to distinguish the quizzes from the retrieval practice that students perform using COLT, achieving the sort of integration of testing and learning that educators often seek.

7 Conclusions

Theory-driven approaches in psychology and cognitive science excel at chracterizing the laws and mechanisms of human cognition. Data-driven approaches from machine learning excel at inferring statistical regularities that describe how individual vary within a population. In this chapter, we've argued that in the domain of learning and memory, a synthesis of theory- and data-driven approaches inherits the strengths of each. Theory-driven approaches characterize the temporal dynamics of learning and forgetting based on study history and past performance. Data-driven approaches use data from a *population* of students learning a *collection* of items to make inferences concerning the knowledge state of *individual* students for *specific* items.

The models described in this chapter offer more than qualitative guidance to students about how to study. In one respect, they go beyond what even a skilled classroom teacher can offer: they are able to keep track of student knowledge state at a granularity that is impossible for a teacher who encounters hundreds of students over the course of a day. A system such as COLT provides an efficient housekeeping function to ensure that knowledge, once mastered, remains accessible and a part of each student's core competency. COLT allows educators to do what they do best: to motivate and encourage; to help students to acquire facts, concepts, and skills; and to offer creative tutoring to those who face difficulty. To achieve this sort of complementarity between electronic tools and educators, a big data approach is essential.

Appendix: Simulation Methodology For Hybrid Forgetting Model

Each of the five forgetting models was cast in a hierarchical Bayesian generative framework, as specificed in Table 2. We employed Markov chain Monte Carlo to draw samples from the posterior, specifically Metropolis-within-Gibbs (Patz & Junker, 1999), an extension of Gibbs sampling wherein each draw from the model's full conditional distribution is performed by a single Metropolis-Hastings step.

Inference on the two sets of latent traits in the HYBRID BOTH model—$\{a_s\}$ and $\{d_i\}$ from IRT, $\{\tilde{a}_s\}$ and $\{\tilde{d}_i\}$ from HYBRID DECAY—is done jointly, leading to possibly a different outcome than the one that we would obtain by first fitting IRT and then inferring the decay-rate determining parameters. In essence, the HYBRID BOTH model allows the corrupting influence of time to be removed from the IRT variables, and allows the corrupting influence of static factors to be removed from the forgetting-related variables.

The hierarchical Bayesian models impose weak priors on the parameters. Each model assumes that latent traits are normally distributed with mean zero and an unknown precision parameter shared across the population of items or students. The precision parameters are all given Gamma priors. Through Normal-

Gamma conjugacy, we can analytically marginalize them before sampling. Each latent trait's conditional distribution thus has the form of a likelihood (defined in Equations 2–5) multiplied by the probability density function of a non-standardized Student's t-distribution. For example, the ability parameter in the HYBRID SCALE model is drawn via a Metropolis-Hastings step from the distribution

$$p(a_s \mid \mathbf{a}_{\neg \mathbf{s}}, \mathbf{d}, h, m, R) \propto \prod_i P(r_{si} \mid a_s, d_i, h, m) \times \left(1 + \frac{a_s^2}{2(\psi_2 + \frac{1}{2}\sum_{j \neq s} a_j)}\right)^{\psi_1 + \frac{n_S - 1}{2}} \tag{10}$$

where the first term is given by Equations 3 and 5. The effect of the marginalization of the precision parameters is to tie the traits of different students together so that they are no longer conditionally independent.

Hyperparameters ψ of the Bayesian models were set so that all the Gamma distributions had shape parameter 1 and scale parameter .1. For each run of each model, we combined predictions from across three Markov chains, each with a random starting location. Each chain was run for a burn in of 1,000 iterations and then 2,000 more iterations were recorded. To reduce autocorrelation among the samples, we thinned them by keeping every tenth one.

Why did we choose to fit models with hierarchical Bayesian (HB) inference instead of the more standard maximum likelihood (ML) estimation? The difference between HB and ML is that HB imposes an additional bias that, in the absence of strong evidence about a parameter value—say, a student's ability or an item's difficulty—the parameter should be typical of those for other students or other items. ML does not incorporate this prior belief, and as a result, it is more susceptible to overfitting a training set. For this reason, we were not surprised when we tried training models with ML and found they did not perform as well as with HB.

Acknowledgments

The research was supported by NSF grants SBE-0542013, SMA-1041755, and SES-1461535 and an NSF Graduate Research Fellowship to R. L. We thank Jeff Shroyer for his support in conducting the classroom studies, and Melody Wisehart and Harold Pashler for providing raw data from their published work and for their generous guidance in interpreting the spacing literature.

IRT	HYBRID DECAY	HYBRID SCALE
$r_{si} \mid a_s, d_i \sim \text{Bernoulli}(p_{si})$	$r_{si} \mid \tilde{a}_s, \tilde{d}_i, m, h, t_{si} \sim \text{Bernoulli}(m\tilde{p}_{si})$	$r_{si} \mid a_s, d_i, \tilde{a}_s, \tilde{d}_i, h, t_{si} \sim \text{Bernoulli}(p_{si}\tilde{p}_{si})$
$p_{si} = (1 + \exp(d_i - a_s))^{-1}$	$\tilde{p}_{si} = (1 + ht_{si})^{-\exp(\tilde{a}_s - \tilde{d}_i)}$	$\tilde{p}_{si} = (1 + ht_{si})^{-f}$
$a_s \mid \tau_a \sim \text{Normal}(0, \tau_a^{-1})$	$\tilde{a}_s \mid \tau_{\tilde{a}} \sim \text{Normal}(0, \tau_{\tilde{a}}^{-1})$	$f \sim \text{Gamma}(\psi_{f1}, \psi_{f2})$
$d_i \mid \tau_d \sim \text{Normal}(0, \tau_d^{-1})$	$\tilde{d}_i \mid \tau_{\tilde{d}} \sim \text{Normal}(0, \tau_{\tilde{d}}^{-1})$	
$\tau_a \sim \text{Gamma}(\psi_{a1}, \psi_{a2})$	$\tau_{\tilde{a}} \sim \text{Gamma}(\psi_{\tilde{a}1}, \psi_{\tilde{a}2})$	
$\tau_d \sim \text{Gamma}(\psi_{d1}, \psi_{d2})$	$\tau_{\tilde{d}} \sim \text{Gamma}(\psi_{\tilde{d}1}, \psi_{\tilde{d}2})$	All other parameters are same as IRT and HYBRID DECAY
	$h \sim \text{Gamma}(\psi_{h1}, \psi_{h2})$	
	$m \sim \text{Beta}(\psi_{m1}, \psi_{m2})$	

Table 2: Distributional assumptions of the generative Bayesian response models. The HYBRID BOTH model shares the same distributional assumptions as the HYBRID DECAY and HYBRID SCALE models.

References

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

Anderson, J. R., Conrad, F. G., & Corbett, A. T. (1989). Skill acquisition and the LISP tutor. *Cognitive Science*, *13*, 467–506.

Andrade, D. F., & Tavares, H. R. (2005). Item response theory for longitudinal data: population parameter estimation. *Journal of Multivariate Analysis*, *95*, 1–22.

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

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). Cambridge, MA: MIT Press.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review*, *24*, 369–378.

Carpenter, S. K., 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. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Journal of Experimental Psychology*,

56, 236–246.

Cepeda, N. J., Pashler, H., Vul, E., & Wixted, J. T. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin & Review*, *132*, 364–380.

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

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

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.

DeBoek, 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.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Grimaldi, P. J., Pyc, M. A., & Rawson, K. A. (2010). Normative multitrial recall performance, metacognitive judgments and retrieval latencies for Lithuanian-English paired associates. *Behavioral Research Methods*, *42*, 634–642.

Kang, S. H. K., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). *Retrieval practice over the long term: Expanding or equal-interval spacing?* (Vol. 21).

Khajah, M., Lindsey, R. V., & Mozer, M. C. (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.*

Koedinger, K. R., & Corbett, A. T. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.), *The cambridge handbook of the learning sciences* (pp. 61–78). Cambridge UK: Cambridge University Press.

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

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *IEEE Computer*, *42*, 42–49.

Lan, A. S., Studer, C., & Baraniuk, R. G. (2014). Time-varying learning and content analytics via sparse factor analysis. In *ACM SIGKDD conf. on knowledge disc. and data mining.* Retrieved from `http://arxiv.org/abs/1312.5734`

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

Lindsey, R. V. (2014). *Probabilistic models of student learning and forgetting* (Unpublished doctoral dissertation). Computer Science Department, University of Colorado at Boulder.

Lindsey, R. V., Lewis, O., Pashler, H., & Mozer, M. C. (2010). Predicting students' retention of facts from feedback during training. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 2332–2337). Austin, TX: Cognitive Science Society.

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.

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

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

Metcalfe, J., & Finn, B. (2011). People's hypercorrection of high confidence errors: Did they know it all along? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 437–448.

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

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

Mozer, M. C., Pashler, H., Cepeda, N., Lindsey, R. V., & 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).

Mullany, A. (2013). *A Q&A with Salman Khan.* Retrieved 2014-12-23, from `http://live.fastcompany.com/Event/A_QA_With_Salman_Khan`

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146–178.

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

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

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

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

Pavlik, P. I., 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.

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

Rickard, T., Lau, J., & Pashler, H. (2008). Spacing and the transition from calculation to retrieval. *Psychonomic Bulletin & Review*, *15*, 656–661.

Ridgeway, K., Mozer, M. C., & Bowles, A. (2016). Forgetting of foreign-language skills: A corpus-based analysis of online tutoring software. *Cognitive Science Journal*. (Accepted for publication)

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology*, *20*, 1209–1224.

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement*, *44*, 293–311.

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.

Sohl-Dickstein, J. (2013). *Personalized learning and temporal modeling at Khan Academy.* Retrieved from `http://lytics.stanford.edu/datadriveneducation/slides/sohldickstein.pdf`

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

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).

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.

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, *55*, 235-269.

Wixted, J. T., & Carpenter, S. K. (2007). The Wickelgren power law and the Ebbinghaus savings function. *Psychological Science*, *18*, 133–134.

Woźniak, P. (1990). *Optimization of learning* (Unpublished master's thesis). Poznan University of Technology, Poznan, Poland.