Probabilistic Models of Student Learning and Forgetting

by

Robert Lindsey

B.S., Rensselaer Polytechnic Institute, 2008

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

2014

This thesis entitled:
Probabilistic Models of Student Learning and Forgetting
written by Robert Lindsey
has been approved for the Department of Computer Science

Michael Mozer

Aaron Clauset

Vanja Dukic

Matt Jones

Sriram Sankaranarayanan

Date

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

IRB protocol #0110.9, 11-0596, 12-0661

Lindsey, Robert (Ph.D., Computer Science)

Probabilistic Models of Student Learning and Forgetting

Thesis directed by Prof. Michael Mozer

This thesis uses statistical machine learning techniques to construct predictive models of human learning and to improve human learning by discovering optimal teaching methodologies. In Chapters 2 and 3, I present and evaluate models for predicting the changing memory strength of material being studied over time. The models combine a psychological theory of memory with Bayesian methods for inferring individual differences. In Chapter 4, I develop methods for delivering efficient, systematic, personalized review using the statistical models. Results are presented from three large semester-long experiments with middle school students which demonstrate how this "big data" approach to education yields substantial gains in the long-term retention of course material. In Chapter 5, I focus on optimizing various aspects of instruction for populations of students. This involves a novel experimental paradigm which combines Bayesian nonparametric modeling techniques and probabilistic generative models of student performance. In Chapters 6 and 7, I present supporting laboratory behavioral studies and theoretical analyses. These include an examination of the relationship between study format and the testing effect, and a parsimonious theoretical account of long-term recency effects.

Acknowledgements

I would like to thank Mike Mozer, Jeff Shroyer, and Cathie Knutson for their help. I am also indebted to Hal Pashler and Sean Kang for their advice, and to my parents for their support.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Extended Summary

The purpose of this thesis is to develop software tools that improve human learning and performance. The approach we take is to translate qualitative theories of human learning and memory—those which cognitive scientists have developed over the past 150 years—into quantitative "big data" approaches that measure, predict, and optimize human performance.

The software we developed delivers personalized review by deciding what study items each student should review next at any given time based on estimates of memory strength from a hierarchical Bayesian model. A challenge in modeling memory strength over time is the presence of enormous uncertainty: uncertainty in the students' abilities, the items' difficulties, and the rate of forgetting, among other factors. Chapter 2 provides background information on approaches to modeling memory strength over time in the presence of uncertainty. Chapter 3 describes two novel models we created that leverage the large volume of data the software collects across the entire population of students and study items to make robust predictions about the memory strength of individual students on individual study items. The models operate using the same principles that online commerce websites use to deliver product recommendations to a customer based on the habits of other customers: even though the model may not have a sufficient number of observations to leverage to make a highly constrained recommendation for the individual, it can reduce its uncertainty by leveraging observations of similar customers or products. This *collaborative filtering* approach to estimating memory strength over time allows the software to use optimization techniques to select the material for review that it predicts would be most beneficial, despite the

presence of so much uncertainty.

Chapter 4 focuses on improving educational outcomes by tailoring instruction to the needs of individual students. In typical classroom settings, from grade school through graduate school, students traditionally learn material in blocks. They move from one block to the next, taking a test at the end of each block. Understandably, students are most interested in studying for their tests, but this focus has regrettable consequences for long-term learning. Psychologists well appreciate that all knowledge—whether facts, concepts, or skills—is eventually forgotten and that reviewing old material is necessary to mitigate forgetting. This chapter explores improving classroom education through software tools that provide a form of systematic review which reduces forgetting. We focus on integrating spaced, *personalized* review—temporally distributed practice and testing tailored to individuals based on their study and performance history—into real-world classrooms. The three longitudinal experiments we describe took place over three semesters and involved nearly 500 Denver-area middle school students. In the first experiment, we developed software used by 179 Spanish-language students to practice translation of words and phrases for 30 minutes a week across a semester. Incorporating personalized review yielded a jump in post-semester retention of 16.5% over (time matched) current educational practice and 10% over generic, one-size-fits-all spaced review, despite the fact that the experimental manipulation represents only a small fraction of the time the students were engaged with the course material. These experiments demonstrate that integrating adaptive, personalized software into the classroom is practical and yields appreciable improvements in long-term educational outcomes.

Chapter 5 focuses on improving learning through the optimization of training strategies across populations of learners. Psychologists and educators are interested in developing *instructional policies*—strategies which specify the manner and content of instruction—that boost student learning. For example, in concept learning, a strategy may determine the nature of exemplars chosen across a training sequence. Traditional psychological studies compare several hand-selected strategies: for example, contrasting a strategy that selects only difficult-to-classify exemplars with a strategy that gradually progresses over the training sequence from easy to more difficult exemplars (a manipula-

tion known as *fading*). Proposing an alternative to the traditional experimental methodology, we define a metric space of strategies and iteratively search the space to identify the globally optimal strategy. For example, in concept learning, strategies might be described by a fading function that specifies exemplar difficulty over time. Our method for searching strategy spaces uses nonparametric Bayesian regression techniques and a probabilistic model of student performance. Instead of evaluating a few experimental conditions, each with many students, as the traditional methodology does, this method evaluates many experimental conditions one at a time, each with one or a few students. Even though individual students provide only a noisy estimate of the population mean, the optimization method can determine the shape of the strategy space and efficiently identify the optima. We evaluate the method's applicability to optimizing student learning through two behavioral studies, and we also explore the method's applicability to other "human optimization" settings, including human vision and decision-making.

Chapter 6 describes three behavioral experiments evaluating the effectiveness of different study formats. Retrieval practice study—study which involves both quizzing and reviewing—results in stronger and more durable memories than reviewing alone (H. Roediger & Karpicke, 2006a). However, incorporating quizzing into electronic tutoring systems is infeasible for many common types of study materials. These experiments investigate whether students can reap the benefits of retrieval practice study if they merely retrieve the material from memory without an overt behavioral response.

Chapter 7 provides a theoretical analysis of a phenomenon known as *long term recency*. When tested on a list of items, individuals show a recency effect: the more recently a list item was presented, the more likely it is to be recalled. For short interpresentation intervals (IPIs) and retention intervals (RIs), this effect may be attributable to working memory. However, recency effects also occur over long timescales where IPIs and RIs stretch into the weeks and months. These long-term recency effects have intrigued researchers because of their scale-invariant properties and the sense that understanding the mechanisms of LTR will provide insights into the fundamental nature of memory. An early explanation of LTR posited that it is a consequence of memory trace

decay, but this *decay hypothesis* was discarded in part because LTR was not observed in continuous distractor recognition memory tasks (Glenberg & Kraus, 1981; Bjork & Whitten, 1974; Poltrock & MacLeod, 1977). Since then, a diverse collection of elaborate mechanistic accounts of LTR have been proposed. In Chapter 7, we revive the decay hypothesis. Based on the uncontroversial assumption that forgetting occurs according to a power-law function of time, we argue that not only is the decay hypothesis a sufficient qualitative explanation of LTR, but also that it yields excellent quantitative predictions of LTR strength as a function of list size, test type, IPI, and RI. Through fits to a simple model, this chapter aims to bring resolution to the subject of LTR by arguing that LTR is nothing more than ordinary forgetting.

Chapter 2

Modeling background

A student's *knowledge state*—his or her degree of mastery over specific concepts, skills, or facts—fluctuates across time as a result of studying and forgetting. Reliably inferring the current and anticipated future knowledge states of students is necessary in order to tailor effectively instruction to the needs of individuals. However, this inference problem is challenging because behavioral observations are only weakly informative of a student's underlying dynamic knowledge state. For example, suppose that a student solved four out of five specific long-division problems correctly on a quiz. How well would you expect the student to do on a particular long-division problem assigned a month later? This review chapter surveys two contrasting modeling approaches taken in the literature to overcome this problem. We will refer to these approaches as **data driven** and **theory driven**. Data-driven approaches leverage the often large quantity of observations collected across a population of students to make strong predictions about an individual student, while theory-driven approaches make strong predictions by relying on results from psychological research in long-term memory to constrain the temporal and practice-dependent dynamics of knowledge states.

2.1 Knowledge states and forgetting

A student's *knowledge state*—his or her degree of mastery over specific concepts, skills, or facts—fluctuates across time as a result of studying and forgetting. Inferring the current and anticipated future knowledge states of students is a central concern in diverse areas such as educational assessment, intelligent tutoring systems, and psychological research in long-term memory. The

inference problem has immense practical relevance as well: hundreds of thousands of students each year use automated tutoring systems which make decisions about what instructional interventions or study materials to deliver by using quantitative predictions from statistical models of knowledge states over time (Desmarais & Baker, 2012). Models which make reliable predictions are necessary in order to effectively tailor the decisions of such systems to the needs of individual students, and evidence suggests that better models result in a higher quality of education received by this growing segment of the world's student population (Lee & Brunskill, 2012; Cen, Koedinger, & Junker, 2007)

Longitudinal estimation of a student's knowledge state is a challenging problem because available behavioral observations are only weakly informative of the underlying state. The canonical tutoring system presents material to students in a series of cued retrieval practice trials where in each trial a student is given a cue (e.g., a vocabulary word, an algebra problem, etc.) chosen by the system, attempts to produce the correct response (e.g., the word's definition, the solution to the problem, etc.), and then receives corrective feedback. Tutoring systems often only observe from this process whether or not the student produced the correct response in each trial. Dichotomous observations such as response accuracy convey just a single bit of information at a single instant in time about a student's variable mastery over the material.

The knowledge-state estimation problem is difficult also because memory strength is sensitive to its **study history**: when in the past the specific material was studied, as well as the duration and manner of past study. History is particularly relevant because all forms of learning show forgetting over time, and retention is fragile when the material being learned is unfamiliar (Rohrer & Taylor, 2006; Wixted, 2004b). The temporal distribution of practice has an impact on the durability of learning for various types of material (Cepeda et al., 2006; Rickard, Lau, & Pashler, 2008); even relatively minor changes to the time between successive study trials can reliably double retention on a later test (Cepeda et al., 2008). Furthermore, each time a tutoring system administers a retrieval practice trial, it alters the knowledge state, even if no feedback is provided to students (H. Roediger & Karpicke, 2006b). Thus, each time a tutoring system tries to measure the knowledge state, it alters the knowledge state.

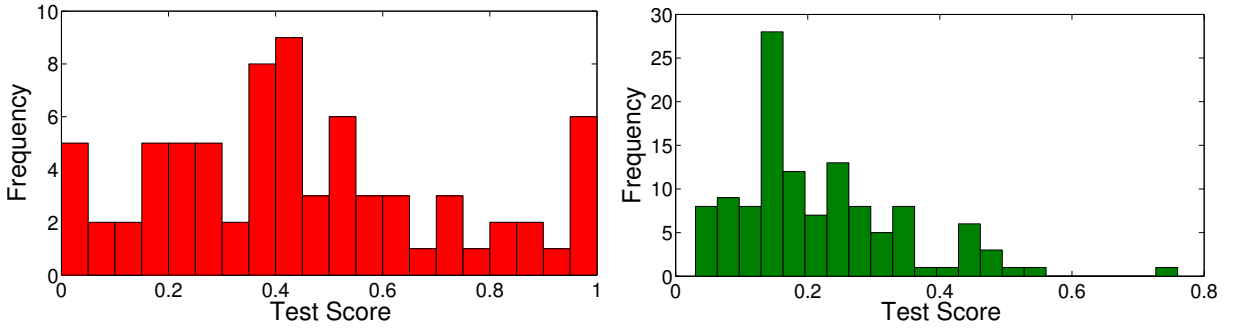


Figure 2.1: (left) Histogram of proportion of items reported correctly on a cued recall task for a population of 60 students learning 32 Japanese-English vocabulary pairs (S. H. K. Kang et al., 2014); (right) Histogram of proportion of subjects correctly reporting an item on a cued recall task for a population of 120 Lithuanian-English vocabulary pairs being learned by roughly 80 students (Grimaldi et al., 2010)

Further complicating the prediction problem is the ubiquity of individual differences in every form of learning. Taking an example from fact learning, Figure 2.1a shows extreme variability in a population of 60 students. These students studied foreign-language vocabulary at four precisely scheduled times over a four-week period. A cued-recall exam was administered after an eight-week retention period and the exam scores were highly dispersed despite the uniformity in materials and training schedules. In addition to inter-student variability, variability between study items is a consideration. For example, 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 2.1b shows the distribution of recall accuracy for 120 Lithuanian-English vocabulary items averaged over a set of students. With a single round of study, an exam administered several minutes later suggests that items show a tremendous range in difficulty (e.g., **krantas→shore** was learned by only 3% of students; **lova→bed** was learned by 76% of students).

This chapter gives an overview of modeling techniques for longitudinal knowledge-state estimation which each address a subset of the aforementioned challenges. The techniques we will discuss gain traction by imposing constraints on the otherwise under-constrained prediction problem. These constraints are generally either *data driven* or *theory driven*. Data-driven modeling

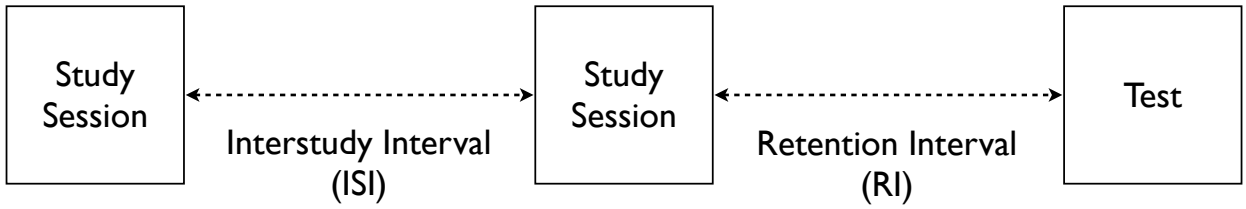


Figure 2.2: Typical spacing experiments have one or more study sessions separated by *interstudy intervals* (ISIs) with a final test administered after a fixed retention interval. Student performance on the test is sensitive to the ISIs and RI.

approaches leverage the often large quantity of observations collected across populations of students and study materials to make strong predictions for the individual. While the knowledge state of an individual student studying an individual item may be too hard to predict in isolation, the model can inform its prediction by looking at how other students have done on the item and at how the student has done on other items. Data-driven approaches to knowledge estimation are often similar to the techniques used in e-commerce to deliver personalized product recommendations based on viewing or purchase history.

In contrast, theory-driven approaches use results from psychological research in long-term memory to constrain the study-history dependent properties of models in order to make strong predictions. Psychologists have long studied the temporal characteristics of learning and memory. The modern consensus is that when a set of materials is learned in a single study session and then tested following some time lag Δt, the probability of recalling the studied material decays according to a generalized power-law function,

$$\Pr(R \ = \ 1) = x_0(1 + h \ \Delta t)^{-d}, \tag{2.1}$$

where $R \in \{0, 1\}$ represents recall success or failure, $0 \leq x_0 \leq 1$ is the degree of learning, $h > 0$ is a scaling factor on time, and $d > 0$ is the memory decay exponent (Wixted & Carpenter, 2007).

Research on the time-course of forgetting following multiple study sessions dates back to the 19th century (Ebbinghaus, 1885 / 1964; Jost, 1897). The ubiquitous *spacing effect*—the finding that temporally spaced study yields enhanced learning as compared to temporally massed study—is

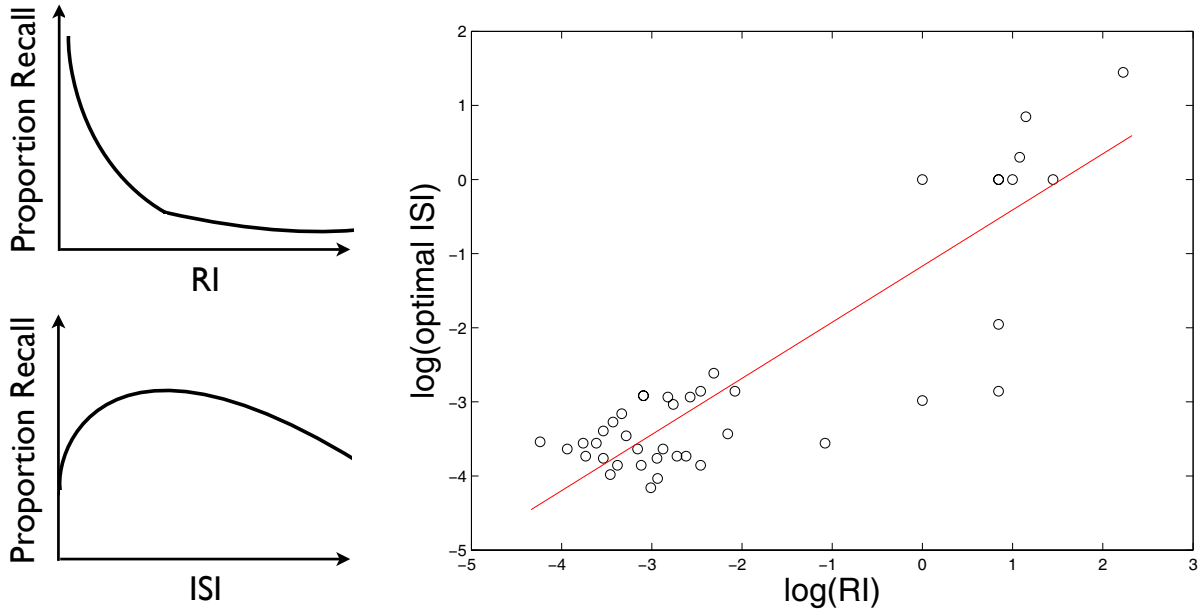


Figure 2.3: (upper left) Illustration of a forgetting curve. Test performance for a population decreases as a power-law function of time (Wixted & Carpenter, 2007). (lower left) Illustration of a spacing curve. For a fixed RI, increased spacing initially improves test performance but then decreases it. The ISI corresponding to the maximum of the spacing curve is termed the *optimal ISI*. (right) The relationship between the RI and optimal ISI from a meta-analysis conducted by Cepeda et al. (2006). Each point represents a spacing effect study. The optimal ISI systematically increases with the RI.

one of the most widely studied and robust phenomena in cognitive psychology (Dempster, 1988). Typical studies of the spacing effect have students study material one or more times, with each study session separated from the next by a temporal lag or *interstudy interval* (ISI) and with students then being tested following a *retention interval* (RI) (Figure 2.2). In many cases, large interstudy intervals can approximately double retention relative to smaller intervals (Cepeda et al., 2008; Melton, 1970). For a given RI, recall performance on the test follows a characteristically concave function often called the *spacing curve*. The ISI corresponding to the maximum of the spacing curve—*optimal ISI*—is known to systematically depend on the RI (Figure 2.3) and almost certainly depends on the student population and study material (Cepeda et al., 2006).

We begin by discussing theory-driven approaches to dynamic knowledge-state estimation. Our emphasis is on quantitative models of forgetting and spacing effects since these effects are

particularly relevant for estimation over educationally relevant timescales. We then discuss different data-driven modeling techniques. Many of the data-driven techniques we cover arose in the field of machine learning and, while applicable to this problem, have rarely been presented in this context. The models we will introduce and critique from the two approaches can in principle be applied to any domain whose mastery can be decomposed into distinct, separable elements of knowledge or items to be learned. Applicable domains range from the concrete to the abstract, and from the perceptual to the cognitive, and in principle span qualitatively different forms of knowledge including:

- declarative (factual) knowledge, e.g., "The German word for dog is *hund*" and "The American Civil War began in 1861";
- procedural (skill) knowledge, e.g., processing columns of digits in multidigit addition from right to left, and specifying unknown quantities as variables as the first step in translating algebraic word problems to equations; and
- conceptual knowledge, e.g., understanding betrayal ("Did Benedict Arnold betray his country?") and reciprocation ("How is the US-Pakistani relationship reciprocal?"), as well as perceptual categorization (e.g., classifying the species of a bird shown in a photo).

Finally, we conclude this chapter with a discussion of potential future research directions.

2.2 Theory-based approaches

Students gradually lose their mastery of study materials over time. Many people view memory as a faulty system, citing the loss of mastery as evidence of its defectiveness (J. Anderson & Schooler, 1991). However, a long tradition of research theorizes that human memory is in fact an *optimal* system, and a consequence of this view is that memory failure arises from the rational behavior of a physical system working to fulfill a particular goal. Humans have evolved to be well-adapted to their environment, and thus presumably human memory is adapted to fulfill the information processing

needs imposed upon it by the environment (Whitehill, 2013). By quantitatively characterizing these processing needs and limitations, we can construct highly constrained models for student knowledge-state estimation which mirror the history-dependent behavior of human memory.

It is often argued that the information-processing task being fulfilled by memory is one of information retrieval: the brain must manage a large collection of information and make relevant information available when needed (J. Anderson & Schooler, 1991). However, it is likely subject to constraints or costs that limit its ability to store or retrieve information—potentially arising, for example, from the metabolic costs of storing information in memory or from other physical considerations. Given the task and limitations, a rational system should make material available with respect to the pattern of past information presentation: material likely to be needed in the future should be made accessible, and material unlikely to be needed in the future should be made less accessible or be discarded.

In this section, we highlight three models of memory based on these ideas of predictive utility. The first, KTS, arose as a rational account of the temporal dynamics of adaptation to motor errors, but it has been shown also to give a coherent explanation of forgetting and spacing effects (Kording, Tenenbaum, & Shadmehr, 2007). The second, MCM, provides an implementation-level account (Marr, 1982) of the phenomena and shares much in common with KTS (Mozer, Pashler, Cepeda, Lindsey, & Vul, 2009). The final model we will discuss is based on ACT-R, an extremely influential cognitive architecture which pioneered the rational analysis of memory (J. Anderson, 1976).

2.2.1 Kording, Tenenbaum, & Shadmehr (2007)

The properties of muscles change over time due to factors such as fatigue, disease, and growth. Kording et al. (2007) proposed a rational model for adaptation to such variation in the motor system; we will refer to their model as KTS (for Kording, Tenenbaum, & Shadmehr). KTS is premised on the idea that the brain, as a kind of motor control system, must minimize movement errors by adapting to motor changes via sending appropriately adjusted control signals to the motor system. The adaptation must be different depending on the nature of the error. Adaptations should

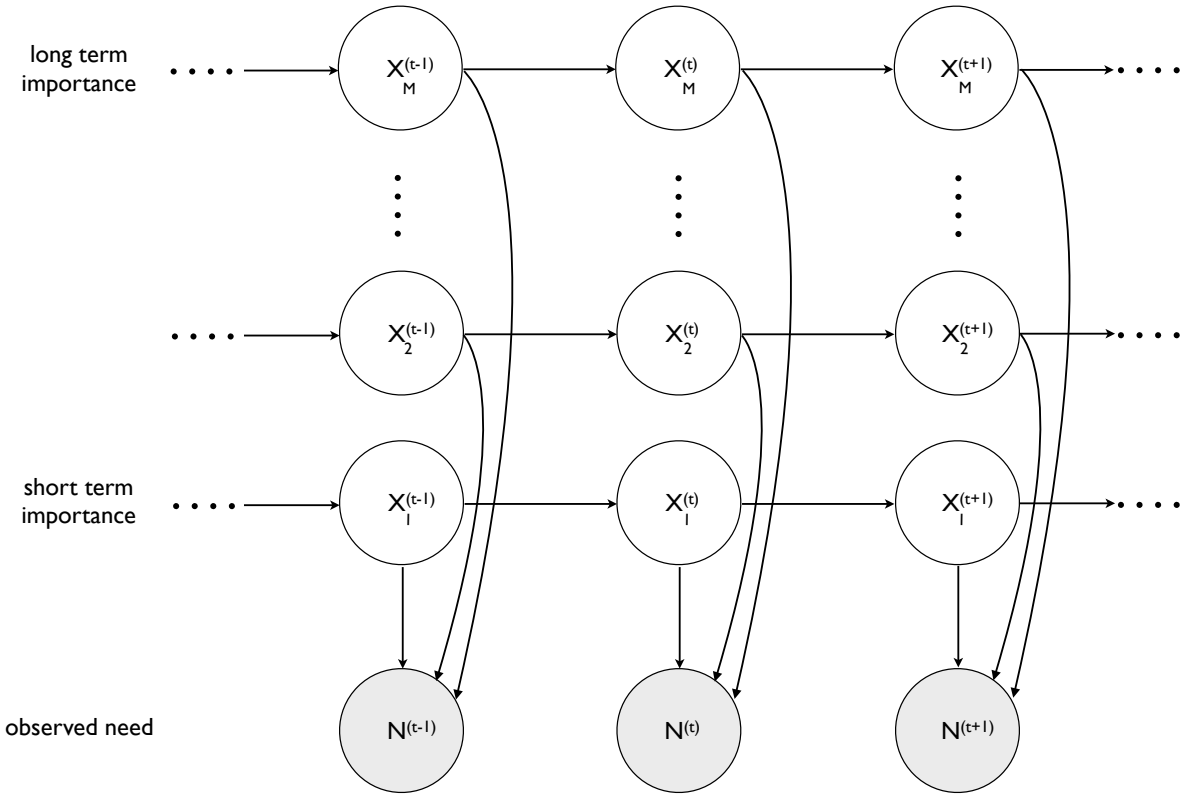


Figure 2.4: The KTS graphical model. An item's importance is assumed to vary over time as a set of independent random walks, each representing a different timescale. A rational learner must attribute an observed need, the noise-corrupted total importance, to the appropriate timescale.

be long-lasting for errors that are expected to persist (e.g., disease), and adaptations should be short-lived but rapidly made for errors that are not expected to persist (e.g., temporary fatigue). A rational motor-control system faces a credit assignment problem: when it observes a motor error, it must attribute the error to the timescale responsible for it so that the system can make the appropriate adaptation.

As Kording et al. (2007) discuss, KTS is equally applicable to predicting spacing effects in memory. The environmental need of a study item presumably changes over time, rising and falling over different timescales due to various factors. The temporal spacing of past exposure to the item provides evidence regarding the timescale over which the item is needed. A short ISI suggests that the item is needed over a short timescale, and a long ISI suggests a need that lasts over a

long timescale. A rational, Bayesian system should use its spacing-dependent estimates of need across time to perform the analogue of motor-control adaptations, making material available for the duration of its expected need.

Formally, KTS assumes that need at a particular timescale j (with $j \in 1, 2, \ldots, M$ and M being the number of timescales) is represented by a random walk over time $\mathbf{X}_j$,

$$X_j^{(t)} \mid X_j^{(t-1)} \sim \text{Normal}\left(\phi_j X_j^{(t-1)}, \sigma_j^2\right), \tag{2.2}$$

where $0 \leq \phi_j \leq 1$ is a decay rate specific to the timescale and σ_j^2 is the walk's variance. It is assumed that the variance is related to the decay rate as $\sigma_j^2 \triangleq c(1-\phi_j)^2$ where $c > 0$ is a free parameter. This parameterization produces low moment-to-moment variability in the importance of items needed over long timescales and high moment-to-moment variability in the importance of items needed over short timescales. In a study trial, it is assumed there is an observed need $N^{(t)}$ given by the noise-corrupted sum of the need on M individual timescales,

$$N^{(t)} \mid X_1^{(t)}, X_2^{(t)}, \ldots, X_M^{(t)} \sim \text{Normal}\left(\sum_{j=1}^{M} X_j^{(t)}, \sigma_w^2\right), \tag{2.3}$$

where σ_w^2 is a free parameter controlling the noise level.

The generative model of KTS is shown in Figure 2.4. Because human memory was not the focus of Kording et al. (2007), the generative model has no mechanism to map an observed need to student recall probability for retrieval practice trials. Kording et al. (2007) treated recall probability as given exactly by need N. That is, if $R \in \{0, 1\}$ is a Bernoulli random variable representing recall success or failure in a trial, then they assumed that $\Pr(R = 1) \approx N$. Subsequent work has used a more general affine transform, $\Pr(R = 1) \approx mN + b$, where m and b are free parameters (Mozer et al., 2009). The model approximates power-law forgetting through the parameterization of $\boldsymbol{\phi}$—though the individual random walks follow an exponential decay process, the sum of the different exponential-decay processes approximates power-law forgetting (R. Anderson & Tweney, 1997).

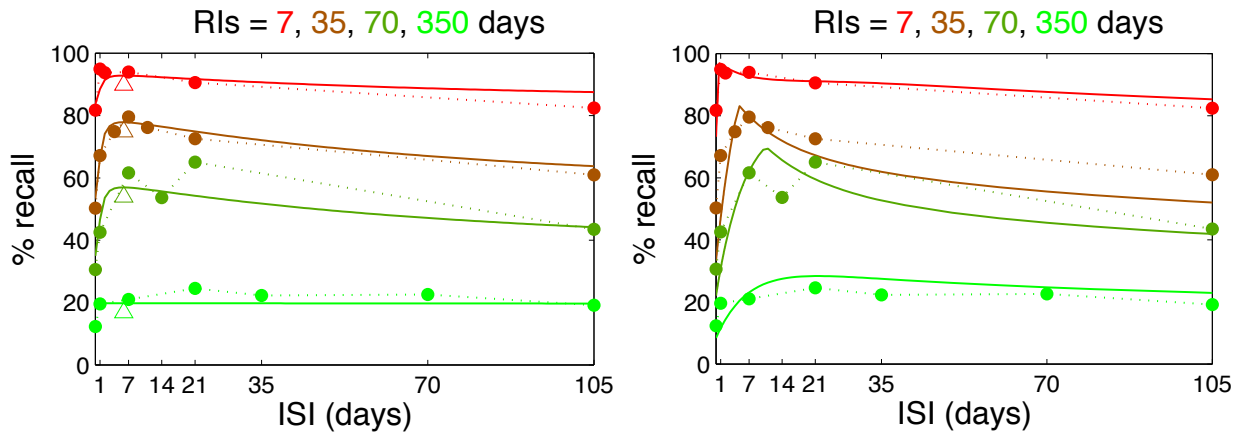


Figure 2.5: (left) We performed a least-squares fit of KTS to the spacing curves from a longitudinal spacing effect study in which subjects underwent two study sessions spaced in time and then a later test (Cepeda et al., 2008). Mean subject test performance is shown as the circles, the model's mean predictions are shown as solid lines, and the predicted optimal ISIs are shown as triangles. The alignment of triangles along the vertical axis suggests that the model is not suitably constrained to have the optimal ISI increase with the RI (recall Figure 2.3) (right) MCM's predictions of the spacing curves when the model is constrained to forgetting curve data (not shown). The model appears to properly increase the optimal ISI with the RI.

Exact, online, probabilistic knowledge-state inference in KTS is possible because it has a linear-Gaussian state space and linear-Gaussian observations. The algorithm used for inference in linear-Gaussian systems is called the Kalman Filter (Kalman, 1960; Welch, 2002); predictions are obtained through the filter's sequentially updated state estimate. To demonstrate KTS's ability to account for spacing effects, we have fit the model to data from a large-scale longitudinal spacing-effect study which used paired associates (Cepeda et al., 2008). In the study, each subject underwent a study session, waited an ISI (ranging from 10 minutes to 105 days), underwent another study session, and then was tested following an RI (ranging from 1 week to 350 days). The ISI and RI were manipulated between subjects. The experimental data and the model fit are shown in Figure 2.5. Despite having few free parameters, the model is able to fit the data tightly and produce spacing curves which exhibit a concavity characteristic of the spacing effect. However, recalling Figure 2.3, we note that the fit does not have qualitative properties entirely consistent with the spacing effect: the predicted optimal ISI (the triangles) should increase with the RI. Although other parameterizations of the model can produce the desired relationship between the optimal ISI and

RI, this simulation suggests that Equations 2.2-2.3 do not make restrictive enough assumptions about memory.

2.2.2 Multiscale context model

Mozer et al. (2009) developed an alternative model of spacing effects cast on Marr's algorithmic level (Marr, 1982) called the Multiscale Context Model (MCM). MCM provides a mechanistic basis for spacing effects and the power-law decay of recall probability over time. Synthesizing key ideas from previous models (Raaijmakers, 2003; Staddon, Chelaru, & Higa, 2002), MCM assumes that each time an item is studied, it is stored in multiple memory traces with each trace decaying at a different rate. Although each trace has an exponential decay, the sum of the traces approximates a power function of time. To highlight similarities betwen MCM and KTS, we present the model as the following discrete-time system:

$$x_j^{(t)} = \begin{cases} x_j^{(t-1)} + \ell_r(1 - n_j^{(t-1)}) & \text{if study trial at timestep } t \\ \phi_j x_j^{(t-1)} & \text{else} \end{cases} \tag{2.4}$$

where $x_j^{(t)}$ is the strength of trace j at time t, $n_j^{(t)} \triangleq \sum_{z=1}^{j} w_z x_z^{(t)}$ is the weighted total strength of traces $1 \ldots j$, w_z is a weight ($\sum_z w_z = 1$), ϕ_j is the decay rate of trace j, and ℓ_r is a recall-dependent parameter which controls the amount of learning that occurs because of a study trial. Mozer et al. (2009) present the continuous-time version of the above system, which is the limiting case as the timestep size goes to zero. Like KTS, MCM maps the total trace strength $n_M^{(t)}$ to recall probability via $\Pr(R = 1) = n_M^{(t)}$, with the additional formalization that recall probability is 1 if $n_M^{(t)} > 1$.

The state-space and observation equations of MCM can be shown to be the deterministic equivalent of KTS's distributional assumptions. However, the two models differ critically in how trace strength is updated after study. Due to the interpretation in KTS of the knowledge-state $\mathbf{X}^{(t)}$ as representing an external environmental need, "learning" amounts to probabilistic inference in the sense that a study trial provides evidence for a Bayesian learner to increase its *estimate* of an item's need, yet the *actual* need state is not causally affected by the study. This disconnect

can result in counter-intuitive predictions from the model. For example, it produces a kind of backwards causality in KTS: a student is more likely to recall an item at a time t if we know he or she was administered a trial at a later time $t' > t$. One would intuit that student recall probability in a study trial should be conditionally independent of the presence of future study trials, but it is not in KTS.

In contrast to KTS, MCM—a deterministic model—assumes that studying causally affects the knowledge state in a retrieval-dependent manner. Thus, studying at a point in time does not affect the posterior predictions for states at earlier times. The retrieval-dependent updating of memory strength occurs through a gradient-ascent rule, the consequence of which is that long-timescale traces are strengthened only if short-timescale traces have decayed away and thus could not have been responsible for the item being needed. This creates a tradeoff wherein increased interstudy spacing increases the amount of learning that occurs following study, but does so at the cost of losing the benefit of the earlier study. This tradeoff produces behavior consistent with the spacing effect.

To demonstrate the model's effectiveness, Mozer et al. (2009) fit all the model parameters except $\ell_{r=1}$ and $\ell_{r=0}$ to forgetting curve data from Cepeda et al. (2008), assumed $\ell_{r=0} \triangleq 1$, and then set $\ell_{r=1}$ by hand to an earlier experiment. This method of constraining the model is noteworthy because forgetting curve data is relatively easy to collect experimentally, unlike spacing curve data. To the extent that the model's predictions are insensitive to the choice of ℓ_r, MCM is well-suited for *predicting* recall performance as it depends on multiple spaced study sessions. This is particularly important for large, educationally relevant ISIs—e.g., weeks to years—for which model-constraining recall performance data following the spaced study cannot feasibly be collected in advance. The predictions MCM makes for the Cepeda et al. (2008) data are shown in Figure 2.5. The model makes very accurate predictions and correctly has the optimal ISI increasing with the RI.

Because the relationship between total trace strength and recall probability is simply a one-to-one mapping (confined to the unit interval), the model can produce qualitatively incorrect predictions. If total trace strength exceeds 1—as may happen when the ISIs are small or when there

are many study sessions—then recall probability is a constant 1 for the period of time till total trace strength decays to be below 1. This changepoint-like behavior is inconsistent with the power-law nature of forgetting. Additionally, from a modeling perspective, it is challenging to elegantly extend MCM for use in Bayesian settings because the model can assign a probability of zero to an observed recall accuracy (which may result in undefined conditional probabilities).

2.2.3 ACT-R

ACT-R is an influential cognitive architecture whose declarative memory module is often used to account for explicit recall following study (J. Anderson & Milson, 1989). It is motivated by the analogy of memory as an information retrieval system and is based on models of book usage in libraries (Burrell, 1980; Burrell & Cane, 1982). ACT-R assumes that a separate memory trace is laid down each time an item is studied. Each trace z decays according to a power law, Δt_z^{-d}, where Δt_z is the age of the trace at the time of current trial and d is the decay rate. On the kth study trial, the individual traces combine to yield a total trace strength m_k as

$$m_k = b + \ln\left(\sum_{z=1}^{k} \Delta t_z^{-d}\right), \tag{2.5}$$

where b represents base-level strength. Recall probability is given by the logistic function,

$$\Pr(R = 1) = \left[1 + e^{-(m_k+\tau)/c}\right]^{-1}. \tag{2.6}$$

where τ and c are additional free parameters. ACT-R's declarative memory module is fundamentally a logistic regression model whose predictor variables are well-motivated by arguments about trace decay. However, although individual traces decay according to a power-law in the model, recall probability does not. This is inconsistent with the power-law nature of forgetting (Wickelgren, 1976; Wixted & Carpenter, 2007). Furthermore, Equation 2.5 cannot account for spacing effects.

J. Anderson and Milson (1989) proposed a rational analysis of memory from which they estimated the future need probability of a stored trace. When an item is studied multiple times with a given ISI, the rational analysis suggests that the need probability drops off rapidly following

the last study once an interval of time greater than the ISI has passed. Consequently, increasing the ISI should lead to a more persistent memory trace. Although this analysis yields a reasonable qualitative match to spacing-effect data, no attempt was made to make quantitative predictions.

Pavlik and Anderson (2005, 2008) addressed spacing effects within ACT-R via the assumption that the decay rate for trace z depends on the total trace strength at the time of the zth study trial,

$$m_k = b + \ln\left(\sum_{z=1}^{k} \Delta t_z^{-d_z}\right) \qquad d_z = c\exp(m_{z-1}) + \alpha \tag{2.7}$$

where c and α are additional free parameters. High memory strength at the time of a study trial causes the new memory trace to have a greater decay rate, and low memory strength at the time of the trial produces a lesser decay rate. This tradeoff between having high memory strength at the time of study and having the benefit of the study persist longer produces behavior qualitatively consistent with the spacing effect.

The assumed relationship between memory strength and decay rate is not guided by any clear theoretical motivations and, to our knowledge, has not been compared to other plausible functional forms. However, we can deduce problems with the relationship by reasoning. For example, items or students with a high base-level activation b (e.g., easy material or smart students) will tend to have a more rapid decay rate than items or students with a low base-level activation (e.g., difficult material or not-so-smart students); this is necessarily the case if the two items or students have identical study histories. This relationship is precisely the opposite of what one might expect—easy material is generally more slowly forgotten than hard material, and smart students often forget more slowly than not-so-smart students. Nevertheless, the model is highly cited and has been the basis of much subsequent work (Pavlik, Presson, & Koedinger, 2007; Pavlik & Anderson, 2008; van Rijn, van Maanen, & van Woudenberg, 2009; Stewart & West, 2007).

A surprisingly common methodological practice in psychological modeling is to evaluate a model solely based on its ability to fit experimental data—some authors have estimated that there are thousands of papers promoting models and theories in the field of psychology exclusively through

this practice (Roberts & Pashler, 2000). In this tradition, the Pavlik and Anderson (2005) model of spacing effects was evaluated solely by calculating goodness-of-fit statistics and never by evaluating its ability to *predict* held-out data. That and the large number of papers citing their work give the impression of a rigorously evaluated model. However, a model's ability to over-fit data is not indicative of its ability to explain or predict phenomena. At the risk of being pedantic, we note that a model which has one free parameter per data point could trivially be constructed to *fit* any dataset perfectly, and the Pavlik and Anderson (2008) model has more than one parameter per data point for datasets that consist of one study trial per student-item pair. The variant of the model in Pavlik and Anderson (2008) was evaluated on held-out data, but it was not compared to any other models. Lindsey, Shroyer, Pashler, and Mozer (2014) demonstrated on one large dataset that the model cannot predict held-out data better than a trivial baseline model under a logarithmic loss function, which suggests that the ACT-R model of the spacing effect—though well-motivated and based on the strong tradition of ACT-R research—needs further refinement.

2.2.4 Discussion

A student's degree of mastery over study material is intricately tied to the amount and timing of past study. Theory-driven approaches to student knowledge-state estimation over time are based on computational models of human memory which quantify this law-like relationship. The psychological plausibility of the models has been demonstrated in the literature through fits to behavioral data from human experimental studies of spaced review. Obtaining a close correspondence between model and data is impressive to the degree that the model has few free parameters relative to the size of the data set being fit. Because minimizing the number of free parameters is key to a compelling account of memory, cognitive models typically fit aggregate data—data averaged over a population of students studying a set of items. However, they face a serious challenge in being useful for predicting the knowledge state of a particular item for a particular student. When a cognitive model is used to predict individual differences, it is usually assumed that each student or item is governed by a separate model parameterization (Navarro, Griffiths, Steyvers, & Lee, 2006).

This proliferation of parameters promotes over-fitting and is an impediment to making strong predictions. Moreover, aggregating behavioral data can produce data that does not accurately reflect the behavior of any individual and may potentially mislead researchers regarding the characteristics of the phenomena under study (Myung, Kim, & Pitt, 2000).

In the next section, we discuss an alternative modeling approach which focuses on modeling non-aggregate data from individual students studying individual items. These approaches are generally agnostic about human memory but specialize in capturing individual differences, often being based on common modeling techniques such as matrix factorization and hidden Markov models.

2.3 Data-driven approaches

Student knowledge-state estimation is a type of *dyadic data* prediction problem: the available behavioral data consist of two sets of entities (students and study items), and observations and predictions of response accuracy are made on *dyads* (student-item pairs) (Hofmann, Puzicha, & Jordan, 1999). Often associated with each dyad is a set of covariates—predictor variables—which influence the observation; for dynamic knowledge-state estimation, the covariates may include statistics of the spacing of past study and the pattern of past responses. Dyadic data arise in many domains beyond student modeling. Viewer-movie ratings, customer-product purchases, keyword-document occurrences, and image-feature observations are all common application domains involving dyadic prediction. Techniques for modeling customer-preference dyadic data broadly fall under the label *collaborative filtering* (Menon & Elkan, 2011) and are a highly active field of research.

In this section, we discuss three areas of dyadic data modeling in the context of knowledge-state estimation. The first is a family of regression models for categorical response data which arose in the psychometrics literature under the title of Item Response Theory. The second is a family of hidden Markov models which arose in the intelligent tutoring systems literature under the title of Knowledge Tracing. The third and final area we outline covers a broad range of clustering and factorial modeling techniques from machine learning. Though clustering and factorial modeling

have rarely been applied to dynamic knowledge-state estimation, they and related dyadic modeling techniques are often domain independent and thus applicable to this estimation problem.

2.3.1 Item response theory

A traditional psychometric approach to student modeling is item response theory (IRT), which is also known as latent trait theory (De Boeck & Wilson, 2004). The focus of IRT is typically on *static* knowledge-state estimation without regard for time- or practice-dependent factors. IRT is generally used to analyze tests and surveys post hoc in order to evaluate the diagnosticity of test items and the skill level of students (Roussos, Templin, & Henson, 2007). A common application of IRT is the analysis and interpretation of results from large standardized tests such as the SAT and GRE. Given a population of students answering a set of test items, IRT decomposes response accuracies into student- and item-specific parameters. The simplest form of IRT is the Rasch model (Rasch, 1961), a logistic regression model for dichotomous responses on dyadic student-item data. The Rasch model has factors representing a student-specific ability, α_s, and an item-specific difficulty, δ_i. Formally, the probability of student s making a correct response to item i is given by the constant

$$\Pr(R = 1) = \left[1 + e^{-(\alpha_s - \delta_i)}\right]^{-1}, \tag{2.8}$$

which is independent of the student's history of study with the item. Given a data set of response-accuracy observations $\mathbf{r}$ across students and items, the model is typically fit by finding the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ which maximize the likelihood of $\mathbf{r}$. Bayesian variants of IRT have been proposed that allow for additional knowledge in the form of hierarchical priors over student ability and item difficulty (Fox, 2010).

There are psychometric models which supplement or replace the difficulty parameter δ_i with parameters representing what subskills or *knowledge components* (KCs) are needed for anyone to correctly respond to the item (Fischer, 1973, 1995; Huguenard, Lerch, Junker, Patz, & Kass, 1997).

They have the general form

$$\Pr(R=1) = \left[1 + e^{-(\alpha_s - \delta_i - \mathbf{q}_i^\top \boldsymbol{\eta})}\right]^{-1}, \tag{2.9}$$

where $\mathbf{q}_i$ is a binary vector whose kth entry denotes whether students need to know KC k to make a correct response to item i, and $\boldsymbol{\eta}$ is a vector whose kth entry is the difficulty associated with KC k. One can think about η_k as summarizing the level of skill or knowledge needed to master a KC k, which in turn is required to answer item i when $q_{ik} = 1$.

There are also psychometric models which supplement the ability parameter by incorporating student-specific additive factors of the form

$$\Pr(R=1) = \left[1 + e^{-(\alpha_s + \mathbf{w}^\top \mathbf{f}_s - \delta_i)}\right]^{-1}, \tag{2.10}$$

where $\mathbf{w}$ is a vector of weights and $\mathbf{f}_s$ is a vector of student-specific factors representing either observable covariates (e.g., age of participant) or latent traits (Mislevy, 1987; Draney, Pirolli, & Wilson, 1995).

Other extensions to IRT have been proposed to allow for a student to have a different ability at different times (Andrade & Tavares, 2005), but many opportunities remain to explore variants of IRT that shift the focus from static to dynamic knowledge-state estimation and integrate the history of study into the predictions. The intelligent tutoring systems literature has recently taken steps in this direction, proposing models such as Learning Factors Analysis (Cen, Koedinger, & Junker, 2006, 2008), Performance Factors Analysis (Pavlik, Cen, & Koedinger, 2009), and Instructional Factors Analysis (Chi, Koedinger, Gordon, Jordan, & van Lehn, 2011). Such models are part of this tradition in IRT of decomposing student- and item-specific parameters into linear combinations of factors; they share the general form

$$\Pr(R=1) = \left[1 + e^{-(\alpha_s + \mathbf{w}^\top \mathbf{f}_s - \delta_i - \mathbf{q}_i^\top \boldsymbol{\eta})}\right]^{-1}, \tag{2.11}$$

which is the combination of Equations 2.9 and 2.10. The models differ primarily by the types of covariates they include—e.g., the amount of past practice, the success of past practice, and the types

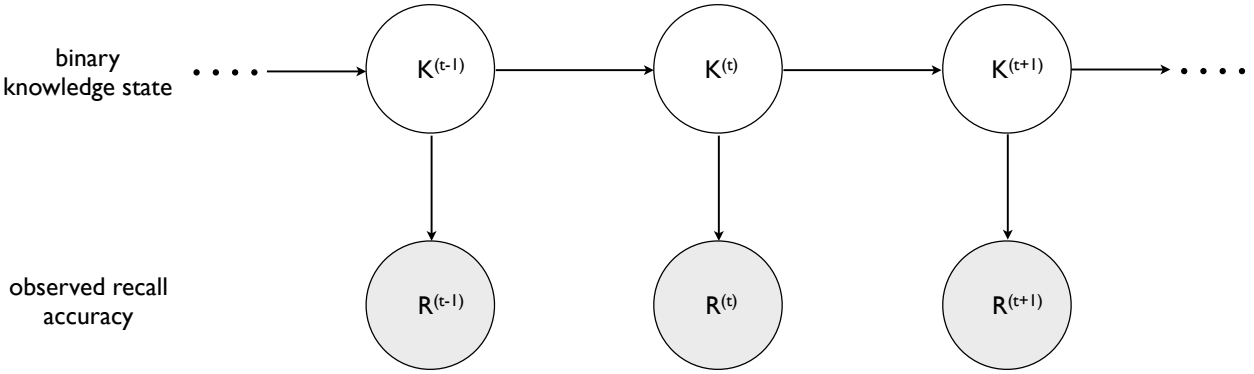


Figure 2.6: The Bayesian Knowledge Tracing (BKT) graphical model. An item is assumed either to be known or unknown in a trial: $K \in \{0, 1\}$. Recall accuracy is determined by a Bernoulli trial with a state-specific success probability.

of instructional intervention. It is through the choice of covariates that these IRT models incorporate a dependence on study history and thus provide time- and practice-dependent individualized predictions.

2.3.2 Bayesian knowledge tracing

A popular technique for dynamic knowledge-state estimation is Bayesian knowledge tracing (BKT) (Corbett & Anderson, 1995). Although originally used for modeling procedural knowledge acquisition, BKT could just as well be used for other forms of knowledge. BKT is a two-state hidden Markov model in which a latent state variable $K^{(t)} \in \{0, 1\}$ represents whether a skill is known by a student at a study trial t. With probability γ, the student can "learn" after the trial by transitioning from the unknown to the known state:

$$K^{(t)} \mid K^{(t-1)} \sim \text{Bernoulli}(\gamma^{1-K^{(t-1)}}). \tag{2.12}$$

It is assumed that the initial knowledge state is uncertain: $\Pr(K^{(1)} = 1) = \psi$ where ψ is a free parameter. Students respond correctly with probability μ_1 when the skill or item is known and with probability μ_0 when it is unknown,

$$R^{(t)} \mid K^{(t)} \sim \text{Bernoulli}(\mu_{K^{(t)}}), \tag{2.13}$$

The model thus has four free parameters (μ_0, μ_1, ψ, and γ). A convenient property of BKT is that predictions can trivially be sequentially calculated in closed form, analogous to—but much simpler than—how KTS updates its state estimate after each trial. Some of the popularity of BKT is likely attributable to this convenience.

By definition, there is no forgetting in BKT; once an item reaches the known state, it can never return to the unknown state. Unsurprisingly then, the model has been shown to consistently over-estimate learning (Corbett & Bhatnagar, 1997). Knowledge tracing's success is likely due to its use in modeling massed practice, an application area which has limited need to account for long-term forgetting and spacing effects. If the model is amended to allow for forgetting—transitions from the known to unknown state—it produces forgetting curves which are exponential and has decay rates which are independent of the past history of study. These properties are inconsistent with current beliefs about long-term memory (Wixted & Carpenter, 2007) and empirical observations concerning spacing effects (Pavlik & Anderson, 2005). There have been limited, orthogonal efforts to account for time within BKT (Qiu, Qi, Lu, Pardos, & Heffernan, 2011).

The parameters of BKT are typically fit to data from a population of students studying a population of items. Nevertheless, the model's predictions are individualized in the sense that they are specific to the history of response accuracies for each student-item dyad. There have been some efforts to individualize some or all of the parameters of BKT (Pardos & Heffernan, 2010; Lee & Brunskill, 2012). However, at the time of this writing, such efforts appear to be limited to fitting each student to a separate model parameterization, an approach which generally promotes over-fitting and limits the predictive power of the model. Further, these attempts at individualizing the BKT model parameters ignore individual differences in items and cannot readily be applied to students who have not undergone a large number of trials.

Practitioners who use BKT are often interested in interpreting or acting upon estimates of the knowledge-state variable $K^{(t)}$ (Baker, Corbett, & Aleven, 2008), in contrast to merely using predictions about future response accuracies given past response accuracies. This is problematic

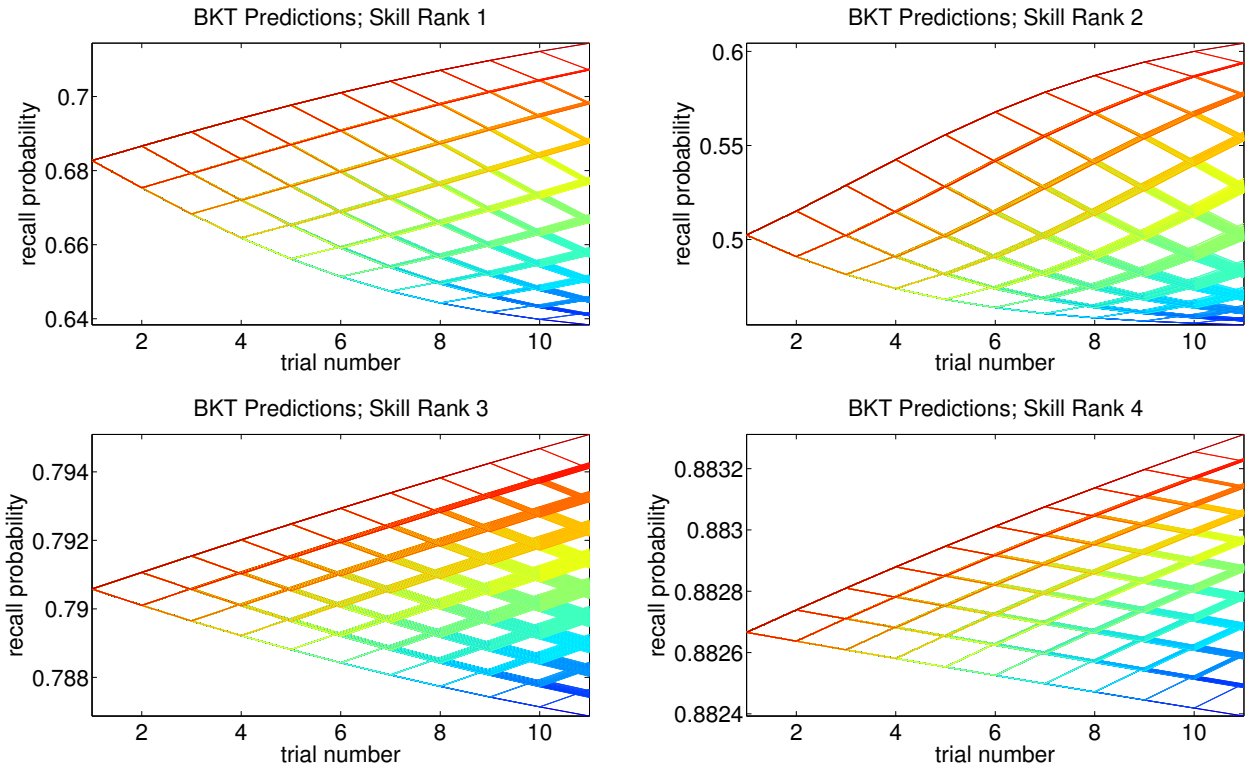


Figure 2.7: A representation of the predictions of BKT under four separate parameterizations. In each trial, BKT makes a prediction of a student's recall probability. After observing a binary recall event, it updates its prediction—thus, on the nth trial, there are 2^{n-1} possible predictions the model can make. Each line in this figure represents one possible trajectory through BKT's prediction space. The predictions are bounded above and below by μ_0 and μ_1.

for a number of reasons, especially when no constraints are imposed on the hyperparameters. For example, if $\mu_0 > \mu_1$, entering into the so-called known state counterintuitively decreases the probability of a correct response. A popular topic in the Knowledge Tracing community is how to solve problems arising from this practice (Beck & Chang, 2007; Beck, 2007; Baker et al., 2008). Because of the restrictive and unrealistic assumptions BKT makes about knowledge-state dynamics and how they relate to observed response accuracy, we suggest that the practice of interpreting the BKT knowledge-state variable as literally representing whether an item is known or not is misguided.

2.3.3 Clustering and factorial models

A general goal in unsupervised machine learning is to characterize a complex dataset by some simpler underlying or hidden structure. For dyadic data, one of the most straightforward methods to capture hidden structure is called *bi-clustering* or *co-clustering* (Hartigan, 1972; Shan & Banerjee, 2008; Kemp, Tenenbaum, Griffiths, Yamada, & Ueda, 2006). In the context of student-item response accuracies, bi-clustering involves assigning each student to a group and assigning each item to a group, then making predictions for a new dyad by looking at the student group and item group the dyad belongs to. For example, we might speculate that there are S types of students and I types of items, and that knowing what type of student is responding to what type of item is sufficient to guess the probability of a correct response being made to the specific student-item. Akin to how IRT models have been extended to account for covariates, bi-clustering methods can be extended to account for covariates as well (Agarwal & Merugu, 2007). There is little to no literature on applying those types of clustering methods to the problem of dynamic knowledge-state estimation. However, student-item modeling is a natural application of these domain-independent techniques. As in IRT, the covariates of these models could be chosen so as to account for the history-dependent nature of recall.

Clustering approaches which assume that each entity in the dyad belongs to a single latent group have limited representational power. A student, for example, might be best characterized as both a sixth grader and an honors student. An item, for example, might be best characterized as involving both long division and algebra. We may want a model which can uncover multiple causes or factors like these that interact to influence a student's response accuracy on an item. This is a type of *factorial learning* problem—the goal is to discover a parsimonious underlying representation which accounts for the observed data and reflects its multiple causes or influencing factors (Ghahramani, 1995).

Unsupervised matrix decomposition is a family of techniques for factorial learning. For dyadic data, these techniques are generally based on the assumption that associated with each student is

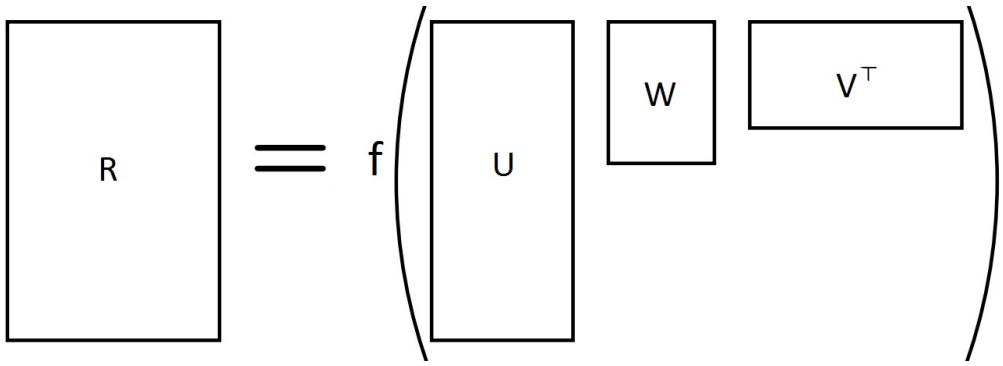


Figure 2.8: Schematic of matrix factorization techniques for knowledge-state estimation models; reproduced from Meeds et al. (2007). The dyadic data matrix of student response accuracy $\mathbf{R}$ is decomposed into three latent matrices: $\mathbf{U}$ contains student features (one row per student), $\mathbf{V}$ contains item features (one column per item), and $\mathbf{W}$ contains interaction weights for each student-item feature combination. A link function f (e.g., the logistic function) is applied element-wise to $\mathbf{UWV}^\top$.

a vector of features, $\mathbf{u}_s$, and likewise associated with each item is a vector of features, $\mathbf{v}_i$. Recall probability is related to these feature vectors via a link function (e.g., the logistic function) applied to the linear inner product of the feature vectors and weights (Meeds et al., 2007). Formally,

$$\text{logit } \mathbf{P} = \mathbf{UWV}^\top \tag{2.14}$$

where $\mathbf{P}$ is a matrix of recall probabilities, $\mathbf{W}$ is a matrix of interaction weights, $\mathbf{U}$ is a matrix such that row s contains the feature vector $\mathbf{u}_s$, and $\mathbf{V}$ is a matrix such that column i contains the feature vector $\mathbf{v}_i$. In student-item modeling, for example, response accuracies could be distributed as

$$R_{si} \mid \mathbf{P} \sim \text{Bernoulli}(P_{si}). \tag{2.15}$$

See Figure 2.8 for an illustration. Models differ significantly by what distributional assumptions they make about $\mathbf{U}, \mathbf{W}$, and $\mathbf{V}$. For example, in binary matrix factorization, $\mathbf{U}$ and $\mathbf{V}$ are given Beta-Bernoulli priors (Meeds et al., 2007). Determining *a priori* the number of features to use in factorial models is challenging. However, models like Meeds et al. (2007) use Bayesian nonparametric priors[1] in which the number of features is not fixed a priori and is allowed to grow with the

[1] They used the Indian Buffet Process, which arises in the limiting case of a particular Beta-Bernoulli prior.

complexity of the dataset. This type of model can easily be extended to handle covariates related to study history (Miller, Jordan, & Griffiths, 2009).

There have been preliminary efforts to apply factorization techniques to educational data mining (Toscher & Jahrer, 2010; Thai-Nghe, Drumond, Horváth, Krohn-Grimberghe, et al., 2011; Thai-Nghe, Drumond, Horváth, Nanopoulos, & Schmidt-Thieme, 2011), but this approach remains largely unexplored. Thai-Nghe, Horváth, and Schmidt-Thieme (2011) reported a slight improvement over BKT on two large datasets through a factorization approach which took into account time. However, they did not use cross validation and it is not clear whether the difference is significant. Because their factorization model is significantly more complex than BKT, it is questionable whether their result is generalizable. There has been limited work to extend other related collaborative filtering techniques to time-dependent student modeling (Cetintas, Si, Xin, & Hord, 2010). However, the study by Cetintas et al. (2010) was very limited in scope: it showed that a weighted-averaging technique which relies on a particular similarity metric between students does better if the metric takes into account all of a student's responses to problems, as opposed to discarding all but the most recent response.

Chapter 3

Modeling students' knowledge states

3.1 Preliminary investigation 1

In educational settings, individuals are often required to memorize facts such as foreign language vocabulary words. A question of great practical interest is how to retain knowledge once acquired. Psychologists have identified factors influencing the durability of learning, most notably the temporal distribution of practice: when individuals study material across multiple sessions, long-term retention generally improves when the sessions are spaced in time. This effect, known as the *distributed practice* or *spacing* effect, is typically studied via an experimental paradigm in which participants are asked to study items over two or more sessions, and the time between sessions—the *interstudy interval* or *ISI*—is varied. Retention is often evaluated via a cued recall test at a fixed lag following the final study session called the *retention interval* or *RI* (Figure 2.2).

Typical experimental results are shown in the data points and dotted lines of Figures 2a (Glenberg, 1976) and 2b (Cepeda et al., 2008). In both experiments, participants studied material at two points in time, with a variable ISI, and then were tested following a fixed RI. The graphs show recall accuracy at test as a function of ISI for several different RIs. The curves, which we will refer to as *spacing functions*, typically show a rapid rise in memory retention as ISI increases, reach a peak, and then gradually drop off. From the spacing function, one can determine the *optimal ISI*, the spacing of study that yields maximal retention. The exact form of the spacing function depends on the specific material to be learned and the RI. The distributed practice effect is obtained over a wide range of time scales: ISIs and RIs in the Glenberg study are on the order

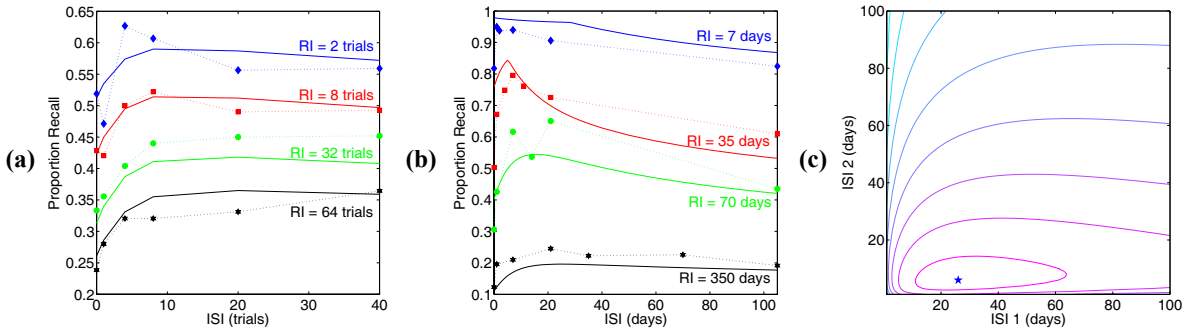


Figure 3.1: Results from **(a)** Glenberg (1976) and **(b)** Cepeda et al. (2008) illustrative of the distributed practice effect. The dotted lines correspond to experimental data. The solid lines in **(a)** and **(b)** are the ACT-R and MCM fits to the respective data. **(c)** A contour plot of recall probability as a function of two ISIs from ACT-R with parameterization in Pavlik and Anderson (2008).

of seconds to minutes, and in the Cepeda et al. study are on the order of weeks to months. On the educationally relevant time scale of months, optimally spaced study can double retention over massed study. Thus, determining the optimal spacing of study can have a tremendous practical impact on human learning.

Pavlik and Anderson (2005; 2008) used the ACT-R declarative memory equations to explain distributed practice effects. ACT-R supposes a separate trace is laid down for each study and that the trace decays according to a power function of time. The key feature of the model that yields the distributed practice effect is that the decay rate of a new trace depends on an item's current memory strength at the point in time when the item is studied. This ACT-R model has been fit successfully to numerous experimental datasets. The solid lines of Figure 3.1a show the ACT-R fit to the Glenberg data.

Mozer, Pashler, Lindsey, and Vul (submitted) have recently proposed a model providing an alternative explanation of the distributed practice effect. In this model, when an item is studied, a memory trace is formed that includes the current *psychological context*, which is assumed to vary randomly over time. Probability of later recall depends in part on the similarity between the context representations at study and test. The key feature of this model that distinguishes it from related past models (e.g., Raaijmakers, 2003) is that the context is assumed to wander on

multiple time scales. This model, referred to as the *multiscale context model* (MCM), has also been successfully fit to numerous empirical datasets, including the Glenberg study. In Figure 3.1b, we show the MCM prediction (solid lines) of the Cepeda et al. data.

Both ACT-R and MCM can be parameterized to fit data post hoc. However, both models have been used in a predictive capacity. Pavlik and Anderson (2008) have used ACT-R to determine the order and nature of study of a set of items, and showed that ACT-R schedules improved retention over alternative schedules. Mozer et al. (submitted) parameterize MCM with the basic forgetting function for a set of items (the function relating recall probability to RI following a single study session) and then predict the spacing function for the case of multiple study sessions. Figure 3.1b is an example of such a (parameter free) prediction of MCM.

Most experimental work involves two study sessions, the minimum number required to examine the distributed-practice effect. Consequently, models have mostly focused on this simple case. However, typical learning situations typically offer more than two opportunities to study material. The models can also predict retention following three or more sessions. In this section, we explore predictions of ACT-R and MCM in order to guide the design of future experiments that might discriminate between the models.

3.1.1 Study Schedule Optimization

A cognitive model of the distributed practice effect allows us to predict recall accuracy at test for a particular study schedule and RI. For example, Figure 3.1c shows ACT-R's prediction of recall probability for a study schedule with two variable ISIs and an RI of 20 days, for a particular parameterization of the model based on Pavlik and Anderson (2008). It is the two-dimensional generalization of the kind of spacing functions illustrated in Figures 3.1a and 2b. Recall probability, shown by the contour lines, is a function of both ISIs. The star in Figure 3.1c indicates the schedule that maximizes recall accuracy.

Models are particularly important for study-schedule optimization. It is impractical to determine optimal study schedules empirically because the optimal schedule is likely to depend on

the particular materials being learned and also because the combinatorics of scheduling $n+1$ study sessions (i.e., determining n ISIs) make it all but impossible to explore experimentally for $n > 1$. With models of the distributed practice effect, we can substitute computer simulation for exhaustive human experimentation.

In real-world learning scenarios, we generally do not know exactly when studied material will be needed; rather, we have a general notion of a span of time over which the material should be retained. Though not the focus of this section, models of the distributed practice effect can be used to determine study schedules that maximize retention not only for a particular prespecified RI, but also for the situation in which the RI is treated as a random variable with known distribution. The method used in this section to determine optimal study schedules can easily be extended to accomodate uncertain RIs.

3.1.2 Models to Evaluate

3.1.2.1 Pavlik and Anderson ACT-R Model

In this section, we delve into more details of the Pavlik and Anderson (2005; 2008) model, which is based on ACT-R declarative memory assumptions. In ACT-R, a separate trace is laid down each time an item is studied, and the trace decays according to a power law, t^{-d}, where t is the age of the memory and d is the power law decay for that trace. Following n study episodes, the activation for an item, m_n, combines the trace strengths of individual study episodes:

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

where t_k and d_k refer to the age (in seconds) and decay associated with trace k, and the additive parameters β_s, β_i, and β_{si} correspond to participant, item, and participant-item factors that influence memory strength, respectively. The variable b_k reflects the salience of the kth study session (Pavlik et al., 2007); larger values of b_k correspond to cases where, for example, the participant self-tested and therefore exerted more effort.

The key claim of the ACT-R model with respect to the distributed-practice effect is that

the decay term on study trial k depends on the item's overall activation at the point when study occurs, according to the expression:

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

where c and α are constants. If spacing between study trials is brief, the activation m_{k-1} is large and consequently the new study trial will have a rapid decay, d_k. Increasing spacing can therefore slow memory decay of trace k, but it also incurs a cost in that traces $1...k-1$ will have substantial decay.

The model's recall probability is related to activation by:

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

where τ and s are additional parameters. The pieces of the ACT-R model relevant to this section include 3 additional parameters, for a total of 10 parameters, including: h, a translation of real-world time to internal model time, u, a descriptor of the maximum benefit of study, and v, a descriptor of the rate of approach to the maximum.

Pavlik and Anderson (2008) use ACT-R activation predictions in a heuristic algorithm for scheduling the trial order *within* a study session, as well as the trial type (i.e., whether an item is merely studied, or whether it is first tested and then studied). They assume a fixed intersession spacing. Thus, their algorithm reduces to determining how to best allocate a finite amount of time within a session.

Although they show a clear effect of the algorithm used for within-session scheduling, we focus on the complementary issue of scheduling the lag between sessions. The ISI manipulation is more in keeping with the traditional conceptualization of the distributed-practice effect. Fortunately, the ACT-R model can be used for both within- and between-session scheduling. To model between-session scheduling, we assume—as is true in controlled experimental studies—that each item to be learned is allotted the same amount of study (or test followed by study) time within a session.

Pavlik and Anderson (2008) describe their within-session scheduling algorithm as optimizing performance, yet we question whether their algorithm is appropriately cast in terms of optimization.

They argue that maximizing probability of recall should not be the goal of a scheduling algorithm, but that activation gain at test should be maximized so as to encourage additional benefits (e.g., improved long-term retention). We believe that had Pavlik and Anderson (2008) sought simply to maximize probability of recall at test and had more rigorously defined their optimization problem, they would have seen results of the ACT-R within-session scheduler even better than what they achieved. In light of these facts, we contend that our work is the first effort to truly optimize memory retention via cognitive models.

3.1.2.2 Multiscale Context Model

One class of theories proposed to explain the distributed-practice effect focuses on the notion of encoding variability. According to these theories, when an item is studied, a memory trace is formed that incorporates the current psychological context. Psychological context includes conditions of study, internal state of the learner, and recent experiences of the learner. Retrieval of a stored item depends partly on the similarity of the contexts at the study and test. If psychological context is assumed to fluctuate randomly, two study sessions close together in time will have similar contexts. Consequently, at the time of a recall test, either both study contexts will match the test context or neither will. A longer ISI can thus prove advantageous because the test context will have higher likelihood of matching one study context or the other.

Raaijmakers (2003) developed an encoding variability theory by incorporating time-varying contextual drift into the Search of Associative Memory (SAM) model and used this model to explain data from the distributed-practice literature. The context consists of a pool of binary-valued neurons which flip their state at a common fixed rate. This behavior results in exponentially decreasing similarity between contexts at study and test time as a function of the study-test lag.

In further explorations, we (Mozer et al., 2009) found a serious limitation of SAM: Distributed-practice effects occur on many time scales (Cepeda et al., 2006). SAM can explain effects for study sessions separated by minutes or hours, but not for sessions separated by weeks or months. The reason is essentially that the exponential decay in context similarity bounds the time scale at which

the model operates.

To address this issue, we proposed a model with multiple pools of context neurons. The pools vary in their relative size and the rate at which their neurons flip state. With an appropriate selection of the pool parameters, we obtain a model that has a power-law forgetting function and is therefore well suited for handling multiple time scales. The notion of multiscale representations comes from another model of distributed-practice effects developed by Staddon et al. (2002) to explain rat habituation. We call our model, which integrates features of SAM and Staddon et al.'s model, the Multiscale Context Model (MCM).

MCM has only five free parameters. Four of these parameters configure the pools of context neurons, and these parameters can be fully constrained for a set of materials to be learned by the the basic forgetting function—the function characterizing recall probability versus lag between a single study opportunity and a subsequent test. Given the forgetting function, the model makes strong predictions concerning recall performance at test time given a study schedule.

MCM predicts the outcome of four experiments by Cepeda et al. (2008). These experiments all involved two study sessions with variable ISIs and RIs. Given the basic forgetting functions for the material under study, MCM accurately predicted the ISI yielding maximal recall performance at test for each RI. Although MCM is at an early stage of development, the results we have obtained are sufficiently promising and robust that we find it valuable to explore the model's predictions and to compare them to the well-established ACT-R model.

3.1.3 Comparing Model Predictions

Having introduced the ACT-R model and MCM, we now turn to the focus of this section: obtaining predictions from the two models to determine whether the models are distinguishable. We focus on the most important, practical prediction that the models can make: how to schedule study to optimize memory retention. We already know that the models make similar predictions in empirical studies with two study sessions (one ISI); we therefore turn to predictions from the models with more than two sessions (two or more ISIs). Even if the models make nonidentical

predictions, they may make predictions that are quantitatively so similar that the models will in practice be difficult to distinguish. We therefore focus our explorations on whether the models make qualitatively different predictions. Constraining our explorations to study schedules with three study sessions (i.e., two ISIs), we test whether the models predict that optimal study schedules have *expanding*, *contracting*, or *equal* spacing, that is, schedules in which ISI 1 is less than, greater than, or equal to ISI 2, respectively. For the sake of categorizing study schedules, we judge two ISIs to be equal if they are within 30% of one another. The key conclusions from our experiments do not depend on the precise setting of this criterion.

In all simulations, we used the Nelder-Mead Simplex Method (as implemented in Matlab's `fminsearch`) for finding the values of ISI 1 and ISI 2 that yield the maximum recall accuracy following a specified RI. Because this method finds local optima, we used random restarts to increase the likelihood of obtaining global optima. We observed some degenerate local optima, but for the most part, it appeared that both models had spacing functions like those in Figures 3.1a and 3.1b with a single optimum.

Our first exploration of the models' spacing predictions uses parameterizations of the models fit to the Glenberg (1976) data (Figure 3.1a for ACT-R, not shown for MCM). Because the models have already been constrained by the experimental data, which involved two study opportunities, they make strong predictions concerning memory strength following three spaced study opportunities. We used the models to predict the (two) optimal ISIs for RIs ranging from ten minutes to one year. We found that both models predict contracting spacing is optimal regardless of RI. The spacing functions obtained from the models look similar to that in Figure 3.1c. Because the models cannot be qualitatively discriminated based on the parameters fit to the Glenberg experiment, we turn to exploring a wider range of model parameterizations.

3.1.4 Randomized Parameterizations

In this section, we explore the predictions of the models across a wide range of RIs and model parameterizations in order to determine whether we can abstract regularities in the models'

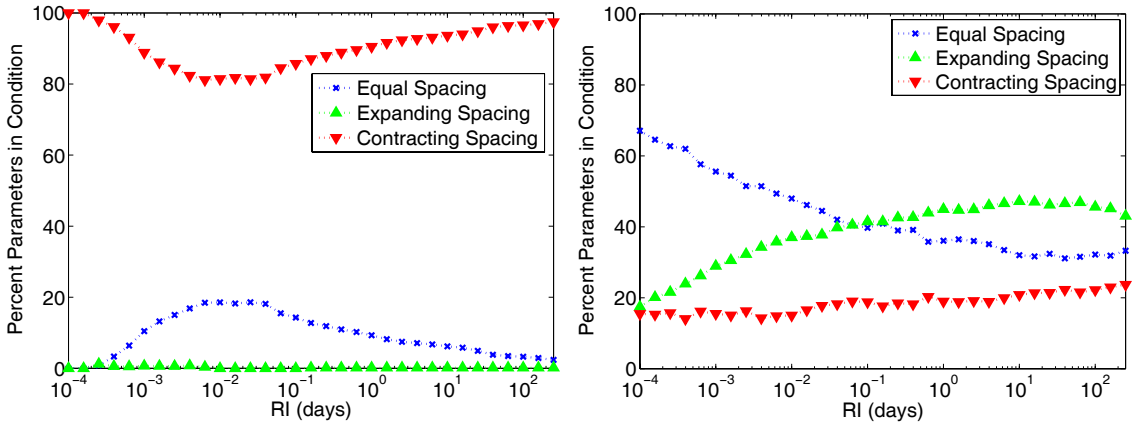


Figure 3.2: The distribution of qualitative spacing predictions of ACT-R (left figure) and MCM (right figure) as a function of RI, for random model variants. Each point corresponds to the percentage of valid model fits that produced a particular qualitative spacing prediction.

predictions that could serve to discriminate between the models. In particular, we are interested in whether the optimality of contracting spacing predicted by both models for the Glenberg paradigm and material is due to peculiarities of that study, or whether optimality of contracting spacing is a robust parameter-independent prediction of both models.

3.1.4.1 Methodology.

We performed over 200,000 simulations for each model. In our simulations, we systematically varied the RIs from roughly 10 seconds to 300 days. We also chose random parameter settings that yielded sensible behavior from the models. We later expand on the notion of "sensible."

For the ACT-R model, we draw the parameters β_i, β_s, β_{si} from Gaussian distributions with standard deviations specified in Pavlik and Anderson (2008). The parameters h, c, and α are drawn from a uniform distribution in [0, 1]. The study weight parameter b is fixed at 1, which assumes test-practice trials (Pavlik & Anderson, 2008). Remaining parameters of the model are fixed at values chosen by Pavlik and Anderson (2008). For MCM, we vary the four parameters that determine the shape of the forgetting function.

To ensure that the randomly generated parameterizations of both models are sensible—i.e., yield behavior that one might expect to observe of individuals studying specific materials—we

observe the forgetting function for an item studied once and then tested following an RI, and place two criteria on the forgetting function: (1) With an RI of one day, recall probability must be less than 0.80. (2) With an RI of thirty days, recall probability must be greater than 0.05. We thus eliminate parameterizations that yield unrealistically small amounts of forgetting and too little long-term memory.

3.1.4.2 Results.

Results of our random-parameter simulations are presented in Figures 3.2 and 3.3. The left graphs of each figure are for the ACT-R model and the right graphs are for MCM. Figure 3.2 shows, as a function of the RI, the proportion of simulations that yield contracting (red curve), expanding (green curve), and equal (blue curve) optimal spacing. The ACT-R model (Figure 3.2, left) strongly predicts that contracting spacing is optimal, regardless of the RI and model parameters. In contrast, MCM (Figure 3.2, right) suggests that the qualitative nature of the optimal study schedule is more strongly dependent on RI and model parameters. As the RI increases, the proportion of expanding spacing predictions slowly increases and the proportion of equal spacing predictions decreases; contracting spacing predictions remain relatively constant. Over a variety of materials to be learned (i.e., parameterizations of the model), MCM predicts that expanding spacing becomes increasingly advantageous as the RI increases.

Each scatter plot in Figure 3.3 contains one point per random simulation, plotted in a log-log space that shows the values of the optimal ISI 1 on the x-axis and the optimal ISI 2 on the y-axis. In other words, each point is like the star (point of optimal retention) of Figure 3.1c, plotted for a unique parameterization and RI. The two solid diagonal lines represent the decision boundary between the different qualitative spacing predictions. Points between the decision boundaries are within 30% of each other (in linear space) and fall under the label of equal spacing. Points above the upper diagonal line are classified as expanding spacing, and points below the lower diagonal line are classified as contracting spacing. The color of the individual points specifies the corresponding RI.

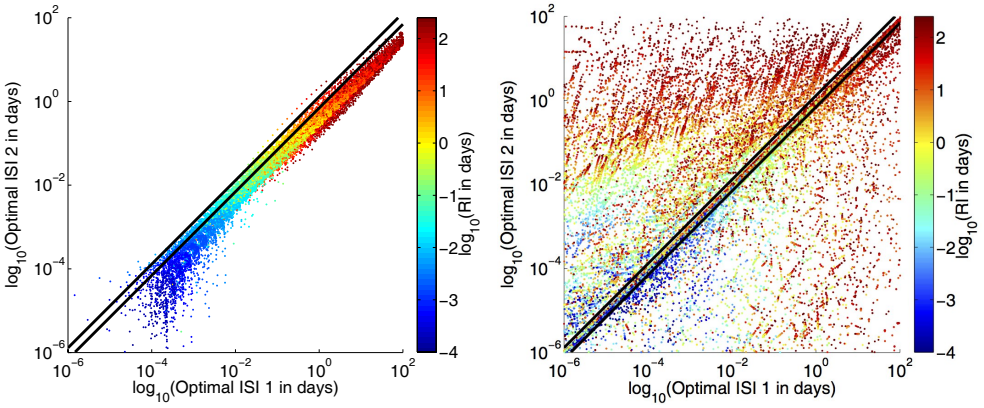


Figure 3.3: Optimal spacing predictions in log-space of ACT-R (left figure) and MCM (right figure) for random parameter settings over a range of RIs. Each point corresponds to a parameter setting's optimal spacing prediction for a specific RI, indicated by the point's color. The black lines indicate the boundaries between expanding, equal, and contracting spacing predictions.

The spacing functions produced by the ACT-R model are fairly similar, which is manifested not only in the consistency of the qualitative predictions (Figure 3.2, left), but also the optimal ISIs (Figure 3.3, left). The relationship between optimal ISI 1 and optimal ISI 2 appears much stronger for the ACT-R model than for MCM, and less dependent on the specific model parameterization. Not only do we observe a parameter-independent relationship between the optimal ISIs, but we also observe a parameter-independent relationship between the RI and each of the ISIs. The apparent linearity in the left panel of Figure 3.3 translates to a linear relationship in log-log space between RI and each of the optimal ISIs. The least-squares regression yields:

$$\log_{10}(ISI_1) = 1.0164 \log_{10}(RI) + 0.5091$$

$$\log_{10}(ISI_2) = 1.0237 \log_{10}(RI) + 0.9738$$

with coefficient of determination (ρ^2) values of 0.89 and 0.90, respectively. We emphasize that these relationships are predictions of a model, not empirical results. The only empirical evidence concerning the relationship between RI and the optimal ISI is found in Cepeda et al. (2006), who performed a meta-analysis of all cogent studies of the distributed-practice effect and observed a roughly log-log linear relationship between RI and optimal ISI for experiments consisting of two study sessions (one ISI). Were this lawful relationship to exist, it could serve as an extremely useful

heuristic for educators who face questions such as: If I want my students to study this material so that they remember it for six months until we return to the same topic, how should I space the two classes I have available to cover the material?

In further contrast with ACT-R, MCM's optimal ISI predictions are strongly parameter dependent (Figure 3.3, right). Is this result problematic for MCM? We are indeed surprised by the model's variability, but there are no experimental data at present to indicate whether such variability is observed in optimal study schedules for different types of material (as represented by the model parameters).

Although ACT-R shows greater regularity in its predictions than MCM, as evidenced by the contrast between the left and right panels of Figure 3.3, note that both models make optimal spacing predictions that can vary by several orders of magnitude for a fixed RI. No experimentalist would be surprised by the prediction of both models that optimal spacing of study for a given RI is material-dependent, but this point has not been acknowledged in the experimental literature, and indeed, the study by Cepeda et al. (2008) would seem to suggest otherwise: two different types of material yielded spacing functions that appear, with the limited set of ISIs tested, to peak at the same ISI.

Another commonality between the models is that both clearly predict the trend that optimal ISIs increase with the RI. This is evidenced in Figure 3.3 by the fact that the long RIs (red points) are closer to the upper right corner than the short RIs (blue points). Although the experimental literature has little to offer in the way of behavioral results using more than two study sessions, experimental explorations of the distributed-practice effect with just two study sessions do suggest a monotonic relationship between RI and the optimal ISI (Cepeda et al., 2006).

3.1.5 Discussion

In this section, we have explored two computational models of the distributed practice effect, ACT-R and MCM. We have focused on the educationally relevant issue of how to space three or more study sessions so as to maximize retention at some future time. The models show some points

of agreement and some points of fundamental disagreement.

Both models have fit the experimental results of Glenberg (1976). With the parameterization determined by this fit, both models make the same basic prediction of contracting spacing being optimal when three study sessions are involved. Both models also agree in suggesting a monotonic relationship between the RI and the ISIs. Finally, to differing extents, both models suggest that optimal spacing depends not only on the desired RI, but also on the specific materials under study.

When we run simulations over the models' respective parameter spaces, we find that the two models make remarkably different predictions. ACT-R strongly predicts contracting spacing is best regardless of the RI and materials. In contrast, MCM strongly predicts that equal or expanding spacing is best, although it shows a greater dependence on both RI and the materials than does ACT-R. This stark difference between the models gives us a means by which the models can be evaluated. One cannot ask for any better set-up to pit one model against the other in an experimental test.

In reviewing the experimental literature, we have found only four published papers that involve three or more study sessions and directly compare contracting versus equal or contracting versus expanding study schedules (Foos & Smith, 1974; Hser & Wickens, 1989; Landauer & Bjork, 1978; Tsai, 1927). *All four studies show that contracting spacing leads to poorer recall at test than the better of expanding or equal spacing.* These findings are consistent with MCM and inconsistent with ACT-R. However, the findings hardly allow us to rule out ACT-R, because it would not be surprising if a post-hoc parameterization of ACT-R could be found to fit each of the experimental studies.

Nonetheless, the sharp contrast in the predictive tendencies of the two models (Figure 3.6) offers us an opportunity to devise a definitive experiment that discriminates between the models in the following manner: We conduct an experimental study with a single ISI and parameterize both models via fits to the resulting data. We then examine the constrained models' predictions regarding three or more study sessions. If ACT-R predicts decreasing spacing and MCM predicts equal or increasing spacing, we can then conduct a follow-on study in which we pit the predictions

of two fully specified models against one another. Without extensive simulation studies of the sort reported in this section, one would not have enough information on how the models differ to offer an approach to discriminate the models via experimental data.

3.2 Preliminary investigation 2

Our ultimate goal is to design intelligent tutoring systems (ITSs) that help students memorize a set of facts, such as the English equivalents of foreign words, that are to be learned before some future test date. An effective way to teach this kind of material is to test students while they are studying (H. Roediger & Karpicke, 2006a). For example, if a student is learning the meanings of foreign words, an appropriately designed ITS would display a foreign word, ask the student to guess the English translation, and then provide the correct answer. In this work, we consider the case where students undergo several rounds of this type of study. By convention, we refer to the group of rounds as a **study session**. At the end of a study session, students have had several encounters with each item being studied.

In addition to promoting robust learning, testing students during study provides valuable information that, in principle, can be used to infer a student's current and future state of memory for the material. Through the use of a student's performance during study to predict recall at a subsequent test, informed decisions can be made about the degree to which individual facts would benefit from further study. In this section, we explore algorithms to predict a student's future recall performance on specific facts using both the accuracy of the student's responses during study, and his or her response latencies—the time it took to produce the responses. In principle, other information is available as well, such as the nature of errors made and the student's willingness to guess a response. However, we restrict ourselves to accuracy and latency data because such data are independent of the domain and the study question format. Thus, we expect that algorithms that base their predictions on accuracy and latency data will be applicable to many domains.

Predicting future recall accuracy from observations during study can be posed as a machine learning problem. Given a group of students for whom we have made observations, we divide the

students into "training" and "test" groups. The training group is used to build predictive models whose performance is later evaluated using the test group. We developed several predictive models and describe them later in this section. Of particular interest is a method we call Bayesian ACT-R (BACT-R). It is based on the popular ACT-R cognitive architecture (J. Anderson et al., 2004), which has equations that interrelate response latency during study, accuracy during study, the time periods separating study sessions from one another and from the test, and the probability of a correct answer at test. However, these equations have a large number of free parameters, which makes it challenging to use the model in a truly predictive manner. BACT-R is a method for using Bayesian techniques to infer a distribution over the free parameters, which makes it possible to use the ACT-R equations to predict future recall.

This section is organized as follows: first, we describe the experiment from which we obtained accuracy and latency data for a group of students studying paired associates. Next, we describe BACT-R and three other models we built to predict student recall in the experiment. Finally, we evaluate and discuss the performance of the algorithms.

Our data are from an unpublished experiment by Pashler, Mozer, and Wixted (Pashler, Mozer, & Wixted, unpublished), in which 56 undergraduates tried to learn the disciplines of 60 relatively obscure Nobel prize winners. During a first pass through the material, subjects were shown the names of the winners paired with their disciplines. Each winner-discipline pair was displayed for five seconds. For each prize winner's name, subjects were given either three or six study opportunities during which they could guess the discipline. For each guess, they received auditory feedback that signaled whether or not the guess was correct. If it was incorrect, the correct answer was displayed on the screen. For these study trials, subjects responded by pressing one of four keys on a keyboard (the experiment involved only three disciplines, and a fourth key indicated a "no guess"). During study, both the accuracies and latencies of the subjects' responses were recorded. Two weeks following study, subjects were evaluated in a cumulative test over all the material. The cumulative test was given in the same format as the study trials.

3.2.1 Approaches to consider

In our machine learning approach to predicting student recall at test, we split subjects into training and test groups. For both the training and test groups, we gave our algorithms access to response accuracies and latencies obtained during the study session. Additionally, we gave the algorithm access to the response accuracies at the cumulative test for only the training group. In this section, we describe four increasingly complex algorithms designed to learn from the training group in order to make predictions about the test group.

We use the information from the training subjects to build a model that we apply to the test subjects to predict the probability that they will answer correctly when tested. The model is then evaluated on the test subjects: for each subject s in the test group and item i being learned, we use the model to predict the probability that s correctly recalled i when tested, and compare this prediction to the observed accuracy. In the future, we will refer to s and i as a "subject-item pair."

Because all subjects learned the same set of items, it is possible to use the performance of the training group on a particular item to inform the predicted performance of the test group on this item. We chose to avoid methods that do this because they are restricted to situations where data are available for a large number of subjects learning the same set of items. In principle, the methods we explore here might work even if individuals learned different items chosen from the same domain.

3.2.1.1 Percentage Classifier

This was the simplest method we examined: given a subject-item pair, the predicted probability of a correct answer at test is the simply the fraction of correct answers given during study. Unlike the other methods we describe in this section, the percentage classifier does not use data from the training subjects.

3.2.1.2 Histogram Classifier

For this method, we specified each subject-item pair by two numbers: the fraction of correct answers during study and the mean latency of the correct answers. We then formed two grids, one for the subject-item pairs that had three trials and another for the pairs that had six. The grids were formed in the following way: one axis had n numbers, such that each interval between two successive numbers contained an equal number of the mean latencies for the training set. n is a parameter of the model and was chosen by cross-validation. The other axis contained either four (for the three-session grid) or seven (for the six-session grid) numbers, such that each interval between two successive numbers contained exactly one of the possible fractions of correct answers. Each training example could then be placed in exactly one of the grid cells. For each cell, we found the number of training examples that fell within the cell and how many of these corresponded to a correct answer at evaluation. This enabled us to find, for each cell, a fraction correct. Given a test subject-item pair, we then found which cell it would fall into based on study performance and predicted that its probability of being correct at evaluation would be that cell's fraction correct. Figure 3.4 shows the grid for the six-trial case. Note that to display the figure, we had to fix the number of bins. In reality, since this number was chosen by cross-validation, it would be different for each test subject.

3.2.1.3 Logistic Regression

Logistic regression is a common method in statistics and machine learning that, in its simplest form, takes the values of some number of predictor variables x_i (which may be either binary or continuous) corresponding to an input and then outputs a prediction of the probability that the input belongs to one of two classes. This probability of membership in one of the classes is given by:

$$f(x_1, \ldots, x_n) = \left[1 + \exp(-\beta_0 - \sum_{i=1}^{n} \beta_i x_i)\right]^{-1}$$

The weights β_i are to be learned. β_0 is an offset term.

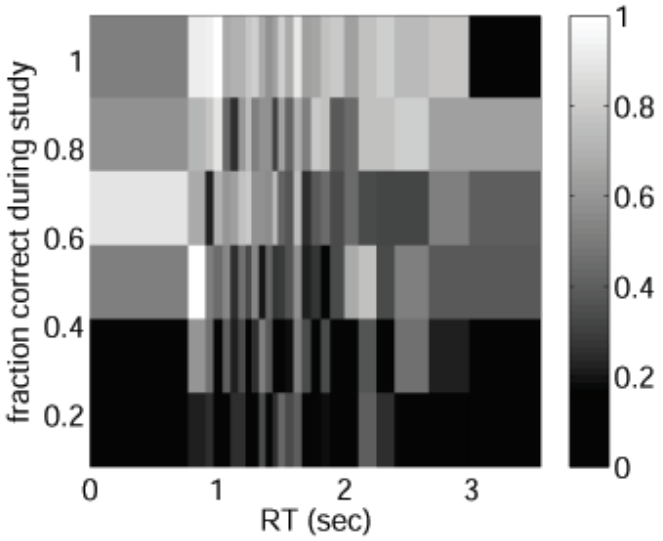


Figure 3.4: The grid used by the histogram classifier for subject-item pairs that had six study trials. Shading indicates the fraction of those subject-item pairs in the cell that had a correct answer at test. In this figure, the number of bins has been fixed. In practice, it is chosen by cross-validation and is unique to each test subject.

In this application, the predictor variables x_i are the latencies and accuracies obtained during study. More specifically, to predict the probability of a correct response at test for a subject-item pair with three study trials, we use six predictor variables. Three of these are binary and indicate whether each of the three answers given during study were correct or incorrect. The other three variables are the response latencies for the study answers and are therefore continuous. The predictor variables are constructed analogously for the six trial cases. The two classes are "correct answer at test" and "incorrect answer at test."

3.2.1.4 BACT-R

ACT-R is an influential cognitive architecture whose declarative memory module is often used to model recall follow a series of study sessions (e.g., Pavlik and Anderson (2008)). ACT-R assumes a separate trace is laid down each time an item is studied. Each trace decays according to a power law, t^{-D}, where t is the age of the memory and D is the decay rate. Following N study episodes, the activation for an item combines the trace strength of individual study episodes. It is

governed by the equation:

$$A(\mathbf{t}, D, B, c) = \log\left(\sum_{j=1}^{N} t_j^{-D}\right) + B + \epsilon, \qquad \epsilon \sim f(x; c)$$

where A is activation, B is a base activation level, ϵ is a noise term drawn from a logistic distribution with mean zero. That is, ϵ has the density function $f(x; c) = \frac{1}{4c}\text{sech}^2\frac{x}{2c}$, where c is a free parameter. Recall probability is related to activation by:

$$\text{P}\left(\text{correct recall} \mid A; \tau, c\right) = \left[1 + \exp\left(\frac{\tau - A}{c}\right)\right]^{-1}$$

where τ is a free parameter. According to the model, latency (RT) is related to activation by:

$$\text{RT}(A, F, f) = Fe^{-fA}$$

where F and f are free parameters. In total, there are six free parameters whose values we must estimate from the data: D, B, c, τ, F, f. Of these, we assume that c, τ, F, f are to be chosen for each subject-item pair, while the trace decay D and base-level activation term B are fixed for each subject.

For each subject-item pair we have a set of study-trial accuracies and latencies, and we can use ACT-R to compute the likelihood of these data for any parameter vector. To do this, we plug the parameters into the equations to generate predictions for study trials and then compare these predictions to actual results of the study trials. More explicitly, we do likelihood-weighted sampling. For a given test subject, we take n_S samples from prior distributions of the six parameters. For each item, we compute the likelihood L of each set of parameters that have been generated. The final prediction of the probability of a correct answer at test is then:

$$\hat{P} = \sum_{i=1}^{n_S} P([D, B, c, \tau, F, f]_i) \; \frac{L([D, B, c, \tau, F, f]_i)}{\sum_{j=1}^{n_S} L([D, B, c, \tau, F, f]_j)}$$

where $\hat{P}$ is the prediction. The likelihood of a set of parameters with respect to a given subject-item pair is given by the product of its likelihood on each study trial:

$$\text{L}(D, B, c, \tau, F, f) = \prod_{i=1}^{n_{trials}} l_{acc}^i l_{RT}^i,$$

where i runs over study trials, and l_{acc} and l_{RT} denote the contribution to the likelihood of the accuracy and response latency. The l_{acc}

$$l^i_{acc} = \begin{cases} P(\text{correct recall}|\hat{A};\tau,c) & \text{if response } i \text{ is accurate} \\ 1 - P(\text{correct recall}|\hat{A};\tau,c) & \text{otherwise} \end{cases}$$

Here, $\hat{A} = A(\mathbf{t}, D, B, c)$.

$$l^i_{RT} = \begin{cases} \frac{1}{4c}\text{sech}^2\frac{\hat{\epsilon}}{2c} & \text{if response } i \text{ is accurate} \\ 1 & \text{otherwise} \end{cases}$$

where $\hat{\epsilon} = \log\left(\frac{\text{RT}^i}{\text{RT}(\hat{A},F,f)}\right)$ and RT^i is the observed latency on the ith study trial. The intuition is that for a given set of parameters, we calculate how much noise would be necessary for these parameters to produce the observed latency and then take the likelihood to be the probability of observing this noise level. We used 250 samples for likelihood-weighted sampling. We found that increasing this number did not noticeably improve performance.

To define priors for the six parameters, we use the fact that the framework above allows us to find, for each subject, maximum likelihood estimates for the parameter values. We do this for a group of training subjects and compile the results in a histogram. We then fit the results for each parameter to a probability distribution which is then that parameter's prior. In practice, the optimization routine we used to do the likelihood maximization did not converge for all subjects. The subjects for which it failed to converge were left out of the calculation of the prior. An example set of histograms used to choose priors is shown in Figure 3.5.

3.2.2 Results

To evaluate the different methods we tried, we used leave-one-out cross-validation. This means that each subject in turn was held out as a test subject, and a prediction for that subject was

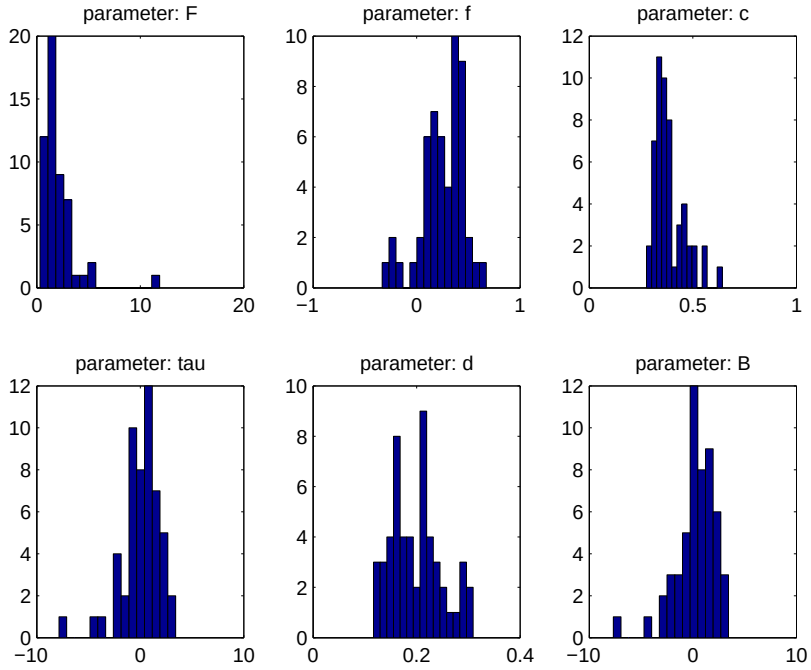


Figure 3.5: A set of histograms used to set the priors for the parameters in BACT-R. To set these priors, we find the maximum likelihood parameter values for each of the subjects in the training group, compile these estimates into histograms, and then fit the data for each parameter to a continuous probability distribution.

made by models trained on all the other subjects. This prediction takes the form of a probability between zero and one. Because the data with which we have to compare these predictions are binary—a subject's response is either correct or incorrect—we threshold the probability so that the predictions also become binary. After thresholding, the models' predictions are either true positive, false positive, true negative, or false negative. Adjusting the threshold changes the number of predictions that fall into each of these categories. In Figures 3.6-**??** (to be described shortly), we summarize the threshold manipulation with an ROC curve, which plots the false positive rate versus the true positive rate for various thresholds. If the ROC curve falls exactly on the dashed diagonal line in the figures, then the method achieves results equivalent to chance prediction. In general, the more bowed the ROC curve, the better the performance of the model.

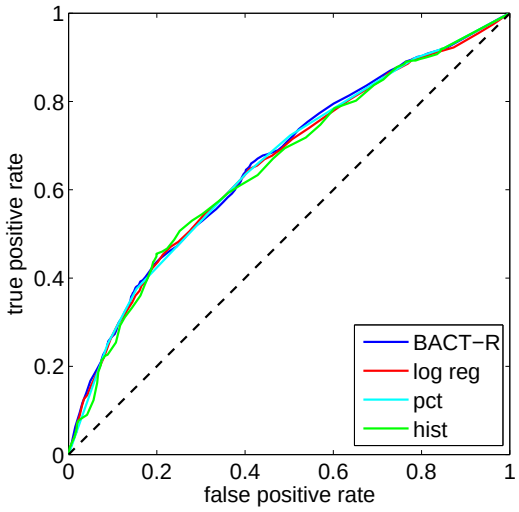


Figure 3.6: ROC curves for the methods we tried. A comparison shows that all methods perform similarly.

3.2.2.1 Comparison of methods

The results obtained by the various methods we tried are shown in Figure 3.6. As this figure shows, all the methods performed almost equally well. In particular, BACT-R did not outperform other methods we tried. It is interesting to note that this implies that the *order* of correct and incorrect responses, which is information to which BACT-R had access and the percentage classifier did not, seems not to have enabled BACT-R to outperform the percentage classifier.

We next examined how much information, if any, is contained in the latency data. Our findings are mixed. On the one hand, logistic regression and BACT-R performed just as well with the latency information removed as with it included (see Figures 3.7 and 3.8, respectively). On the other hand, when provided only with latency information, logistic regression yielded results significantly better than chance (Figure 3.7).

We also examined the weights given by logistic regression to latency and accuracy features. (The inputs to logistic regression are normalized so that it is meaningful to compare the magnitudes of these weights.) The mean magnitudes of the weights for accuracy and latency data are 0.3884 and 0.0751, respectively. The mean weight for the latencies is considerably smaller than the mean

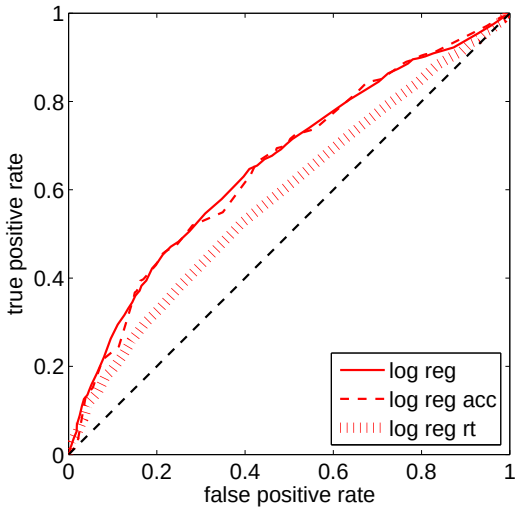


Figure 3.7: ROC curves for logistic regression, when the model was trained with all available data ("log reg"), only accuracy data ("log reg acc"), and only latency data ("log reg RT"). Removing the latency information does not degrade logistic regression's performance. However, using only the latency information gives results that are significantly better than random. We conclude that the latencies contain information, but that this information is redundant with the accuracy information, and does not help with classification.

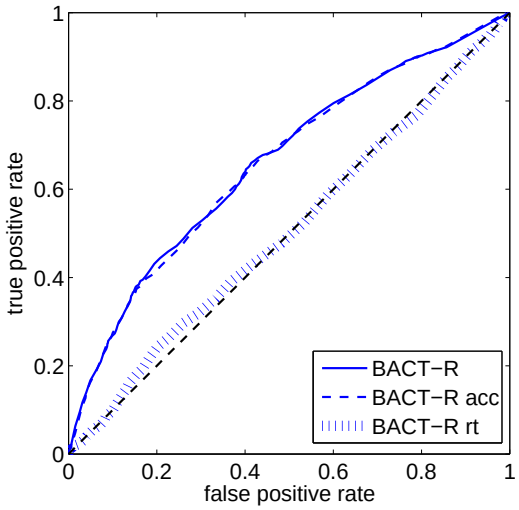


Figure 3.8: ROC curves for BACT-R when the method uses all available data, only the accuracy data, and only latency data. As with logistic regression (Figure 3.7), removing latencies does not noticeably hurt the performance of BACT-R. Using only latencies with BACT-R gives worse performance than it does with logistic regression.

weight for the accuracies; yet, it is not negligible. Thus, there is information in the latencies, but

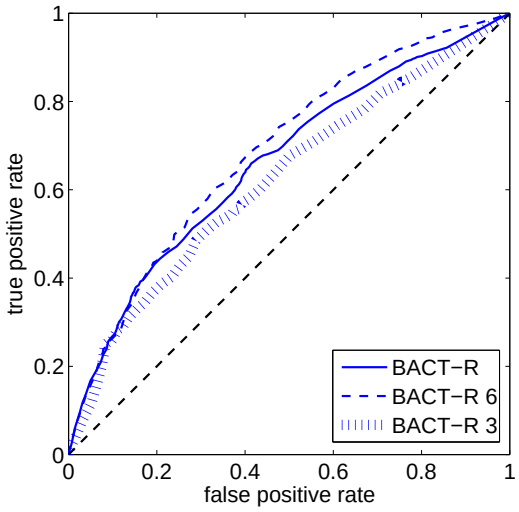


Figure 3.9: This figure compares the performance of BACT-R on the three-trial and six-trial subject-item pairs. Since six study trials give more feedback than three study trials, we expected BACT-R to perform better for these cases. As the figure shows, this is what we observed. Also as expected, we see that the three-study trial cases gave worse performance. However, BACT-R's performance on the three-study trial cases was not sufficiently degraded to conclude that these trials are responsible for BACT-R's inability to outperform the other methods we studied.

it is to a large extent redundant with the information from the accuracies.

The fact that latency information does not improve the performance of our methods may shed some light on the fact that all our methods had more or less equivalent performance: no method took advantage of the latency information; all the information present in the accuracy information reduced to the percentage correct during study. Therefore, all methods did almost exactly as well as the percentage classifier.

3.2.2.2 Number of Study Trials

Figure 3.9 shows the performance of BACT-R when restricted to only the three- or the six-trial study conditions. As expected, BACT-R performed somewhat better with six trials than with three, but the difference is not drastic. This result is significant because it rules out the possibility that the BACT-R's performance was being dragged down by the three-session cases.

Another experiment we did involved applying logistic regression to only the first study session.

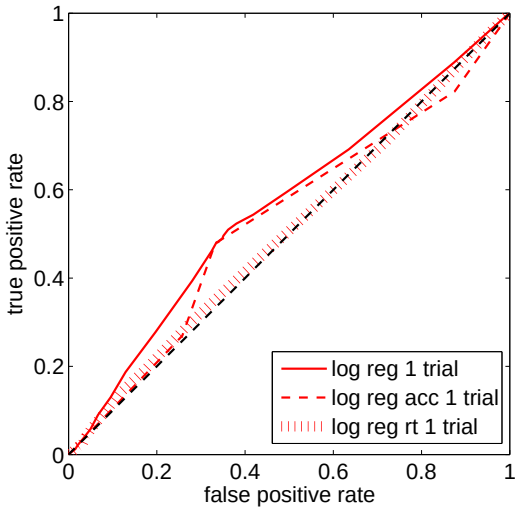


Figure 3.10: ROC curves for logistic regression when this method was applied to data from only the first study trial for each subject-item pair. If we look at only one study trial, we see that using latency information gives a substantial improvement in performance over the model trained with accuracy data alone. We also observe that, when using both pieces of data, we obtain reasonably good prediction performance, even on the basis of only one study trial.

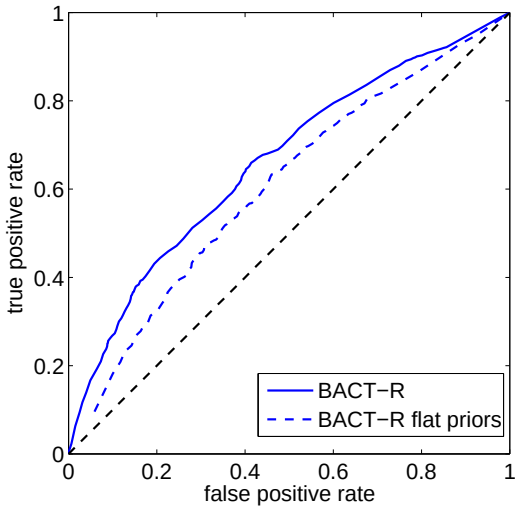


Figure 3.11: This shows the results we obtain if, rather than using the maximum likelihood priors described above, we use flat, uninformative priors. Using the maximum likelihood priors for BACT-R gives substantially better performance than using uniform priors.

In general, we have data from either three or six study trials for each subject-item pair. For this experiment, we used only the first of these. Apart from this, logistic regression was applied in the

same way as before. The motivation for this experiment was the hypothesis that even if the accuracy information dominated the latency information when we used all the trials available, perhaps it would contribute more if we used only one trial. In fact, this was what we observed, as is shown in Figure 3.10, which indicates that, in the one-trial case, adding the latency information to the accuracy information gives a substantial improvement in performance. In addition, we see that it is possible to get reasonably good predictive performance even when we use information from only one trial.

3.2.2.3 Effect of Priors on BACT-R

In order to examine how much information was contained in the priors we used for BACT-R, we tried replacing the priors chosen by maximum likelihood with uniform priors having mean zero and length four. As Figure 3.11 shows, the results were noticeably worse than the results obtained with the maximum likelihood priors. This is a validation of the Bayesian approach, since it shows that the performance of the model was due, at least in part, to the knowledge contained in the prior distributions used for the parameters.

3.2.2.4 Variants

In addition to the methods described above, we tried several variants. For example, we tried replacing raw latencies with z-scores and including latencies from incorrect trials. We also tried assigning greater weight to information from later trials, since these were closer to the test time. No variant we tried significantly altered the performance of the models.

3.2.3 Discussion

The best way for students to learn arbitrary facts is not by rereading the facts, but by testing themselves on the facts. Testing has a side benefit: it produces feedback from the student which potentially could inform an intelligent tutoring system (ITS) about how well the student has learned the facts. In this work, we described an experiment in which feedback was collected from

students learning to identify the disciplines of 60 Nobel Prize winners. This feedback took the form of response accuracy and latency during a study session in which each fact was reviewed multiple times. Using data from the study session, we are able to predict memory for individual facts after a two-week retention interval.

We found that latency data alone was predictive. To the best of our knowledge, this finding has not been reported before. However, we also found that adding latency data to accuracy data did not improve the performance of our models, suggesting that the latency information was redundant with the accuracy information.

We found that all the predictive models had similar performance, including a model based on ACT-R, which is one of the best developed and evaluated high-level theories of human memory. Although BACT-R did not outperform other models, we believe that the addition of Bayesian uncertainty integration to the ACT-R framework is a promising idea that should be explored in other contexts. We also believe that the use of latency information for prediction of future recall warrants further study, especially when the feedback data are sparse (e.g., Figure 3.10, which shows the benefit of latencies when we have feedback from only one trial). Further, it would be interesting to see if the latency information from an experiment specially designed to elicit fast latencies would be more informative than the latencies from this experiment.

In one sense, our conclusions are not astonishing: accuracy of recall during study predicts accuracy of recall at a subsequent test. However, it is important that we have made this intuitively obvious relationship quantitative and that we have explored multiple approaches that can exploit the relationship to make concrete predictions of future recall performance.

3.3 Individualized modeling of forgetting following one study session

Effective teaching requires an understanding of the **knowledge state** of students—what material the student already grasps well, what material can be easily learned, and what material is fragile and likely to be forgotten without additional teaching effort. Based on the knowledge state, individualized teaching policies can be constructed that present highly relevant information and

maximize instructional effectiveness. State-of-the-art software tutors (e.g., J. R. Anderson, Conrad, & Corbett, 1989; K. R. Koedinger & Corbett, 2006; Martin & van Lehn, 1995) incorporate models of the student in order to make inferences about latent state variables. These models are typically expert system based and are constructed through extensive handcrafted analysis of the teaching domain and by means of iterative evaluation and refinement.

We describe a complementary approach to inferring the knowledge state of students that is fully automatic and independent of the content domain. Our approach applies in any domain whose mastery can be decomposed into distinct, separable **elements** of knowledge or **items** to be learned.

What does it mean to infer a student's knowledge state, especially in a domain-independent way? The knowledge state consists of unobservable aspects of a student's cognitive architecture such as the decay rate of a specific declarative memory, the strength of an association, or the boundary of a concept in semantic space. Such representations cannot be validated and therefore have little value except insofar as they can be used to make meaningful predictions. In particular, they have implications for education: being able to **predict** a student's future skill and knowledge. Our work thus focuses on comparing models in terms of the accuracy of their predictions.

Given inter-student differences, the abilities of a particular student cannot be determined without sufficient experience teaching that student; given inter-item differences, the challenge of a new item cannot be determined without sufficient experience teaching that item. By the point at which this experience is acquired, it may be too late for it to be useful in teaching. We propose a solution to this dilemma that leverages a **population** of students learning a **population** of items to make inferences concerning the knowledge state of **individual** students for **specific** items. (We refer to this pair as a **student-item**.)

Our approach is a form of collaborative filtering in which we predict whether a student who has mastered items X and Y will likely have mastered Z, based on the performance of other students for the same item and the performance of that student for other items. This approach fundamentally needs to address the dynamic nature of latent knowledge states. Dynamics differentiate our task

from canonical collaborative-filtering tasks (e.g., movie preference prediction) in three respects. First, canonical tasks require predictions only about the present, but effective teaching requires predictions about the future performance of a student in order to select appropriate material at the present. Second, canonical tasks may allow for nonstationarity—for example, a change in movie preferences over time—but as we argued earlier, the current knowledge state is causally dependent on the distribution, frequency, and type of past study. Third, canonical tasks make predictions (e.g., about whether Fred will like the movie **Borat**) without any direct past evidence from Fred about **Borat**, whereas in learning scenarios, each student typically has a history of encountering and being evaluated on a specific fact, skill, or concept in the past.

3.3.1 Models for predicting student performance

Our work is based on item-response theory (IRT), the classic psychometric approach to inducing latent traits of students and items based on exam scores (De Boeck & Wilson, 2004). Whereas IRT assumes static states of knowledge, we are concerned with states that depend on the temporal history of study. We thus propose novel models that incorporate this history and in general better embody the dynamics of student learning and retention.

3.3.1.1 Item response theory (IRT)

Among other applications, 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:

$$Pr(R_{si} = 1) = \frac{1}{1 + \exp(d_i - a_s)}. \tag{3.1}$$

(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. We explored the 2PL model, but found for our data sets that it was indistinguishable from the 1PL model.)

The free parameters of IRT are typically fit by maximum likelihood. Bayesian variants of IRT have been proposed that allow for additional knowledge in the form of hierarchical priors over student ability and item difficulty (Fox, 2010).

IRT is generally used to analyze tests and surveys post hoc, in order to evaluate the diagnosticity of test items and the skill level of students (Roussos et al., 2007). Extensions have been proposed to allow for a student to have a different ability at different times (Andrade & Tavares, 2005), but plenty of opportunity remains to explore dynamic variants of IRT that predict future performance of students, integrate the longitudinal history of study, and, instead of directly predicting behavioral outcomes, do so through latent knowledge state variables (such as memory decay rate or concept boundaries). We take first steps in this direction by incorporating the latent traits of IRT into a theory of forgetting.

3.3.1.2 Theories of forgetting

Psychologists have spent well over a century analyzing the temporal characteristics of learning and memory. The modern consensus is when a set of materials are learned in a single study session and then tested following some lag t, the probability of recalling the studied material decays according to a generalized power-law function of t,

$$Pr(\text{recall}) = m(1 + ht)^{-f}, \tag{3.2}$$

where $0 \leq m \leq 1$ is the degree of learning, $h > 0$ is a scaling factor on time, and $f > 0$ is the memory decay exponent (Wixted & Carpenter, 2007).

The form of this curve is supported by data from populations of students and/or populations of items. The forgetting curve cannot be measured for a single student-item due to the observer effect and the all-or-none nature of forgetting, but we will assume the functional form of the curve for a student-item is the same. However, 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, 2005), we could individuate forgetting for each student-item by setting the decay exponent based on latent IRT-like traits:

$$Pr(R_{si} = 1) = m(1 + ht_{si})^{-\exp(\tilde{a}_s - \tilde{d}_i)}, \tag{3.3}$$

where t_{si} denotes the **retention interval**—the time between initial presentation of item i to student s and a later recall test. We have added 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 3.1, and using $f \equiv \exp(\tilde{a}_s - \tilde{d}_i)$ ensures that f remains nonnegative.

Another alternative we consider is individuating the degree-of-learning parameter instead of d. This gives the model

$$Pr(R_{si} = 1) = \frac{(1 + ht_{si})^{-f}}{1 + \exp(d_i - a_s)}. \tag{3.4}$$

As a final alternative, we can individuate both the forgetting parameter f and degree-of-learning parameter m. This yields a hybrid model:

$$Pr(R_{si} = 1) = \frac{(1 + ht_{si})^{-\exp(\tilde{a}_s - \tilde{d}_i)}}{1 + \exp(d_i - a_s)}. \tag{3.5}$$

Both this hybrid model and Equation 3.4 simplify to 1PL (Equation 3.1) at $t = 0$. For $t > 0$, recall probability decays as a power-law function of time.

3.3.1.3 A space of models to explore

We explored five models whose probability of recall for individual student-items was determined by the models presented in Equations $1 - 5$:

- IRT: the 1PL IRT model (Equation 3.1);
- MEMORY: a power-law forgetting model with population-wide parameters (Equation 3.2);
- HYBRID DECAY: a power-law forgetting model with decay rates based on latent student and item traits (Equation 3.3);
- HYBRID SCALE: a power-law forgetting model with the degree-of-learning based on latent student and item traits (Equation 3.4); and
- HYBRID BOTH: a power-law forgetting model that individuates both the decay rate and degree-of-learning (Equation 3.5).

Each of these models was trained in one of two ways: (1) using maximum likelihood (ML) fits of model parameters to the training data, and (2) using a hierarchical Bayesian approach (BAYES) that makes weak distributional assumptions about the parameters (Table 3.1). Inference on the two sets of latent traits in the HYBRID BOTH model—$\{a_s\}$ and $\{d_i\}$ from 1PL, $\{\tilde{a}_s\}$ 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 the 1PL 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 1PL variables, and allows the corrupting influence of static factors to be removed from the forgetting-related variables.

3.3.1.4 Simulation methodology

We employed Markov chain Monte Carlo techniques for posterior inference in the Bayesian models presented in Table 3.1. Gibbs sampling is not feasible in our models, but we can use Metropolis-within-Gibbs (Patz & Junker, 1999), an extension of Gibbs sampling wherein each

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

Table 3.1: 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.

Study name	$\mathcal{S}_1$	$\mathcal{S}_2$
Source	(S. H. K. Kang et al., 2014)	(Pashler et al., unpublished)
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 3.2: Experimental data used for simulations

draw from the model's full conditional distribution is performed by a single Metropolis-Hastings step.

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 term (defined in the previous section) 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

$$p(a_s \mid \mathbf{a}_{\neg \mathbf{s}}, \mathbf{d}, h, m, R) \propto \prod_i P(r_{si} \mid a_s, d_i, h, m) \times \left(1 + \frac{a_s^2}{2(\psi_2 + \frac{1}{2}\sum_{j \neq s} a_j)}\right)^{\psi_1 + \frac{n_S - 1}{2}} \tag{3.6}$$

where the first term is given by Equation 3.4. 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.

For the maximum likelihood models, we found fits using standard gradient-based nonlinear optimization techniques (Matlab's **fminunc** function). To find a fit, we ran the optimization method with five randomized starting locations and took the best solution.

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.

3.3.2 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 3.2. 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 to produce the appropriate response to each cue. Whether the student was successful or not, the correct response was then displayed. Following training and a delay t_{si} that was specific to each student and each item, an exam was administered, obtaining the r_{si} binary value for that student-item.

To evaluate the models, we performed 50-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 of recall probability at the exam given t_{si}, conditioned on the training data,

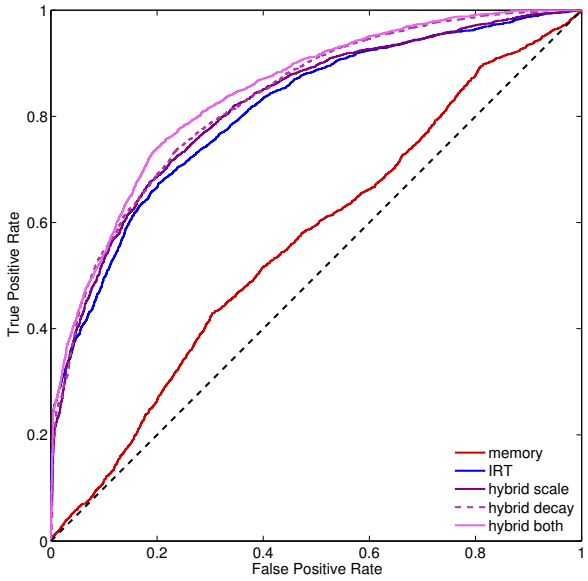


Figure 3.12: Mean ROC curves for the Bayesian models on held-out data from Study $\mathcal{S}_1$.

which can be compared against the held-out data. Each model's ability to discriminate successful and unsuccessful recall trials was assessed with a signal-detection analysis technique (Green & Swets, 1966).

Figure 3.12 shows the ROC curves for Study $\mathcal{S}_1$ for the Bayesian versions of the models. Each curve is the mean across validation folds for a particular model. The area under the ROC curve (hereafter, **AUC**) is a measure of the model's predictive ability: the more bowed the curve, the better the model is at predicting a particular student's recall success on a specific item after a given lag. The figure includes the models described earlier, including the baseline IRT model that ignores the time lag between study and test, and the baseline MEMORY model that assumes power law forgetting but assumes parameters of the power function that are independent of the student and the item.

The top panel of Figure 3.13 summarizes the AUC values for Study $\mathcal{S}_1$. The baseline MEMORY model is trounced by the other models ($p < .01$ for all pairwise comparisons with MEMORY by a two-tailed t test), suggesting that the other models have successfully recovered latent student and item traits that can be used to improve inference about the knowledge state of a particular student-

item. Though performance is high for all the non-baseline models, the HYBRID BOTH model does better than its peers.

The middle panel of Figure 3.13 presents the AUC values for Study $\mathcal{S}_2$. These results are consistent with our findings for $\mathcal{S}_1$. First, MEMORY fails to predict as well as any of the models that accommodate individual differences ($p < .01$ for all pairwise comparisons with MEMORY by a two-tailed t test). Second, the HYBRID BOTH model outperforms the other models. This suggests that allowing for individual differences both in degree of learning and rate of forgetting is appropriate even on the short timescale of Study $\mathcal{S}_2$.

The ML models are compared to the BAYES models in the bottom panel of Figure 3.13 for study $\mathcal{S}_1$. For the IRT and MEMORY models, BAYES provides no benefit. However, HYBRID BOTH BAYES yields significantly better discrimination than HYBRID BOTH ML ($p < .01$ by paired t test). In the Bayesian models, ability parameters of each student s, a_s and $\tilde{a}_s$, are constrained by the distribution of abilities of the other students, via a hierarchical prior; likewise, the difficulty parameters of each item i, d_i and $\tilde{d}_i$, are similarly constrained by their population distributions. These constraints bias inference in the right direction so long as assumptions concerning the qualitative shape of the population distributions are appropriate. The two findings we have presented—(1) that systematic individual (student and item) differences exist that can be used for predicting knowledge state, and (2) that the population distributions are useful for prediction—are not incompatible.

3.3.2.1 Generalization to new material

The previous simulations held 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

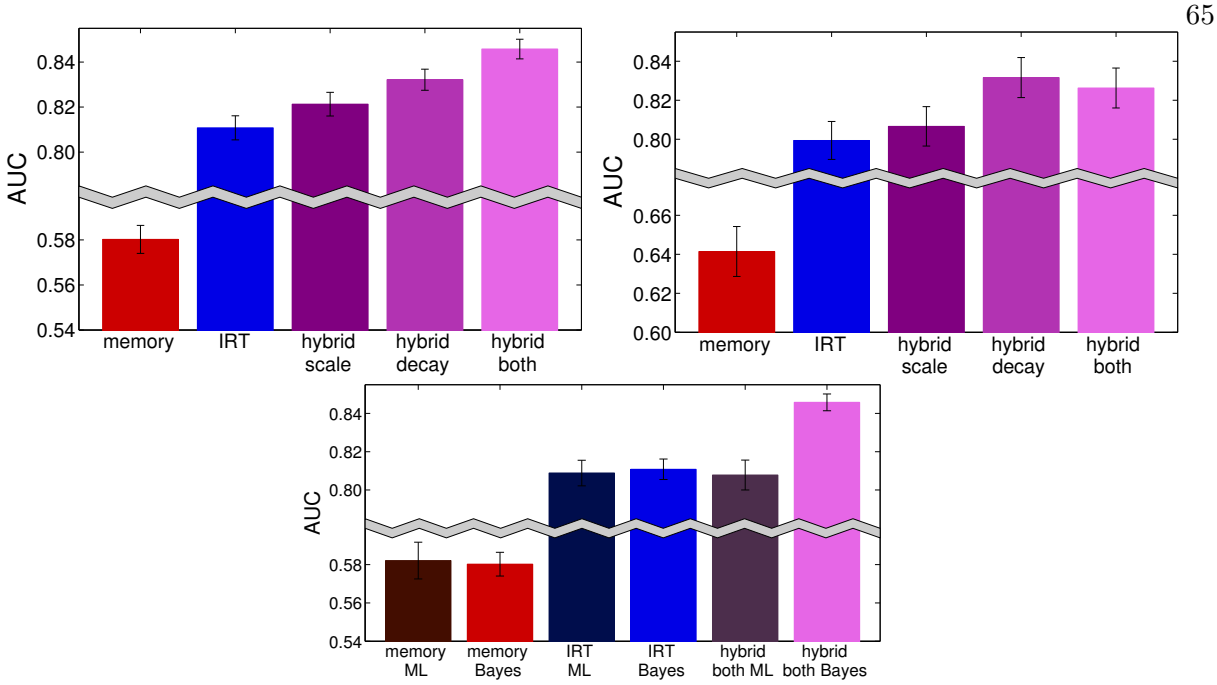


Figure 3.13: The top left and top right graphs show mean AUC values on the five BAYES models trained and evaluated on Studies $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. The bottom graph compares BAYES and ML versions of three models on Study $\mathcal{S}_1$. The error bars indicate a 95% confidence 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.

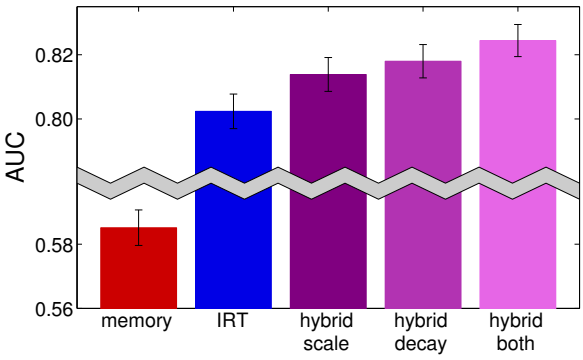


Figure 3.14: Mean AUC values when random items are held out during validation folds, Study $\mathcal{S}_1$

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?

To model this situation, we conducted further validation tests in which, instead of holding

out random student-item pairs, we held out random items for all students. Figure 3.14 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.

3.3.3 Discussion

Psychological models of human memory have typically been used to characterize the aggregate performance of a population of students learning a collection of items (Pavlik & Anderson, 2005). Psychometric models of individual differences have been used to recover static latent characteristics of students and items. We have shown that by combining a dynamical model of human memory with a static latent-state model of individual differences, we can significantly improve predictions about the performance of individual students for specific items. Via collaborative filtering, we recover information about the time-varying unobservable knowledge state of a particular student for specific material by leveraging data collected from populations of students and collections of material. Our approach has enormous potential to improve electronic tutoring systems, which rely on accurate models of student knowledge state to tailor instruction to the needs of individuals.

3.4 Individualized modeling of forgetting following multiple study sessions

To personalize review in electronic tutoring systems wherein students study material across multiple study sessions, we must infer a student's *knowledge state*—the dynamically varying strength of each atomic component of knowledge (KC) as the student learns and forgets. Knowledge-state inference is a central concern in fields as diverse as educational assessment, intelligent tutoring systems, and long-term memory research. We briefly resummarize two contrasting approaches taken in the literature, *data driven* and *theory driven*, and propose a synthesis then propose a synthesis combining the power of the two approaches. We refer the reader to chapter 2 for a more thorough discussion of the two approaches.

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

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

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

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

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

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

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

The psychological plausibility of MCM and ACT-R is demonstrated through fits of the models to behavioral data from laboratory studies of spaced review. Because minimizing the number of free

parameters is key to a compelling account, cognitive models are typically fit to aggregate data—data from a population of students studying a body of material. They face a serious challenge in being useful for modeling the state of a particular KC for a particular student: a proliferation of parameters is needed to provide the flexibility to characterize different students and different types of material, but flexibility is an impediment to making strong predictions.

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

$$\Pr(R_{si} = 1 \mid \alpha_s, \delta_i, \boldsymbol{\phi}, \boldsymbol{\psi}) = \text{logistic}\Big[\alpha_s - \delta_i + \sum_w \phi_w \log(1 + c_{siw}) - \psi_w \log(1 + n_{siw})\Big], \qquad (3.7)$$

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

We will explain the selection of time windows shortly, but we first provide an intuition for the specific form of the factors. The difference of factors inside the summation of Equation 3.7 determines a power law of practice. Odds of correct recall improve as a power function of: the number of correct trials with $\phi_w > 0$ and $\psi_w = 0$, the number of study trials with $\psi_w < 0$ and $\phi_w = 0$, and the proportion of correct trials with $\phi_w = \psi_w$. The power law of practice is a ubiquitous property of human learning incorporated into ACT-R. Our two-parameter formulation allows for a

wide variety of power function relationships, from the three just mentioned to combinations thereof. The formulation builds a bias into DASH that additional study in a given time window helps, but has logarithmically diminishing returns. To validate the form of DASH in Equation 3.7, we fit a single-window model to data from the first week of our experiment, predicting performance on the end-of-chapter quiz for held-out data. We verified that Equation 3.7 outperformed variations of the formula which omitted one term or the other or which expressed log-odds of recall directly in terms of the counts instead of the logarithmic form.

To model effects of temporally distributed study and forgetting, DASH includes multiple time windows. Window-specific parameters (ψ_w, ϕ_w) encode the dependence between recall at the present moment and the amount and outcome of study within the window. Motivated by theories of memory, we anchored all time windows at the present moment and varied their spans such that the temporal span of window w, denoted s_w, increased with w. We chose the distribution of spans such that there was finer temporal resolution for shorter spans, i.e., $s_{w+2} - s_{w+1} > s_{w+1} - s_w$. This distribution allows the model to efficiently represent rapid initial forgetting followed by a more gradual memory decay, which is a hallmark of the ACT-R power-function forgetting. This distribution is also motivated by the overlapping time scales of memory in MCM. ACT-R and MCM both suggest the elegant approach of exponentially expanding time windows, i.e., $s_w \propto \mathrm{e}^{\rho w}$. Lindsey et al. (2014) roughly followed this suggestion, with three caveats. First, we did not try to encode the distribution of study on a very fine scale—less than an hour—because the fine-scale distribution is irrelevant for retention intervals on the order of months (Cepeda et al., 2006, 2008). Second, we wished to limit the number of time scales so as to minimize the number of free parameters in the model to prevent overfitting and to allow for sensible generalization early in the semester when little data existed for long-term study. Third, we synchronized the time scales to the natural periodicities of student life. Taking these considerations into account, we chose five time scales: $\mathbf{s} = \{1/24, 1, 7, 30, \infty\}$.

3.4.1 Other models that consider time

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

3.4.2 Hierarchical distributional assumptions

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

$$\begin{aligned} \alpha_s &\sim \text{Normal}(\mu_\alpha, \sigma_\alpha^2) \\ (\mu_\alpha, \sigma_\alpha^{-2}) &\sim \text{Normal-Gamma}(\mu_0^{(\alpha)}, \kappa_0^{(\alpha)}, a_0^{(\alpha)}, b_0^{(\alpha)}) \\ \delta_i &\sim \text{Normal}(\mu_\delta, \sigma_\delta^2) \\ (\mu_\delta, \sigma_\delta^{-2}) &\sim \text{Normal-Gamma}(\mu_0^{(\delta)}, \kappa_0^{(\delta)}, a_0^{(\delta)}, b_0^{(\delta)}) \end{aligned} \tag{3.8}$$

where the Normal-Gamma distribution has parameters $\mu_0, \kappa_0, a_0, b_0$. Individual ability parameters α_s are drawn independently from a normal distribution with unknown population-wide mean μ_α

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

3.4.3 Gibbs-EM inference algorithm

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

$$P(R_{sik} = 1 \mid \alpha_s, \delta_i, \mathbf{t}_{1:k}, \mathbf{r}_{1:k-1}, \boldsymbol{\theta}) = \sigma(\alpha_s - \delta_i + h_{\boldsymbol{\theta}}(\mathbf{t}_{s,i,1:k}, \mathbf{r}_{s,i,1:k-1})) \tag{3.9}$$

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

$$h_{\boldsymbol{\theta}} = \sum_{w=0}^{W-1} \theta_{2w+1} \log(1 + c_{si,w+1}) + \theta_{2w+2} \log(1 + n_{si,w+1}) \tag{3.10}$$

where the summation is over W time windows.

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

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \iint P(\mathbf{r}|\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\theta}) p(\boldsymbol{\alpha}) p(\boldsymbol{\delta}) \; d\boldsymbol{\alpha} \; d\boldsymbol{\delta} \tag{3.11}$$

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

(1) Initialize $\boldsymbol{\theta}^{(0)}$ and set $i = 1$

(2) Iteration i

- **E-step:** Draw N samples $\left\{\boldsymbol{\alpha}^{(\ell)}, \boldsymbol{\delta}^{(\ell)}\right\}_{\ell=1}^{N}$ from $p(\boldsymbol{\alpha}, \boldsymbol{\delta} \mid \mathbf{r}, \boldsymbol{\theta}^{(i-1)})$ using a Gibbs sampler
- **M-step:** Find

$$\boldsymbol{\theta}^{(i)} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{\ell=1}^{N} \log P(\mathbf{r}, \boldsymbol{\alpha}^{(\ell)}, \boldsymbol{\delta}^{(\ell)} | \boldsymbol{\theta}) \tag{3.12}$$

(3) $i \leftarrow i + 1$, go to 2 if not converged.

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

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

$$\begin{aligned} p(\alpha_s \mid \boldsymbol{\alpha}_{\neg s}, \boldsymbol{\delta}, \boldsymbol{\theta}, \mathbf{r}) &\propto p(\alpha_s \mid \boldsymbol{\alpha}_{\neg s}) \prod_{i,k} P(r_{sik} \mid \alpha_s, \delta_i, \boldsymbol{\theta}) \\ p(\delta_i \mid \boldsymbol{\delta}_{\neg i}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \mathbf{r}) &\propto p(\delta_i \mid \boldsymbol{\delta}_{\neg i}) \prod_{s,k} P(r_{sik} \mid \alpha_s, \delta_i, \boldsymbol{\theta}) \end{aligned} \tag{3.13}$$

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

M-Step. Let S be the number of students, I be the number of items, and n_{si} be the number of trials undergone by student s on item i. By assumption, the hyperparameters of the normal-gamma

distributions are not part of $\boldsymbol{\theta}$. Thus, the M-step is equivalent to finding the hyperparameters which maximize the expectation of the data log-likelihood,

$$\boldsymbol{\theta}^{(i)} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{\ell=1}^{N} \log P(\mathbf{r} \mid \boldsymbol{\alpha}^{(\ell)}, \boldsymbol{\delta}^{(\ell)}, \boldsymbol{\theta}) \tag{3.14}$$

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

$$\mathcal{L}^{(\ell)} = \sum_{s=1}^{S} \sum_{i=1}^{I} \sum_{k=1}^{n_{si}} r_{sik}\gamma^{(\ell)} - \log\left(1 + e^{\gamma^{(\ell)}}\right) \tag{3.15}$$

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

$$\frac{\partial \mathcal{L}^{(\ell)}}{\partial \theta_j} = \sum_{s=1}^{S} \sum_{i=1}^{I} \sum_{k=1}^{n_{si}} (r_{sik} - \sigma(\gamma^{(\ell)})) \frac{\partial h}{\partial \theta_j} \tag{3.16}$$

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

$$\frac{\partial^2 \mathcal{L}^{(\ell)}}{\partial \theta_z \theta_j} = \sum_{s=1}^{S} \sum_{i=1}^{I} \sum_{k=1}^{n_{si}} (r_{sik} - \sigma(\gamma^{(\ell)})) \frac{\partial^2 h}{\partial \theta_z \partial \theta_j} - \sigma(\gamma^{(\ell)})(1 - \sigma(\gamma^{(\ell)})) \frac{\partial h}{\partial \theta_z} \frac{\partial h}{\partial \theta_j} \tag{3.17}$$

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

3.4.4 Simulation results

This section describes the procedure used to evaluate the models. The models were trained on all data up to a given point in time on the 597,990 retrieval practice trials recorded across a semester-long experiment (Lindsey et al., 2014) involving a population of students studying a set of study items over time (described in Chapter 4). We divided these time-ordered trials into contiguous segments with each segment containing 1% of the trials. We then tested each model's ability to predict a segment n given segments $1 \ldots n-1$ as training data, for $n \in \{2 \ldots 100\}$. We scored each model's across-segment average prediction quality using cross entropy[1] and mean per-trial prediction error[2] . The former method more strongly penalizes held-out trials for which the model assigned low probability to the observed recall event.

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

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

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

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

- ACT-R. Pavlik and Anderson (Pavlik & Anderson, 2005, 2008) extended the ACT-R memory model to account for the effects of temporally distributed study; we will refer to their model as ACT-R. The model includes parameters similar to the ability and difficulty factors in IRT that characterize individual differences among students and among KCs. Further, the model allows for parameters that characterize each student-KC pair. Whereas DASH is fully specified by eight parameters,[3] the number of free parameters in the ACT-R model increases multiplicatively with the size of the student pool and amount of study material. To fit the data recorded in this experiment, the model requires over forty thousand free parameters, and there are few data points per parameter. Fitting such a high-dimensional and weakly constrained model is an extremely challenging problem. Pavlik and Anderson had the sensible idea of inventing simple heuristics to adapt the parameters as the model is used. We found that these heuristics did not fare well for our experiment. Therefore, in our simulation of ACT-R, we eliminated the student-KC specific parameters and used Monte

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

Carlo maximum likelihood estimation, which is a search method that repeatedly iterates through all the model parameters, stochastically adjusting their values so as to increase the data log-likelihood.[4]

- IRT. We created a hierarchical Bayesian version of the Rasch Item-Response Theory model with the same distributional assumptions over $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ as made in DASH. We will refer to this model as IRT. It corresponds to the assumption that $h_{\boldsymbol{\theta}} = 0$ in Equation 3.9.

- DASH[ACT-R]. We experimented with a version of DASH which does not have a fixed number of time windows, but instead—like ACT-R—allows for the influence of past trials to continuously decay according to a power-law. Using the DASH likelihood equation in Equation 3.9, the model is formalized as

$$h_{\boldsymbol{\theta}} = c \log(1 + \sum_{k'<k} m_{r_{k'}} t_{k'}^{-d}) \tag{3.18}$$

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

- DASH[MCM]. Motivated by the Multiscale Context Model (MCM), a model of the spacing effect we developed which has a fixed set of continuously, exponentially decaying memory traces (Mozer et al., 2009), we experimented with a version of DASH which has a fixed

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

number of continuously decaying windows. The model assumes that the counts n_{siw} and c_{siw} are incremented at each trial and then decay over time at a window-specific exponential rate τ_w. Formally,

$$h_{\boldsymbol{\theta}} = \sum_{w=0}^{W-1} \theta_{2w+1} \log(1 + \tilde{c}_{si,w+1}(t)) + \theta_{2w+2} \log(1 + \tilde{n}_{si,w+1}(t)) \tag{3.19}$$

where

$$\tilde{n}_{siw}^{(k)} = 1 + \tilde{n}_{siw}^{(k-1)} \exp(-\tfrac{t_k - t_{k-1}}{\tau_w}) \qquad \tilde{c}_{siw}^{(k)} = r_{sik} + \tilde{c}_{siw}^{(k-1)} \exp(-\tfrac{t_k - t_{k-1}}{\tau_w}) \tag{3.20}$$

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

* The smallest half-life should be about 30 minutes, roughly the time between COLT prediction updates. Thus, $t_1^{(1/2)} = .0208$ and so $\tau_1 = .0301$.
* The largest half-life should be about the length of the experiment. Thus, $t_W^{(1/2)} = 90$ and so $\tau_W = 129.8426$.
* The half-lives should be exponentially increasing. It is important to be able to differentiate between, for example, whether a trial is 1 or 2 days old. Differentiating between, for example, trials that are 60 vs. 61 days old is less important. Thus, we want $t_w^{(1/2)} = ct_{w-1}^{(1/2)}$ where c is a constant.

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

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

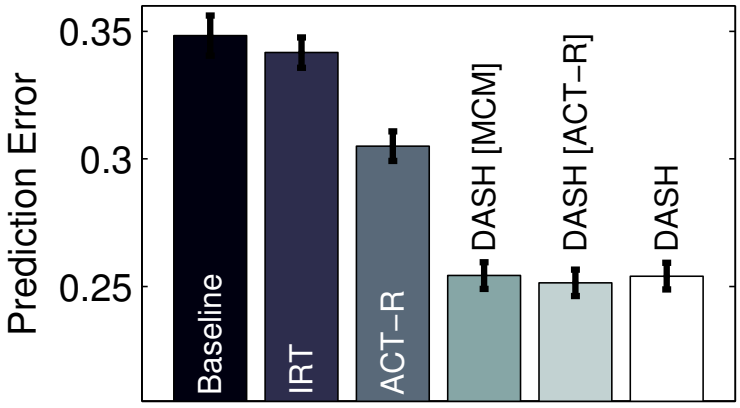


Figure 3.15: Accumulative prediction error of DASH and five alternative models using the data from the semester-long experiment. Error bars indicate ±1 standard error of the mean.

For ACT-R, we ran 1500 iterations of the stochastic hill-climbing algorithm and kept the maximum likelihood solution.

Figure 3.15 compares DASH against the five alternatives: a *baseline* model that predicts a student's future performance to be the proportion of correct responses the student has made in the past, a Bayesian form of *item-response theory* (IRT) (De Boeck & Wilson, 2004), a model of spacing effects based on the memory component of ACT-R (Pavlik & Anderson, 2005), and two variants of DASH that incorporate alternative representations of study history motivated by models of spacing effects (ACT-R, MCM).

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.

Chapter 4

Improving students' long-term knowledge retention through personalized review

Human memory is imperfect; thus, periodic review is required for the long-term preservation of knowledge and skills. However, students at every educational level are challenged by an evergrowing amount of material to review and an ongoing imperative to master new material. We developed a method for efficient, systematic, personalized review that combines statistical techniques for inferring individual differences with a psychological theory of memory. In the first of three experiments, 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.

4.1 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 (Custers, 2010), and 80–85% after 25 years (Custers & Ten Cate, 2011).

Forgetting is influenced by the temporal distribution of study. For over a century, psychologists have noted that temporally spaced practice leads to more robust and durable learning than massed practice (Cepeda et al., 2006). Although spaced practice is beneficial in many tasks beyond rote memorization (Kerfoot et al., 2010) and shows promise in improving educational outcomes (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013), the reward structure of academic programs seldom provides an incentive to methodically revisit previously learned material. Teachers commonly introduce material in sections and evaluate students at the completion of each section; consequently, students' grades are well served by focusing study exclusively on the current section. Although optimal in terms of students' short-term goals, this strategy is costly for the long-term goal of maintaining accessibility of knowledge and skills. Other obstacles stand in the way of incorporating distributed practice into the curriculum. Students who are in principle willing to commit time to review can be overwhelmed by the amount of material, and their metacognitive judgments about what they should study are likely to be unreliable (Nelson & Dunlosky, 1991; Zechmeister & Shaughnessy, 1980). Moreover, though teachers recognize the need for review, the time demands of restudying old material compete against the imperative to regularly introduce new material.

4.2 Main Experiment

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

For each chapter of course material, students engaged in three 20–30 minute sessions with

COLT during class time. The first two sessions began with a study-to-proficiency phase for the current chapter and then proceeded to a review phase. On the third session, these activities were preceded by a quiz on the current chapter, which counted toward the course grade. During the review phase, study items from all chapters covered so far in the course were eligible for presentation. Selection of items was handled by three different 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 M. S. Cohen, Yan, Halamish, and Bjork (2013). [1]

A *generic-spaced* scheduler selected one previous chapter to review at a spacing deemed to be optimal for a range of students and a variety of material according to both empirical studies (Cepeda et al., 2006, 2008) and computational models (Khajah, Lindsey, & Mozer, 2014; 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 chapter, giving priority to the least recently studied (Figure 4.1).

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

[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.

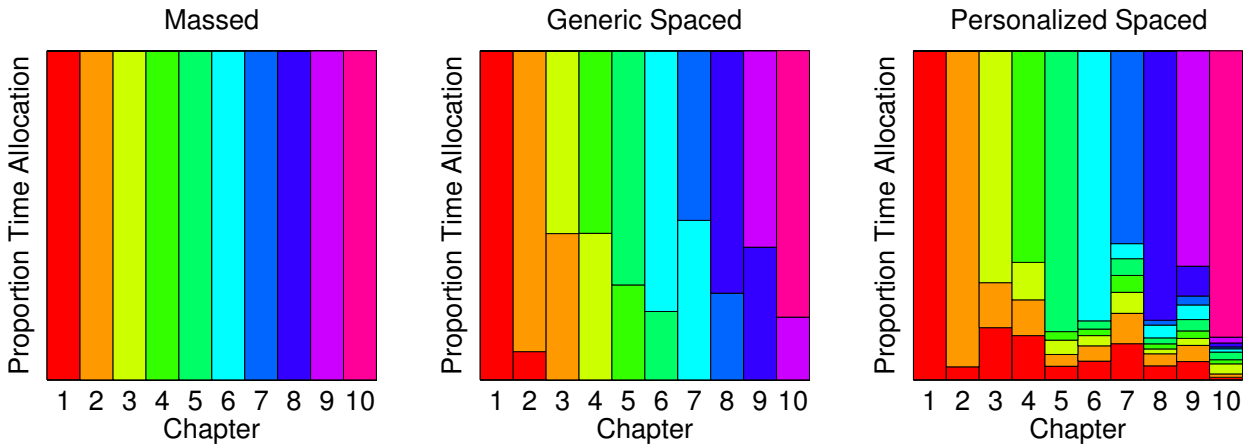


Figure 4.1: Time allocation of the three review schedulers. Course material was introduced one chapter at a time, generally at one-week intervals. Each vertical slice indicates the proportion of time spent in a week studying each of the chapters introduced so far. Each chapter is indicated by a unique 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.

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, DASH characterizes both individual differences and the temporal dynamics of learning and forgetting. Chapter 3 describes DASH in detail.

The scheduler was varied within participant by randomly assigning one third of a chapter's items to each scheduler, counterbalanced across participants. During review, the schedulers alter-

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

Table 4.1: Presentation statistics of individual student-items over entire experiment

nated 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 4.1 and Table 4.1 present student-item statistics for each scheduler over the time course of the experiment.

4.2.1 Results

Two proctored cumulative exams were administered to assess retention, one at the semester's end and one 28 days later, at the beginning of the following semester. Each exam tested half of the course material, randomized for each student and balanced across chapters and schedulers; no corrective feedback was provided. On the first exam, the personalized spaced scheduler improved retention by 12.4% over the massed scheduler ($t(169) = 10.1$, $p < .0001$, Cohen's $d = 1.38$) and by 8.3% over the generic spaced scheduler ($t(169) = 8.2$, $p < .0001$, $d = 1.05$) (Figure 4.2, upper). 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 4.2, lower), 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.

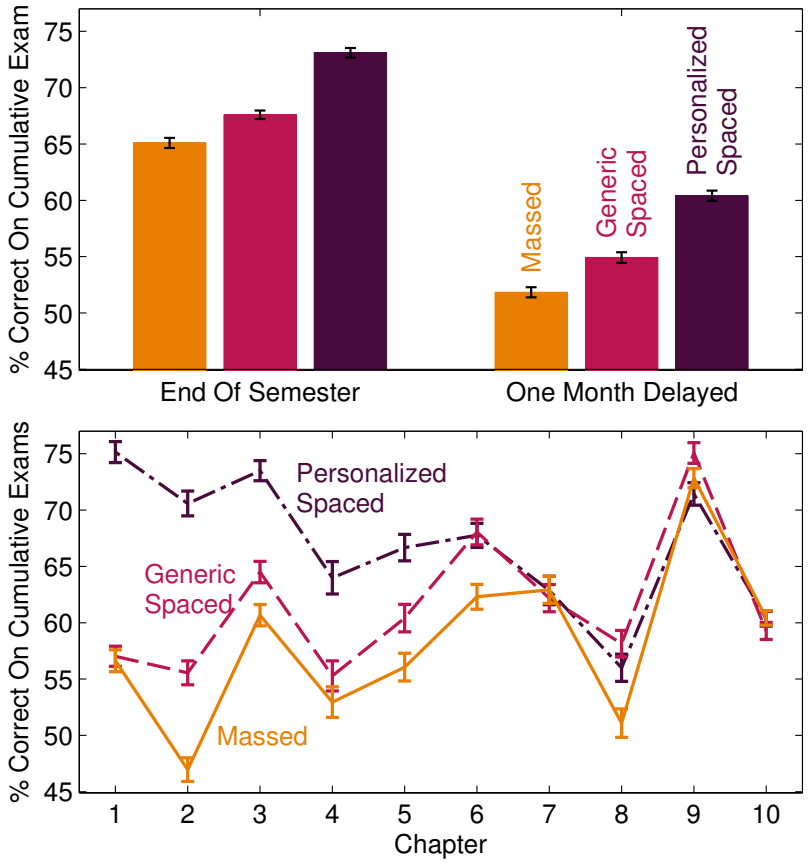


Figure 4.2: `(upper)` Mean scores on the two cumulative end-of-semester exams, taken 28 days apart. `(lower)` Mean score of the two exams as a function of the chapter in which the material was introduced. The personalized-spaced scheduler produced a large benefit for early chapters in the semester and did so without sacrificing efficacy on later chapters. All error bars indicate ± 1 within-student standard error (Masson & Loftus, 2003).

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 4.3), 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. As described in section 3.4.4, to evaluate the

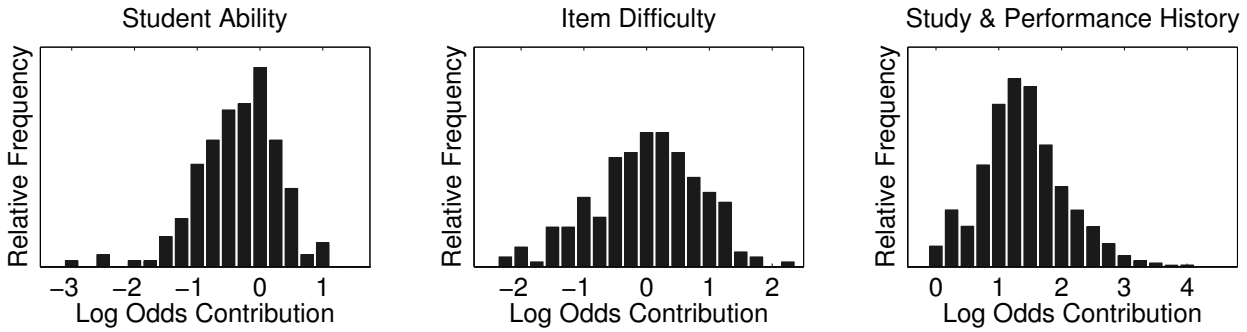


Figure 4.3: Histogram of three sets of inferred factors, expressed in their additive contribution to predicted log-odds of recall. Each factor varies over three log units, corresponding to a possible modulation of recall probability by 0.65.

quality of DASH's predictions, we compared DASH against alternative models by dividing the 597,990 retrieval practice trials recorded over the semester into 100 temporally contiguous disjoint sets, and the data for each set was predicted given the preceding sets. The *accumulative prediction error* (Wagenmakers, Grünwald, & Steyvers, 2006) was computed using the mean deviation between the model's predicted recall probability and the actual binary outcome, normalized such that each student is weighted equally. Figure 3.15 compares DASH against five alternatives: a *baseline* model that predicts a student's future performance to be the proportion of correct responses the student has made in the past, a Bayesian form of *item-response theory* (IRT) (De Boeck & Wilson, 2004), a model of spacing effects based on the memory component of ACT-R (Pavlik & Anderson, 2005), and two variants of DASH that incorporate alternative representations of study history motivated by models of spacing effects (ACT-R, MCM).

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.

4.2.2 Discussion

Our work builds on the rich history of applied human-learning research by integrating two distinct threads: classroom-based studies that compare massed versus spaced presentation of material (Carpenter, Pashler, & Cepeda, 2009; Seabrook, Brown, & Solity, 2005; Sobel, Cepeda, & Kapler, 2011), and laboratory-based investigations of techniques that select material for an individual to study based on that individual's past study history and performance, known as *adaptive scheduling* (e.g., R. C. Atkinson, 1972; Leitner, 1972; Woziak & Gorzelanczyk, 1994).

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 by Budé, Imbos, van de Wiel, & Berger, 2011, 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 (Metzler-Baddeley & Baddeley, 2009; Pavlik & Anderson, 2008; van Rijn et al., 2009), but these evaluations have been confined to the laboratory and have spanned a relatively short time scale. The most ambitious previous experiment (Pavlik & Anderson, 2008) involved three study sessions in one week and a test the following week. This compressed time scale limits the opportunity to manipulate spacing in a manner that would influence long-term retention (Cepeda et al., 2008). Further, brief laboratory studies do not deal with the complex issues that arise in a classroom, such as the staggered introduction of material and the certainty of exposure to the material outside of the experimental context.

Whereas previous studies offer in-principle evidence that human learning can be improved by the timing of review, our results demonstrate in practice that integrating personalized-review software into the classroom yields appreciable improvements in long-term educational outcomes. Our experiment goes beyond past efforts in its scope: it spans the time frame of a semester, covers

the content of an entire course, and introduces material in a staggered fashion and in coordination with other course activities. We find it remarkable that the review manipulation had as large an effect as it did, considering that the duration of roughly 30 minutes a week was only about 10% of the time students were engaged with the course. The additional, uncontrolled exposure to material from classroom instruction, homework, and the textbook might well have washed out the effect of the experimental manipulation.

4.2.2.1 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 4.1, Table 4.1). 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 3.15).

Apart from the academic literature, two traditional adaptive-scheduling techniques have attracted a degree of popular interest: the Leitner (1972) system and SuperMemo (Woziak & Gorzelanczyk, 1994). Both aim to review material at the point of *desirable difficulty* (Bjork, 1994)— 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 (Landauer & Eldridge, 1967; Lindsey, Mozer, Cepeda, & Pashler, 2009). 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, for the few scheduling algorithms we have been able to investigate, we doubt the claims that they optimize long-term retention.

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.

4.2.2.2 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* (van Lehn, Jordan, & Litman, 2007) and that observed student behavior permits inferences about the state of these KCs. The approach is flexible, allowing for any problem posed to a student to depend on arbitrary combinations of KCs. The approach is also general, having application beyond declarative learning to domains focused on conceptual, procedural, and skill learning.

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

4.2.2.3 Personalized review scheduling

DASH obtains a posterior predictive distribution for each student-KC pair over the probability that recall will succeed if the KC were to be presented to the student at the current moment in time. These predictions are necessary in order to schedule review optimally but unfortunately are not sufficient. Although these predictions are 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). Consequently, a heuristic policy is required for selecting review material. The heuristic policy we used for personalized review within COLT is motivated by two distinct arguments, summarized here.

Using simulation studies, Khajah et al. (2014) examined policies that approximate the optimal policy found by exhaustive combinatorial search. To serve as a proxy for the student, in their simulations, they used a range of parameterizations of MCM and ACT-R, two of the best established models of memory for temporally distributed study. Their simulations were based on a set of assumptions approximately true for COLT, including a 10-week experiment in which new material is introduced each week, and a limited, fixed time allotted for review each week. They incorporated additional simplifying assumptions, including: all material had the same difficulty and the learning of one KC did not interact with the learning of another. Under these assumptions, exact optimization could be performed for a student who behaved according to a particular parameterization of either MCM or ACT-R. Comparing long-term retention under alternative policies, the optimal policy obtained performance only slightly better than a simple heuristic policy that prioritizes for review the item whose expected recall probability is closest to a threshold θ, with the threshold $\theta = 0.33$ being best over a range of conditions. Note that with $\theta > 0$, DASH's student-ability parameter, α_s, influences the *relative* prioritization of items.

An independent argument can be made for a threshold-based scheduler from Bjork's (1994) notion of *desirable difficulty*, which suggests that material should be restudied as it is on the verge

of being forgotten. The more difficult a correct answer is for a student to produce in a retrieval practice trial, the more the student's memory will be enhanced by the trial. However, difficult trials in which the student fails to recall the answer are less beneficial than difficult trials in which recall succeeds (the *retrieval effort hypothesis* (Pyc & Rawon, 2009)), hence memory is best served by reviewing material just before it is forgotten. The qualitative prescription that a KC should be studied when it is "about to be forgotten" maps naturally into the quantitative threshold-based policy, assuming one has a memory model like DASH that can accurately predict for individual students and KCs.

4.2.3 Additional information

This section provides provide additional details and analyses related to the Main Experiment as presented in section 4.2.

4.2.3.1 Software

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

When students click the *Begin Studying* button, they are taken to another web page which presents English-Spanish flashcards through *retrieval-practice trials*. At the start of a retrieval-practice trial, students are prompted with a **cue**—an English word or phrase. Students then attempt to type the corresponding **target**—the Spanish translation—after which they receive feedback (Figure 6.1). The feedback consists of the correct translation and a change to the screen's background color: the tint shifts to green when a response is correct and to red when it is incorrect. This form of study exploits the *testing effect*: when students are tested on material and can successfully recall it, they will remember it better than if they had not been tested (H. Roediger &

Karpicke, 2006b). Translation was practiced only from English to Spanish because of approximate associative symmetry and the benefit to students from their translating in the direction of the less familiar orthography (Kahana & Caplan, 2002; Schneider, Healy, & Bourne, 2002).

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

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

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

4.2.3.2 Implementation

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

Figure 4.4: Interface students used in the experiment. The left figure shows the start of a retrieval-practice trial. The right figure shows consequence of an incorrect response.

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

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

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

Table 4.2: Calendar of events throughout the Main Experiment.

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

4.2.3.3 Semester calendar

The experiment proceeded according to the calendar in Table 4.2. The table shows the timeline of presentation of the chapters of material and the cumulative end-of-semester exams, along with the amount of material associated with each chapter. The amount of material is characterized in terms of both the number of unique words or phrases (column 4) and the number of KCs (column 5).

The course was organized such that in-class introduction of a chapter's material was coordinated with practice of the same material using COLT. Typically, students used COLT during class time for three 20-30 minute sessions each week, with exceptions due to holiday schedules or special classroom activities. New material was typically introduced in COLT on a Friday, followed by

additional practice the following Tuesday. In Experiment 1, this was followed by an end-of-chapter quiz on either Wednesday or Thursday. Experiments 2-4 had no weekly quizzing. In addition to the classroom sessions, students were allowed to use COLT at their discretion from home. Each session at home followed the same sequence as the in-class sessions. Figure 4.5 presents pseudocode outlining the selection of items for presentation within each session.

During experiment 1, the quizzes were administered on chapters 1-9 and counted toward the students' course grade. The instructor chose the variants of a KC that would be tested. For all but the chapter 8 quiz, the instructor selected material only from the current chapter. The chapter 8 quiz had material from chapters 7 and 8. Quizzes typically tested most of the KCs in a chapter (column 6 of Table 4.2).

Two cumulative final exams were administered following introduction of all chapters. Cumulative exam 1 occurred around the end of the semester; cumulative exam 2 occurred four weeks later. For experiments 1 and 3, the second cumulative exam followed an intersemester break. Students were not allowed to use COLT between semesters.

4.2.3.4 Materials

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

The 409 words and phrases were reassembled into 221 clusters. Following terminology of the intelligent tutoring community, we refer to a cluster as a **knowledge component** or **KC**. However,

% Study to Proficiency Phase

Let $c \leftarrow$ the current chapter

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

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

Let $y \leftarrow$ a random permutation of x

For each KC i **in** y

Execute a retrieval practice trial on i

If the student answered correctly

Remove i from x

% Review Phase

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

Let $z \leftarrow$ a random permutation of m

Let $k \leftarrow 0$

Until the student quits

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

If $z_k =$ MASSED

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

Else If $z_k =$ GENERIC

If $c > 0$

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

Else

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

Else $z_k =$ PERSONALIZED

Let $i \leftarrow$ the KC in w and in any of chapters $0 \ldots c$ whose current posterior mean recall probability for the student is closest to the desirable difficulty level d

Execute a retrieval practice trial on i

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

Figure 4.5: Pseudocode showing the sequence of steps that each student underwent in a study session in the Main Experiment. Students begin in a study-to-proficiency phase on material from the chapter currently being covered in class. If students complete the study-to-proficiency phase, they proceed to a review phase. During the review phase, trials alternate between schedulers so that each scheduler receives an equal number of review trials. The graded end-of-chapter quizzes did not follow this pseudocode and instead presented the same sequence of instructor-chosen retrieval practice trials to all students, ensuring that all students saw the same questions and had them in the same order.

earlier in this chapter we used the term **item** as a synonym to avoid introducing unnecessary jargon. The course organization was such that all variants of a KC were introduced in a single chapter. During practice trials, COLT randomly drew one variant of a KC.

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

4.2.3.5 Procedure

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

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

Quizzes were administered through COLT using retrieval-practice trials. From a student's

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

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

4.2.3.6 Participants

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

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

From the students' perspective, COLT was simply a replacement for the software they had been using and a substitute for pencil-and-paper quizzes. Students were not aware of the details of

our experimental manipulation, beyond the notion that the software would spend some portion of study time reviewing older vocabulary items.

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

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

4.2.3.7 Additional analyses

The amount of use of COLT varied by chapter due to competing classroom activities, the amount of material introduced in each chapter, the number of class days devoted to each chapter, and the amount of at-home use of COLT. Figure 4.6 presents the median number of retrieval practice trials undergone by students, broken down by chapter and response type (correct, incorrect, and "I don't know") and by in-class versus at-home use of COLT.

Figure 4.7 graphs the proportion correct recall on the two final exams by class section and review scheduler. The class sections are arranged in order from best to worst performing. An Analysis of Variance (ANOVA) was conducted on each exam with the dependent variable being proportion recalled on the exam and with three factors: class period, scheduler (massed, generic spaced, personalized spaced), and chapter of course (1-10). The main effect of the scheduler is highly reliable in both exams (exam 1: $F(2, 328) = 52.3$, $p < .001$; exam 2: $F(2, 340) = 55.1$, $p < .001$); the personalized-spaced scheduler outperforms the two control schedulers. The main effect of class period is significant in both exams (exam 1: $F(5, 164) = 6.77$, $p < .001$; exam

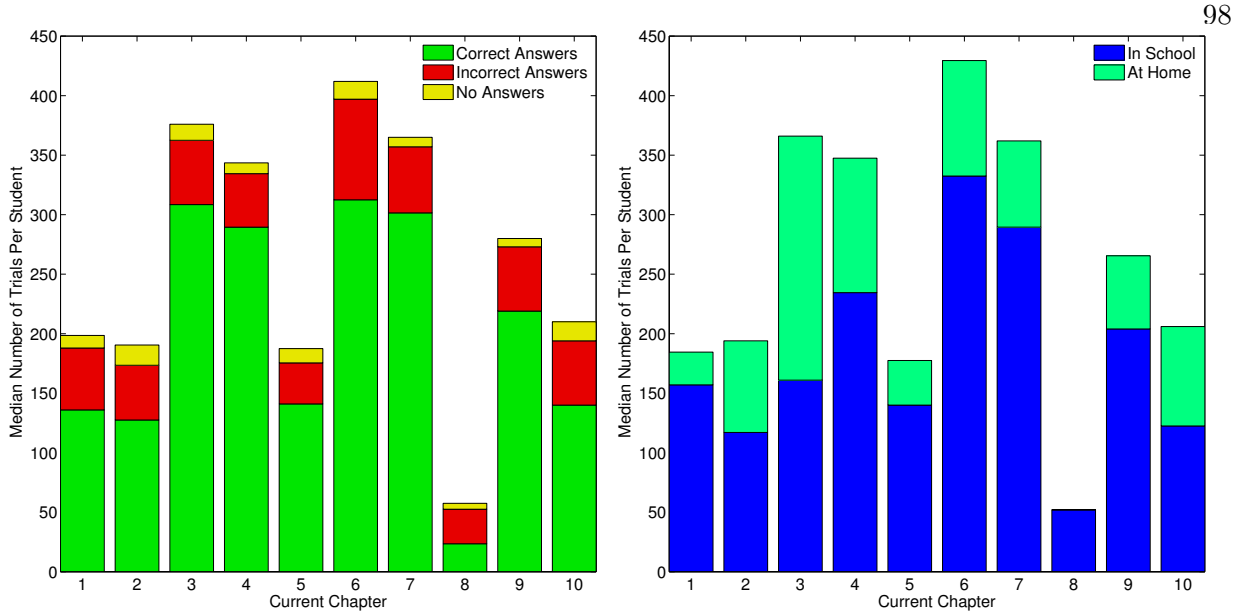


Figure 4.6: Median number of study trials undergone while each chapter was being covered in class. In the left panel, the number is broken down by whether the student responded correctly, responded incorrectly, or clicked "I don't know." In the right panel, the number is broken down by whether the trial happened on a weekday during school hours or not. Chapter 8 has few trials because it was covered in class only the day before a holiday break and the day after it.

2: $F(5, 170) = 9.72$, $p < .001$): some sections perform better than others. A scheduler $\times$ chapter interaction is observed (exam 1: $F(18, 2952) = 8.90$, $p < .001$; $F(9, 1530) = 29.67$, $p < .001$), as one would expect from Fig. 4.7: the scheduler has a larger influence on retention for the early chapters in the semester. The scheduler $\times$ period interaction is not reliable (exam 1: $F(10, 328) = 1.44$, $p = .16$; exam 2: $F(10, 340) = 1.36$, $p = .20$), nor is the three-way scheduler $\times$ period $\times$ chapter interaction (exam 1: $F(90, 2952) < 1$; exam 2: $F(90, 3060) < 1$).

Figure 4.9 splits performance separately on the end-of-semester exam and the exam administered 28 days later. As the ANOVAs in the previous paragraph suggest, the qualitative pattern of results is similar across the two exams. Note that this figure shows students who took either exam. Only a few students missed both exams.

Figure 4.8 shows the mean quiz scores on each chapter for the three conditions. Except for the chapter 8 quiz, all quizzes were on only the current chapter. Ignore chapter 8 for the moment, and also ignore chapter 1 because the three conditions were indistinguishable the first week of the

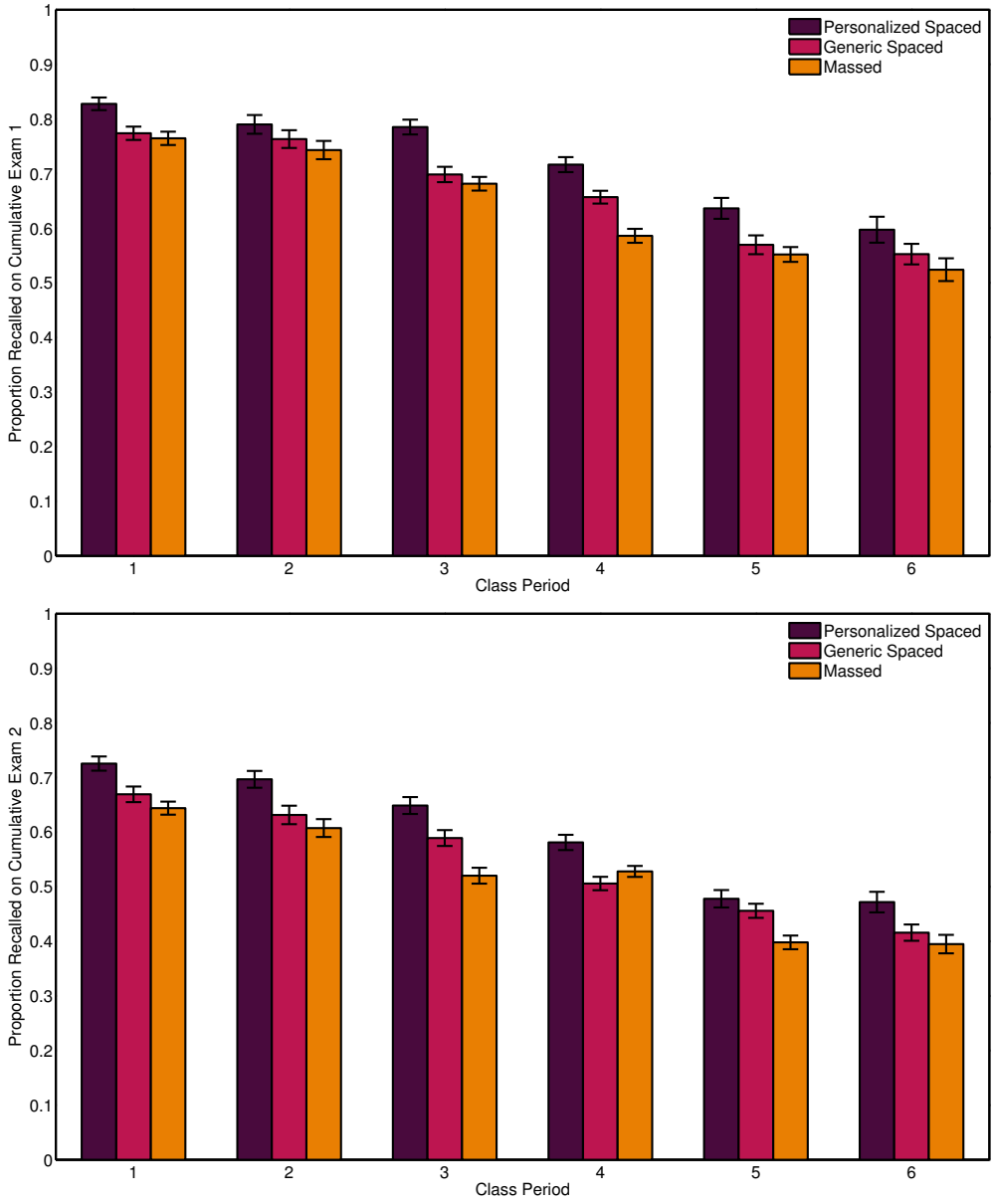


Figure 4.7: Scores on cumulative exams 1 and 2 for each class period. Each group of bars is a class period. The class periods are presented in rank order by their mean Exam 1 score.

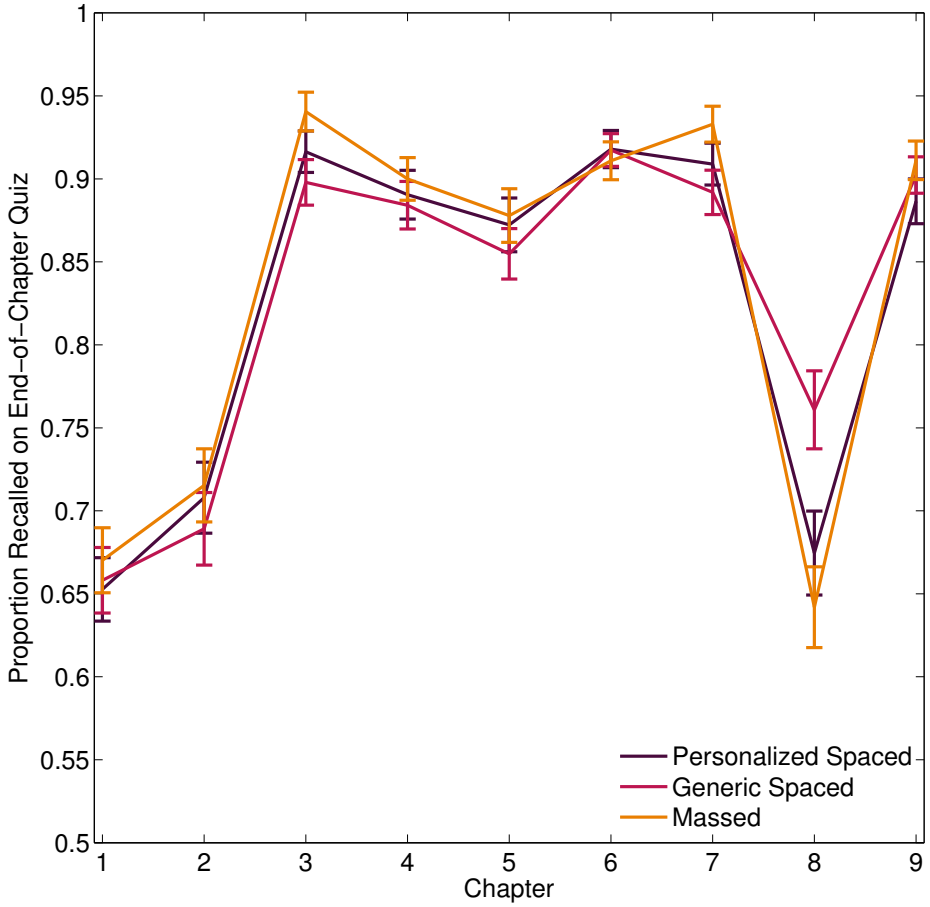


Figure 4.8: End-of-chapter quiz scores by chapter. Note that the chapter 8 quiz included material from chapter 7, but all the other quizzes had material only from the current chapter. There was no chapter 10 quiz.

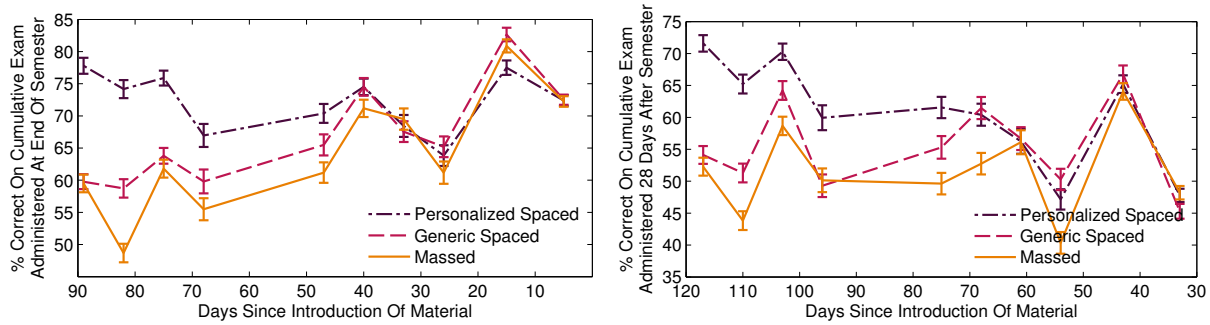


Figure 4.9: Mean score on each of the two exams as a function of the number of days that had passed since the material was introduced. The two exams show similar results by scheduler and chapter.

semester. An ANOVA was conducted with the dependent variable being proportion correct on a

quiz and with the chapter number (2-7, 9) as a factor. Only the 156 students who took all seven of these quizzes were included. The main effect of review scheduler is significant ($F(2, 310) = 11.8$, $p < .001$): the massed scheduler does best on the quizzes—89.4% versus 87.2% and 88.1% for the generic and personalized spaced schedulers—because it provided the largest number of study trials on the quizzed chapter. The main effect of the chapter is significant ($F(6, 930) = 49.0$, $p < .001$), and the scheduler $\times$ chapter interaction is not reliable ($F(12, 1860) = 1.56$, $p = .096$). The simultaneous advantage of the massed condition on immediate tests (the chapter quizzes) and the spaced conditions on delayed tests (the final exams) is consistent with the experimental literature on the distributed-practice effect.

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

4.3 Followup Experiment 1

Section 4.2 described how we used a statistical model of student learning and forgetting to improve long-term retention for students in a Denver-area middle school (see also Lindsey et al., 2014). The within-subject design had a personalized-spaced condition (i.e., the model-based scheduler), a generic-spaced condition (i.e., the review 1-chapter-ago scheduler), and a massed condition (i.e., the scheduler that focused on the current chapter). Although the generic spaced and massed control conditions are well motivated by psychological theory and current educational practices, respectively, they did not shed much light on how effective the model-based scheduler is compared to less intelligent schedulers. In the final weeks of the semester, items from chapters covered early in the semester could be reviewed on COLT only if they were in the personalized-spaced condition. Although students periodically encountered old material in the classroom, it could be expected that any experimental manipulation which allows for the review of the oldest material (i.e., the personalized spaced scheduler) will outperform any experimental manipulation which does not allow

	Textbook Section	Day of Study	# Words & Phrases	# KCs
Chapter 1 Introduced	6-2	1	44	27
Chapter 2 Introduced	6-3	10	46	22
Chapter 3 Introduced	7-1	17	42	28
Chapter 4 Introduced	7-2	24	26	21
Chapter 5 Introduced	7-3	31	28	19
Chapter 6 Introduced	8-1	52	34	34
Chapter 7 Introduced	8-2	59	34	34
Chapter 8 Introduced	9-1	80	33	33
Cumulative Exam 1	-	92-93	-	-
Cumulative Exam 2	-	120-121	-	-

Table 4.3: Calendar of events throughout Followup Experiment 1.

it. Whether we had scheduled review through our statistical model or by some less sophisticated means, it is possible that we would have observed qualitatively similar results. Therefore, we ran a followup experiment with the same teacher and students in the semester following the original experiment. Of the 179 students from the original experiment, 178 participated in this followup experiment.

Followup Experiment 1 was a replication of the original experiment, except that we replaced the poorly performing massed scheduler with a *random* scheduler. The random scheduler selected material for review at random from the set of all vocabulary items introduced so far in the class. Thus, even late in the semester, this scheduler would often have students review material from early in the semester. We used the data collected in the original experiment to help constrain our model's estimates of the students' abilities and the rate at which they forget material. Thus, even on the first day of the experiment, the model had a strong estimate of each student's aptitude.

The instructor provided 287 Spanish-English words and phrases which covered 8 blocks of material. The material came from the same textbook that was used in the original experiment. Chapters were introduced one at a time according to the schedule in Table 4.3. The cumulative final exams were separated by 28 days, during which time students were still exposed to the material in class but did not use COLT and did not learn any new material.

4.3.1 Results and Discussion

The average test scores on exams 1 and 2 are presented in Figure 4.10. The personalized spaced scheduler gives a reliable improvement of 3.4% over the random scheduler ($t(167) = 2.29$, $p = 0.02$, Cohen's $d = 0.18$) and a reliable improvement of 4.8% over the generic spaced scheduler ($t(167) = 3.039$, $p < 0.01$, $d = 0.23$) on the first exam. On the second exam, the personalized spaced scheduler gives no reliable improvement over the random scheduler ($t(166) = 1.6359$, $p = 0.10376$, $d = 0.12659$) and a reliable improvement of 4.6% over the generic spaced scheduler ($t(166) = 2.2717$, $p = 0.024389$, $d = 0.17579$). As in the original experiment, the primary impact of the schedulers was for material introduced earlier in the semester (Figure 4.11), which is sensible because that material had the most opportunity for being manipulated via review

We ran separate ANOVAs on Exams 1 and 2 with the chapter and review scheduler as factors. Exam 1 showed a main effect of the review scheduling condition ($F(2, 334) = 10.95$, $p < .001$) and an interaction with the chapter ($F(14, 2338) = 16.40$, $p < .001$). Exam 2 also showed a main effect of the review scheduling condition ($F(2, 332) = 4.23$, $p = .015$) and also an interaction with the chapter ($F(14, 2324) = 12.0$, $p < .001$). We also ran an ANOVA on both exams in which we included exam as a factor and included data only from students who took both exams. The joint ANOVA shows no interaction between the review scheduling condition and the exam ($F(2, 316) < 1$). Thus, the pattern of results was not reliably different for the two exams. In the joint ANOVA, there is a main effect of the scheduler ($F(2, 316) = 13.88$, $p < .001$), an interaction between the scheduler and the chapter ($F(14, 2212) = 23.97$, $p < .001$), and a weakly significant three-way interaction involving the chapter, scheduler, and exam ($F(14, 2212) = 1.86$, $p = .026$).

We encountered one significant problem while administering the semester-long experiment. As with the Main Experiment, this experiment was run through a website. When a student logged in to the website, he or she first underwent a study-to-criterion stage. Only after the student completed that stage did the website begin selecting material for review according to the schedulers. Unbeknownst to us during the experiment, a significant portion of the students learned

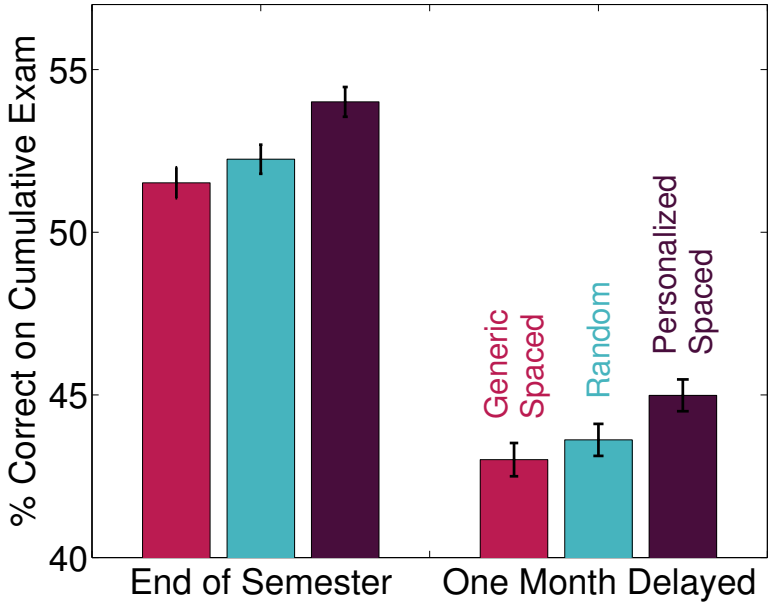


Figure 4.10: Mean scores on the two cumulative end-of-semester exams in Followup Experiment 1, taken 28 days apart. All error bars indicate ±1 within-student standard error (Masson & Loftus, 2003). The relative difference between the personalized and generic schedulers (8.1%) is approximately the same as the relative difference between them in the Main Experiment (8.3%).

that they could avoid the review stage entirely by logging out and back in to the website. This would permanently keep them in the study-to-criterion stage. Students are rationally interested in maximizing their grades on the course's weekly non-cumulative quizzes. Undergoing review of previous chapters takes time away from cramming for the weekly quiz, thus is something they wanted to avoid.

The students' exploitation of the study-to-criterion stage makes the observed benefit of the personalized spaced scheduler all the more surprising. Figure 4.12 shows a breakdown of the final test scores by the amount of time each student spent in the review stage. There is a positive correlation between the time spent in review and the within-student advantage of the personalized-spaced condition over the random-spaced condition. This suggests that the effect of the personalized review scheduler would have been larger had the students not skipped the review stage so often.

Our results again demonstrate that integrating personalized-review software into the classroom yields appreciable improvements in long-term educational outcomes. This study replicates

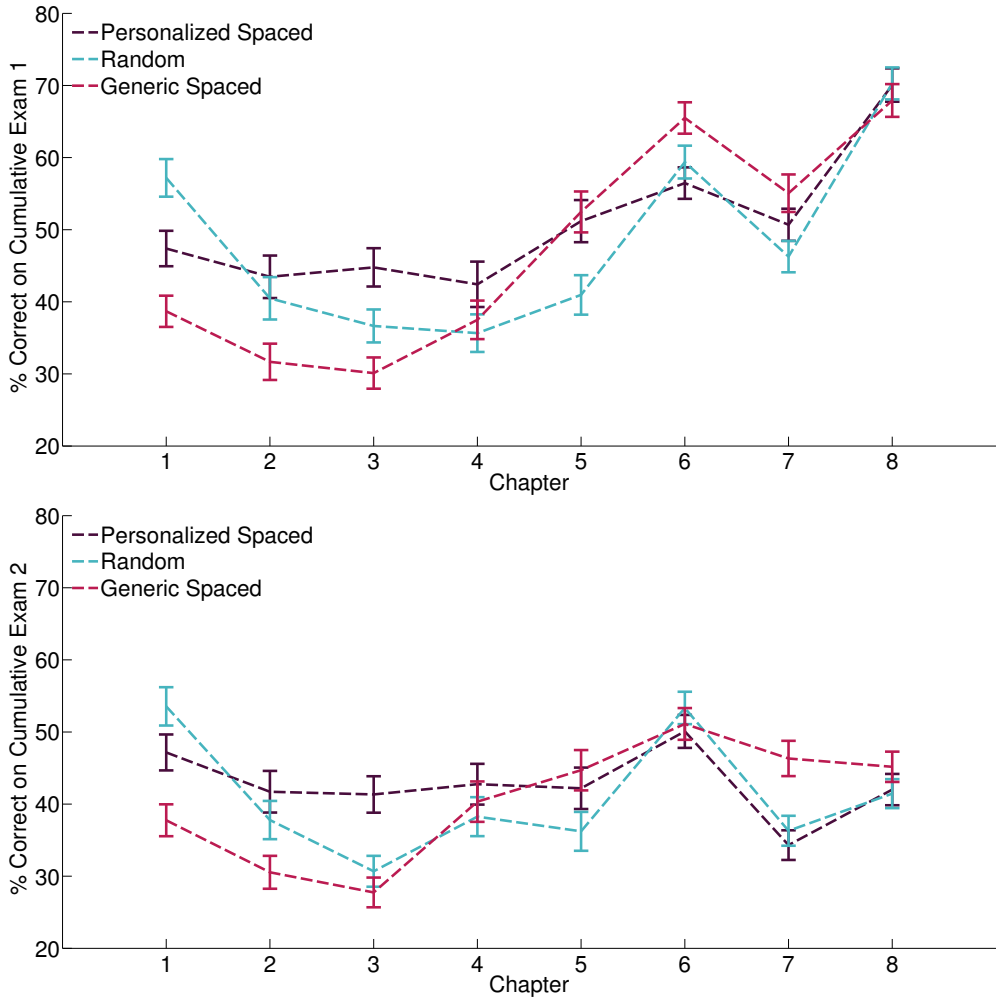


Figure 4.11: Mean scores on the two cumulative exams in Followup Experiment 1 as a function of the chapter number.

the original study's finding of an improvement of delivering personalized, spaced review over simply delivering spaced review based on qualitative advice from the psychology literature. This study's results concerning the random scheduler provide evidence that the improvement of the personalized review scheduler is attributable to its making intelligent decisions about what material should be reviewed.

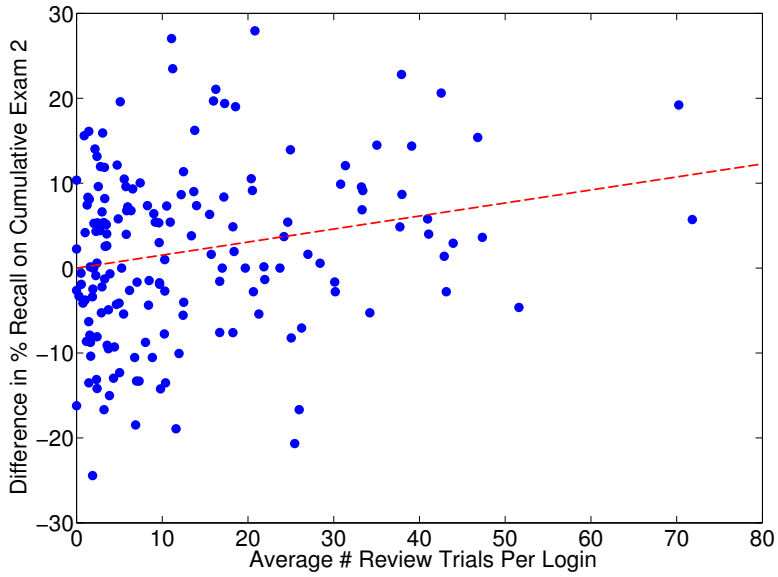


Figure 4.12: Each point represents a student that took the first cumulative exam in Followup Experiment 1. The horizontal axis shows the average number of review-stage trials a student underwent per login session. The vertical axis shows the within-student difference in percent recall between the personalized spaced and random spaced conditions on the second cumulative exam. The regressor line has an assumed intercept of 0, and a fitted slope of 0.15334 ($t(167) = 3.94, p = .0001$). Some students found a way to bypass the review stage in the experiment. This is partly evident by the observation that most students have a small average number of review trials per login. Nevertheless, this figure demonstrates that the within-student benefit of the personalized scheduler over the random scheduler grows with the number of review trials undergone.

4.4 Followup Experiment 2

We again incorporated systematic, temporally distributed review into third-semester Spanish foreign language instruction using COLT, our web-based flashcard tutoring system. Throughout a semester, 250 students used COLT to drill on 13 chapters of Spanish words and phrases. The students used COLT to complement the practice of newly introduced material and the review of previously studied material, both of which they also received in class and through software provided by their textbook's manufacturer. For each chapter of course material, students engaged in two 15-minute sessions with COLT during class time. The students were required to answer at least 100 retrieval practice trials correctly per week. If a student did not meet that requirement, he or she had to finish by working from home.

Selection of items was handled by four different schedulers. This experiment did not have one study-to-criterion stage per login, unlike the previous experiments. Whenever a student logged in to the website, COLT immediately began alternating among the four review scheduling conditions. However, if a new item had been introduced on the website in the condition and had never been correctly recalled, the system overrode the choice of the review scheduler and presented the new item for study. Thus, the experiment avoided the problem we encountered in Followup Experiment 1, wherein students found a way to avoid reviewing old material. In previous iterations of the experiment, the different conditions could undergo slightly different numbers of trials because of the study-to-criterion stage. This iteration of the experiment exactly matched the total number of trials within-student.

A *massed* scheduler continued to select items from the current chapter (Figure 4.13, upper), and a *personalized spaced* scheduler used our DASH-ACT-R model (see section 3.4.4) to select items for review (Figure 4.14, upper). Both schedulers followed the same procedure as in the Main Experiment. The massed scheduler represents current educational practice, and the personalized spaced scheduler represents a model-based approach to incorporating spaced review into the classroom. Note that students still were exposed to material in the massed condition outside of COLT: they encountered it in lectures, through their textbook, and through their textbook's online practice software.

A *SuperMemo* scheduler selected material according to the heuristics used by the commercially available software named SuperMemo (Figure 4.14, lower). The SuperMemo scheduler is based on a complex set of heuristics and assumptions, but the system is at its core a Leitner box system (Leitner, 1972). It progressively increases the time between successive presentations of an item (Woziak & Gorzelanczyk, 1994). When a student fails to recall an item, the item's spacing is reset and the progression starts anew. A black-box implementation of SuperMemo's scheduler was graciously provided to us by the SuperMemo company, and we interfaced it with our experiment website. In addition to providing students with a progressively expanding spacing schedule, it reportedly provides a personalized experience to each student by adapting its spacing schedules

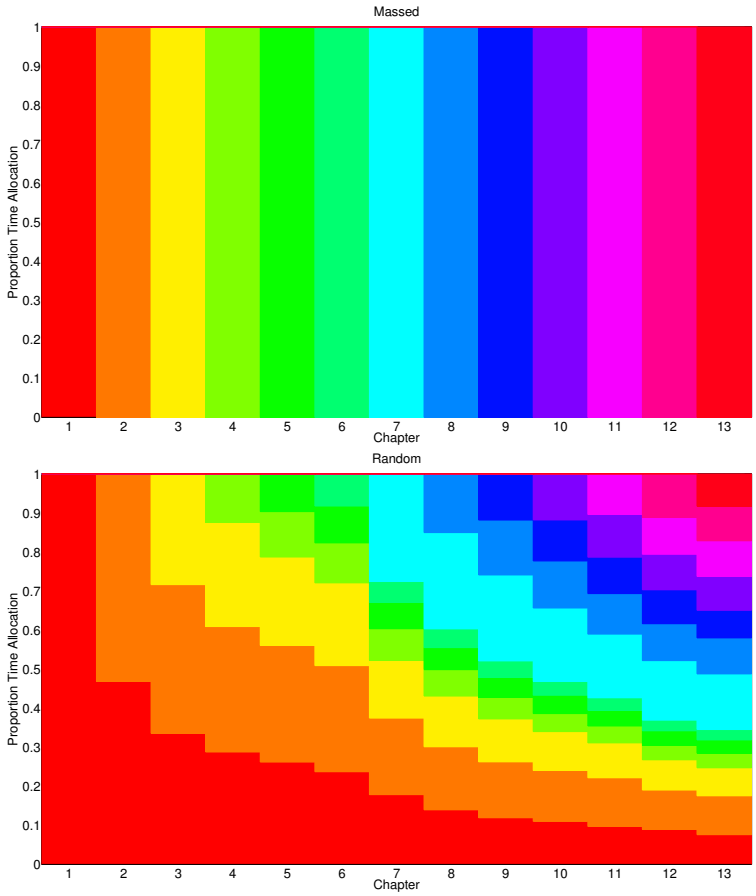


Figure 4.13: Time allocation of the massed and random review schedulers in Followup Experiment 2. As in the original experiment, course material was introduced one chapter at a time. Each vertical slice indicates the proportion of time spent studying each of the chapters introduced so far throughout the period of time the current chapter was being covered. Each chapter is indicated by a unique color. The random condition selected an old KC to review uniformly at random from among the KCs that had been introduced so far.

based on each student's across-item performance.

A *random* scheduler selected an item uniformly at random from among the set of items that had been introduced so far in the class (Figure 4.13, lower). The random scheduler provided students with systematic review of old material, but which item it selected for study was not influenced by the students' responses.

The scheduler was varied within-student by randomly assigning one quarter of a chapter's items to each scheduler, counterbalanced across students. As in the preceding experiments, the schedulers alternated in selecting items for retrieval practice; each selected from among the items

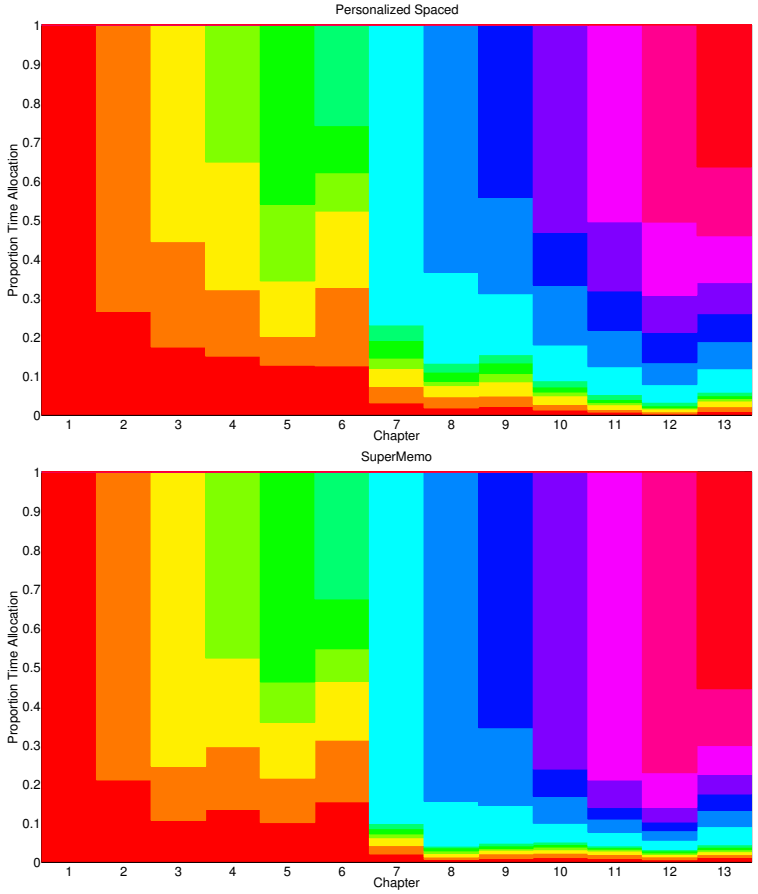


Figure 4.14: Time allocation of the personalized spaced and SuperMemo schedulers in Followup Experiment 2. Both schedulers made granular decisions about what each student should study.

assigned to it, ensuring that all items had equal opportunity and that all schedulers administered an equal number of review trials. For more information about the experimental procedures, see section 4.4.2.

4.4.1 Results

Across the semester, we recorded 633,796 retrieval practice trials: 430,416 correct responses, 137,894 incorrect responses, and 65,486 non-responses. The four schedulers each administered approximately 120,000 review trials, with the difference between schedulers being less than the number of students in the experiment. Two proctored cumulative exams were administered to assess retention, one immediately following an intrasemester break and one 45 days later at the

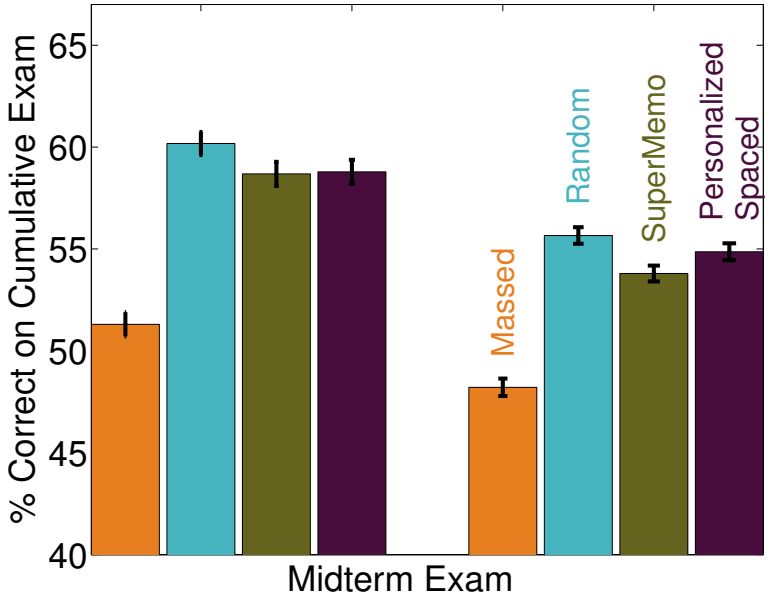


Figure 4.15: Mean scores on the cumulative mid-semester exam and the end-of-semester exam in Followup Experiment 2, taken 45 days apart. All error bars indicate ±1 within-student standard error (Masson & Loftus, 2003).

end of the semester. No corrective feedback was provided to students during either exam. The mid-semester exam was taken by 239 students and tested approximately half of the material from the chapters that had been covered by that point in the semester. The end-of-semester exam was taken by 230 students. Due to a programmer error, the end-of-semester exam was systematically unbalanced: some students underwent no trials in certain conditions for certain chapters. This error makes the interpretation of Exam 2 results difficult. We are still trying to understand exactly how this unbalancing biased the results of the experiment.

The average scores in the four conditions on each exam are shown in Figure 4.15. Paired t-tests show no reliable differences between the personalized scheduler and either the SuperMemo or random schedulers on either exam. The personalized scheduler provides a highly reliable 14.6% relative improvement over the massed scheduler on the midterm exam ($t(238) = 7.66$, $p <$ 1e-12, $d = 0.50$) and a highly reliable 13.8% relative improvement over the massed scheduler on the end-of-semester exam ($t(229) = 9.77$, $p <$ 1e-18, $d = 0.64$).

The differences between the review schedulers are very pronounced for the material introduced

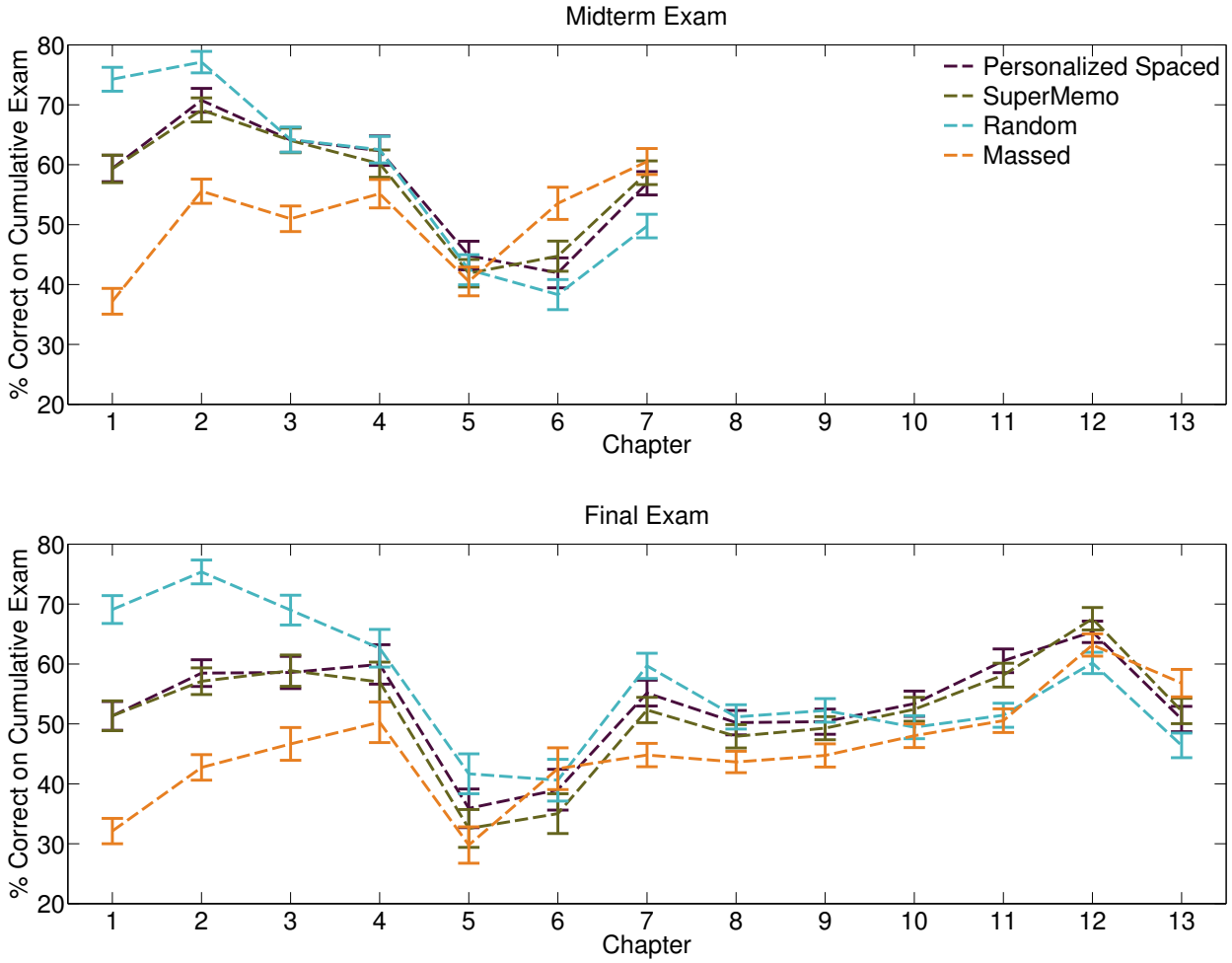


Figure 4.16: Mean scores on the cumulative mid-semester exam and the end-of-semester exam in Followup Experiment 2 as a function of the chapter number. Chapters were typically introduced at one-week intervals, and the final exam occurred 120 days after the introduction of Chapter 1.

early in the semester. Figure 4.16 shows test performance as a function of the chapter number. Note that the experimental manipulations represent a small portion of the time students were engaged with the material. Students encountered material—even from the massed condition—throughout the semester via lectures, homework, and class projects. In light of that, the differences in long-term retention between conditions is remarkable. The students' retention of the chapter 1 material in the massed condition—which represents current educational practice—was less than half that of the random condition.

Because of the complexity of the experiment, its problems, and how recently we obtained

the results, we have yet to arrive on a coherent interpretation of the outcome. However, given the improvement seen in Followup Experiment 1 of the personalized scheduler over the random scheduler, we suspect that there may have been some other problem in this experiment. The difference between this experiment's results and the previous experiments' results may also be attributable to a combination of the change in the study-to-criterion stage, to the change in the model we used, to rampant cheating, or to the addition of an extra teacher into the experiment (who did not as closely monitor students), among other possibilities.

4.4.2 Additional information

This section provides provide additional details and analyses related to Followup Experiment 2 as presented in section 4.4.

4.4.2.1 Semester Calendar

Followup Experiment 2 proceeded according to the calendar in Table 4.4. The table shows the timeline of presentation of the chapters of material and the cumulative mid-semester exam and cumulative final exam. As in the preceding experiments, the course was organized such that in-class introduction of a chapter's material was coordinated with practice of the same material using COLT. Typically, students used COLT during class time for two 15-minute sessions per week. The instructors required that each student answer 100 trials correctly on COLT per week (see Figure 4.17). Students who did not complete their weekly quota in class were required to finish at home, and students who wished to go beyond their quota were allowed to do so at their own discretion.

The cumulative midterm exam was administered following the introduction of the first 7 chapters of material, immediately following a week-long mid-semester vacation. The cumulative final exam was administered at the end of the semester.

	Textbook Section	Day of Study	# Words & Phrases	# KCs
Chapter 1 Introduced	4-1	1	24	20
Chapter 2 Introduced	4-2	11	24	23
Chapter 3 Introduced	4-3	18	32	18
Chapter 4 Introduced	4-4	25	30	9
Chapter 5 Introduced	4-5	32	26	8
Chapter 6 Introduced	4-6	39	24	7
Chapter 7 Introduced	4-7	52	33	33
Midterm Exam	-	74-75	-	-
Chapter 8 Introduced	5-1	76	24	22
Chapter 9 Introduced	5-2	81	19	19
Chapter 10 Introduced	5-3	88	23	21
Chapter 11 Introduced	5-4	95	23	22
Chapter 12 Introduced	5-5	102	23	23
Chapter 13 Introduced	5-6	109	18	18
Cumulative Exam	-	120	-	-

Table 4.4: Calendar of events throughout Followup Experiment 2.

4.4.2.2 Participants

Participants were eighth graders from the suburban Denver middle school that participated in the Main Experiment. A total of 250 students were divided among 9 class periods of a third-semester Spanish course taught in parallel by two instructors. Four class periods met Mondays, Wednesdays, and Fridays, and three class periods met Tuesdays, Thursdays, and Fridays. The Monday, Wednesday, Tuesday, and Thursday classes met for 94 minutes, and the Friday classes met for 47 minutes. Instructor 1 had 110 students—62 male and 48 female—across 4 class periods. Instructor 2 had 140 students—75 male and 65 female—across 5 class periods. Instructor 1 is the same instructor that participated in the Main Experiment and Followup Experiment 1, and Instructor 2 is a veteran of 15 years of teaching Spanish as a foreign language.

Students were not made aware of the details of our experimental manipulation. The instructors were aware of the manipulation, but did not know which study items were assigned to what conditions for any of the students.

The instructors reported to us that at some point in the semester, students learned how to cheat on our web-based tutoring system by manipulating the Javascript via web-developer tools.

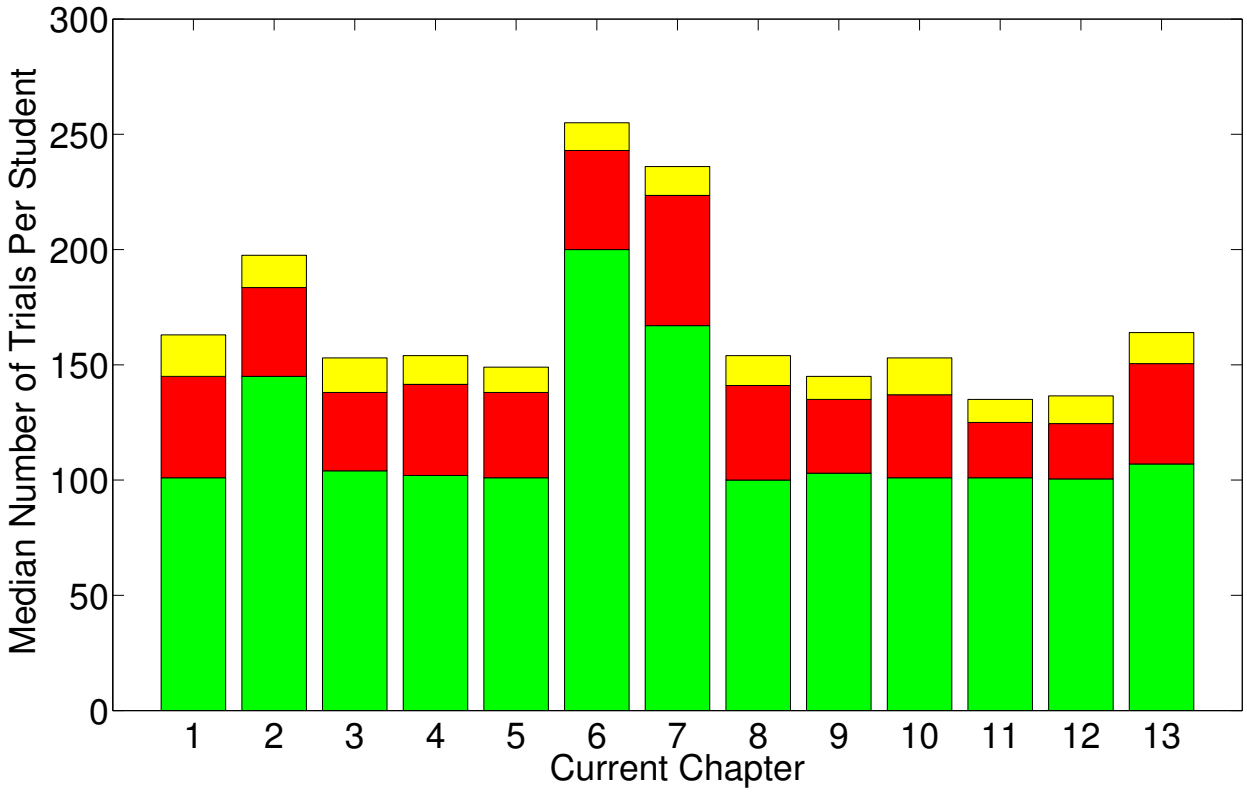


Figure 4.17: Median number of study trials undergone while each chapter was being covered in class in Followup Experiment 2. Each number is broken down by whether the student responded correctly (green), responded incorrectly (red), or clicked "I don't know" (yellow). Students were required to answer a minimum of 100 trials correctly each week. Since a new chapter was typically introduced each week and since students did not typically study more than was required, most green bars are at approximately 100.

It is not clear how widespread this cheating was or how long it lasted, but the instructors do not believe that many students cheated in this manner. It was also reported to us that students sometimes would often cheat by looking up the correct responses to a retrieval practice trial through their textbook, and they would also cheat by working collaboratively or by using online translation programs.

4.4.2.3 Materials

The instructors provided 323 Spanish-English words and phrases, covering 13 sections of material. The material came from the course's electronic textbook, *Descubre*, which every student

had access to. As in earlier experiments, rather than treating minor variants of words and phrases as distinct and learned independently, we manually formed clusters of highly related words and phrases which were assumed to roughly form equivalence classes (e.g., all conjugations of a verb were assumed to form an equivalence class). The 323 words and phrases were thus grouped together into 243 clusters (KCs). All variants of each KC were introduced in the same section. During practice trials, the website drew one variant of a selected KC at random. For each chapter, KCs were assigned to one of the four scheduling conditions for each student, using the same criteria as in the Main Experiment. See section 4.2.3.4 for more information.

4.4.2.4 Procedure

In each COLT session in the Main Experiment, students began with a study-to-proficiency stage with material from only the current chapter, and then moved on to a review stage after all items had been correctly answered once. The study-to-proficiency stage proceeded without regard for the condition assignments of items: it did not match for the total number of trials undergone in each condition. The Main Experiment only matched for the total number of *review* stage trials. Only in the review stage did the schedulers have control over what items were selected for study.

This experiment followed a procedure which balanced the total number of trials undergone in each condition. There was no explicit study-to-proficiency stage. Instead, the website always alternated among the four conditions, each of which was responsible for a random one-quarter of the items from each chapter, assigned on a per-student basis. When a condition needed to select a KC to present in a retrieval-practice trial, it checked whether any introduced items assigned to the condition had never been answered correctly by the student on the website. If there were no such KCs, the website selected an item assigned to the condition in accord with the condition's scheduling algorithm (e.g., the personalized spaced would select an item based on predicted recall probability). Otherwise, the website would randomly select one of the items assigned to the condition that had never been correctly answered. This procedure guaranteed that students would always focus on new material when it was introduced on the website, regardless of the condition assignment. It

also guaranteed that each condition would receive the same number of trials, whereas the Main Experiment's more complicated procedure only matched for the number of *review* trials.

The cumulative midterm and final exams were administered through COLT. In each question on the exams, a student was prompted with a cue and typed in a response. Students did not receive corrective feedback during the exams. The exams were proctored by the instructors, and no students were caught cheating.

Chapter 5

Optimizing instruction for populations of students

5.1 Introduction

What makes teachers effective? A critical factor is their **instructional policy**, which specifies the manner and content of instruction. We use the term 'policy' in the standard sense—as a set of procedures governing action, in this case, rules that guide how a student should be taught. Electronic tutoring systems have been constructed that implement domain-specific instructional policies (e.g., J. R. Anderson et al., 1989; K. R. Koedinger & Corbett, 2006; Martin & van Lehn, 1995). A tutoring system decides at every point in a session whether to present some new material, provide a detailed example to illustrate a concept, pose new problems or questions that are similar to previously presented examples, or lead the student step-by-step to discover an answer. Prior efforts have focused on higher cognitive domains (e.g., algebra) in which policies result from an expert-systems approach involving careful handcrafted analysis and design followed by iterative evaluation and refinement. As a complement to these efforts, we are interested in addressing fundamental questions in the design of instructional policies that pertain to basic cognitive skills. For example, how long should the teacher wait after posing a question before providing an answer? How much time should the teacher spend on each subtopic within a topic? When the teacher asks a question, should the teacher offer additional support in the form of hints or partial answers to provide scaffolding for learning, and what hints should be provided? How difficult a question should the teacher select given the student's study and performance history? Should successive questions concern the same concept/topic, or should a switch be made to a different concept/topic?

Consider a concrete example: training individuals to discriminate between two perceptual or conceptual categories, such as determining whether mammogram x-ray images are negative or positive for an abnormality. In training from examples, should the instructor tend to alternate between categories—as in PNPNPNPN for positive and negative examples—or present a series of instances from the same category—PPPPNNNN (Goldstone & Steyvers, 2001)? Both of these strategies—**interleaving** and **blocking**, respectively—are adopted by human instructors (Khan, Zhu, & Mutlu, 2011). Reliable advantages between strategies has been observed (S. H. K. Kang & Pashler, 2011; Kornell & Bjork, 2008) and factors influencing the relative effectiveness of each have been explored (Carvalho & Goldstone, 2011). Why blocking vs. interleaving? The points of comparison are often selected based on the experimenter's intuition about what will be effective and—in order to obtain a publishable comparison—ineffective.

Empirical evaluation of blocking and interleaving policies involves training a set of human subjects with a fixed-length sequence of exemplars drawn from one policy or the other. During training, exemplars are presented one at a time, and typically subjects are asked to guess the category label associated with the exemplar, after which they are told the correct label. (Jacoby, Wahlheim, and Coane (2010) have shown that actively engaging subjects by requiring them to assign labels yields better learning than passive viewing of labeled exemplars.) Following training, mean classification accuracy is evaluated over a set of test exemplars. Such an experiment yields an intrinsically noisy evaluation of the two policies, limited by the number of subjects and inter-individual variability. Consequently, the goal of a typical psychological experiment is to find a statistically reliable difference between the training conditions, allowing the experimenter to conclude that one policy is superior.

Blocking and interleaving are but two points in a space of policies that could be parameterized by the probability, ρ, that the exemplar presented on trial $t+1$ is drawn from the same category as the exemplar on trial t. Blocking and interleaving correspond to ρ near 1 and 0, respectively. (There are many more interesting ways of constructing a policy space that includes blocking and interleaving—e.g., ρ might vary with t or with a student's running-average classification accuracy—

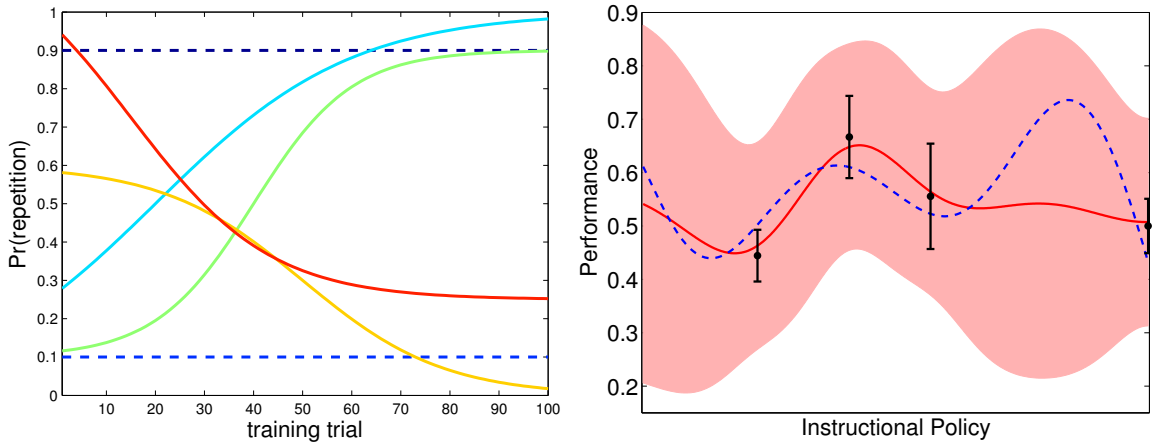


Figure 5.1: (left) Samples from a function space that characterizes policies for choosing the category of training examplars over a sequence of trials; (right) Illustration of a 1D instructional policy space: dashed line is performance as a function of policy; vertical black bars are experiment outcomes with uncertainty; red line and pink shading represent Gaussian Process posterior density

but we will use the simple fixed-ρ policy space for illustration.) Although one would ideally like to explore the policy space exhaustively, limits on the availability of experimental subjects and laboratory resources make it challenging to conduct studies evaluating more than a few candidate policies to the degree necessary to obtain statistically significant differences.

Figure 5.1a shows some examples of the former—time-dependent policies. The fixed interleaved and blocked policies are also depicted (the horizontal lines). These policies have the functional form

$$Pr(\text{category repetition on trial } t+1) = \beta_1 + \frac{\beta_2}{1+\mathrm{e}^{\beta_3 t+\beta_4}}, \tag{5.1}$$

where $\boldsymbol{\beta}$ defines a four-dimensional policy space, which includes time-invariant policies such as blocking and interleaving.

5.2 Optimization of instructional policies

Our goal is to discover the **optimum** in policy space—the policy that maximizes mean accuracy or another measure of performance over a population of students. (We focus on optimizing for a population but later discuss how our approach might be used to address individual differences.)

Our challenge is performing optimization on a budget: each subject tested imposes a time or financial cost. Evaluating a single policy with a degree of certainty requires testing many subjects to reduce sampling variance due to individual differences, factors outside of experimental control (e.g., alertness), and imprecise measurement obtained from brief evaluations and discrete (e.g., correct or incorrect) responses. Consequently, exhaustive search over the set of distinguishable policies is not feasible.

Past research on optimal teaching (Chi, van Lehn, Litman, & Jordan, 2011; Rafferty, Brunskill, Griffiths, & Shafto, 2011; Whitehill & Movellan, 2010) has investigated reinforcement learning and partially observable Markov decision process (POMDP) approaches. These approaches are intriguing but are not typically touted for their data efficiency. To avoid exceeding a subject budget, the flexibility of the POMDP framework demands additional bias, imposed via restrictions on the class of candidate policies and strong assumptions about the learner. The approach we will propose likewise requires specification of a constrained policy space, but does not make assumptions about the internal state of the learner or the temporal dynamics of learning. In contrast to POMDP approaches, the cognitive agnosticism of our approach allows it to be readily applied to arbitrary policy optimization problems. Direct optimization methods that accommodate noisy function evaluations have also been proposed, but experimentation with one such technique (E. J. Anderson & Ferris, 2001) convinced us that the method we propose here is orders of magnitude more efficient in its required subject budget.

Neither POMDP nor direct-optimization approaches model the policy space explicitly. In contrast, we propose an approach based on **function approximation**. From a function-approximation perspective, the goal is to determine the shape and optimum of the function that maps policies to performance—call this the **policy performance function** or **PPF**. What sort of experimental design should be used to approximate the PPF? Traditional experimental design—which aims to show a statistically reliable difference between two alternative policies—requires testing many subjects for each policy. However, if our goal is to determine the shape of the PPF, we may get better value from data collection by evaluating a large number of points in policy space each with few

subjects instead of a small number of points each with many subjects. This possibility suggests a new paradigm for experimental design in psychological science. What makes it particularly feasible is the existence of potential subject populations on the web, e.g., Amazon's Mechanical Turk. Although Mechanical Turk has been used primarily to farm out simple crowdsourcing tasks, a "task" can be defined to be engagement in an entire sequence of experimental trials. Our vision is a completely automated system that selects points in policy space to evaluate, runs an experiment—an evaluation of some policy with one or a small number of subjects—and repeats until a budget for data collection is exhausted.

5.2.1 Surrogate-based optimization using Gaussian process regression

In surrogate-based optimization (e.g., Forrester & Keane, 2009), experimental observations serve to constrain a **surrogate** model that approximates the function being optimized. This surrogate is used both to select additional experiments to run and —when the budget is exhausted— to estimate the optimum. Gaussian process regression (GPR) has long been used as the surrogate for solving low-dimensional stochastic optimization problems in engineering fields (Forrester & Keane, 2009; Sacks, Welch, Mitchell, & Wynn, 1989). Like other Bayesian models, GPR makes efficient use of limited data, which is particularly critical to us because our budget is expressed in terms of the number of subjects required. Further, GPR provides a principled approach to handling measurement uncertainty, which is a problem in any experimental context but is particularly striking in human experimentation due to the range of factors influencing performance. The primary constraint imposed by the Gaussian Process prior—that of function smoothness—can readily be ensured with the appropriate design of policy spaces. To illustrate GPR in surrogate-based optimization, Figure 5.2 depicts a hypothetical 1D instructional policy space, along with the true PPF and the GPR posterior conditioned on the outcome of a set of single-subject experiments at various points in policy space.

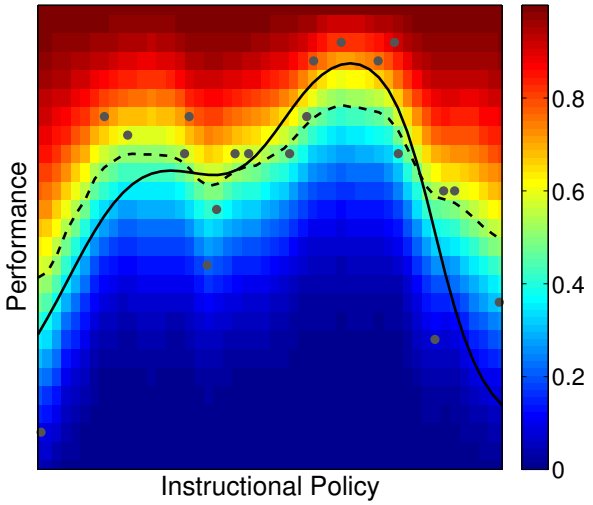


Figure 5.2: A hypothetical 1D instructional policy space. The solid black line represents an (unknown) policy performance function. The grey disks indicate the noisy outcome of single-subject experiments conducted at specified points in policy space. (The diameter of the disk represents the number of data points occurring at the disk's location.) The dashed black line depicts the GP posterior mean, and the coloring of each vertical strip represents the cumulative density function for the posterior.

5.2.2 Generative model of student performance

Each instructional policy is presumed to have an inherent effectiveness for a population of individuals. However, a policy's effectiveness can be observed only indirectly through measurements of subject performance such as the number of correct responses. To determine the most effective policy from noisy observations, we must specify a generative model of student performance which relates the inherent effectiveness of instruction to observed performance.

Formally, each subject s is trained under a policy $\mathbf{x}_s$ and then tested to evaluate his or her performance. We posit that each training policy $\mathbf{x}$ has a latent population-wide effectiveness $f_{\mathbf{x}} \in \mathbb{R}$ and that how well a subject performs on the test is a noisy function of $f_{\mathbf{x}_s}$. We are interested in predicting the effectiveness of a policy $\mathbf{x}'$ across a population of students given the observed test scores of S subjects trained under the policies $\mathbf{x}_{1:S}$. Conceptually, this involves first inferring the effectiveness $\mathbf{f}$ of policies $\mathbf{x}_{1:S}$ from the noisy test data, then interpolating from $\mathbf{f}$ to $f_{\mathbf{x}'}$.

Using a standard Bayesian nonparametric approach, we place a mean-zero Gaussian Process

prior over the function $f_{\mathbf{x}}$. For the finite set of S observations, this corresponds to the multivariate normal distribution $\mathbf{f} \sim \text{MVN}(0, \Sigma)$, where Σ is a covariance matrix prescribing how smoothly varying we expect f to be across policies. We use the squared-exponential covariance function, so that $\Sigma_{s,s'} = \sigma^2 \exp(-\frac{||\mathbf{x}_s - \mathbf{x}_{s'}||^2}{2\ell^2})$, and σ^2 and ℓ are free parameters.

Having specified a prior over policy effectiveness, we turn to specifying a distribution over observable measures of subject learning conditioned on effectiveness. In this paper, we measure learning by administering a multiple-choice test to each subject s and observing the number of correct responses s made, c_s, out of n_s questions. We assume the probability that subject s answers any question correctly is a random variable μ_s whose expected value is related to the policy's effectiveness via the logistic transform: $\mathbb{E}[\mu_s] = \text{logistic}(o + f_{\mathbf{x}_s})$ where o is a constant. This is consistent with the observation model

$$\mu_s \mid f_{\mathbf{x}_s}, o, \gamma \sim \text{Beta}(\gamma, \gamma e^{-(o+f_{\mathbf{x}_s})}) \qquad c_s \mid \mu_s \sim \text{Binomial}(g + (1-g)\mu_s;\ n_s) \tag{5.2}$$

where γ controls inter-subject variability in μ_s and g is the probability of answering a question correctly by random guessing. In this paper, we assume $g = .5$. For this special case, the analytic marginalization over μ_s yields

$$P(c_s \mid f_{\mathbf{x}_s}, \gamma, o, g = .5) = 2^{-n_s} \binom{n_s}{c_s} \sum_{i=0}^{c_s} \binom{c_s}{i} \frac{\text{B}(\gamma + i, n_s - c_s + \gamma e^{-(o+f_{\mathbf{x}_s})})}{\text{B}(\gamma, \gamma e^{-(o+f_{\mathbf{x}_s})})} \tag{5.3}$$

where $\text{B}(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function. Equation 5.3 follows from the likelihood being

$$L(c_s | n_s, \mu_s, g = .5) \quad = \quad \binom{n_s}{k} .5^{n_s} (1+\mu)^{c_s} (1-\mu)^{n_s - c_s} \tag{5.4}$$

and the beta prior being

$$\pi(\mu \mid \gamma, \beta) = \frac{1}{\text{B}(\gamma, \beta)} \mu^{\gamma - 1} (1-\mu)^{\beta - 1} \tag{5.5}$$

where B is the beta function and $\beta \triangleq \gamma e^{-(o+f_{\mathbf{x}_s})}$. The marginal likelihood is defined as

$$P(c_s | \gamma, \beta, n_s) \quad = \quad \int_0^1 L(c_s | n_s, \mu, g)\ \pi(\mu | \gamma, \beta)\ d\mu \tag{5.6}$$

$$= \quad 2^{-n_s} \frac{1}{\text{B}(\gamma, \beta)} \binom{n_s}{c_s} \int_0^1 (1+p)^{c_s} p^{\gamma - 1} (1-p)^{\beta - 1 + n_s - c_s}\ d\mu \tag{5.7}$$

Because c_s is an integer, we can apply the binomial theorem

$$P(c_s|\gamma,\beta,n_s) = 2^{-n_s}\frac{1}{\mathrm{B}(\gamma,\beta)}\binom{n_s}{c_s}\int_0^1\sum_{i=0}^{c_s}\binom{c_s}{i}\mu^i\ \mu^{\gamma-1}(1-\mu)^{\beta-1+n_s-c_s}\ d\mu \tag{5.8}$$

$$= 2^{-n_s}\frac{1}{\mathrm{B}(\gamma,\beta)}\binom{n_s}{c_s}\sum_{i=0}^{c_s}\binom{c_s}{i}\int_0^1\mu^{\gamma+i-1}(1-\mu)^{\beta-1+n_s-c_s}\ d\mu \tag{5.9}$$

The integral in the summation is over an unnormalized Beta$(\gamma+i, n_s+\beta-k)$ density, thus yielding Equation 5.3.

Parameters $\boldsymbol{\theta} \equiv \{\gamma, o, \sigma^2, \ell\}$ are given uniform priors over a large range. The effectiveness of a policy $\mathbf{x}'$ given the number of correct responses made by a set of subjects, $\mathbf{c}$, is estimated via $p(\mathbf{f}_{\mathbf{x}'} \mid \mathbf{c}) \approx \frac{1}{M}\sum_{m=1}^{M} p(\mathbf{f}_{\mathbf{x}'} \mid \mathbf{f}^{(m)}, \boldsymbol{\theta}^{(m)})$, where $p(\mathbf{f}_{\mathbf{x}'} \mid \mathbf{f}^{(m)}, \boldsymbol{\theta}^{(m)})$ is Gaussian with mean and variance determined by the mth sample from the posterior $p(\mathbf{f}, \boldsymbol{\theta} \mid \mathbf{c})$. We are interested in drawing samples from the posterior over $\mathbf{f}$. By Bayes rule, the target distribution for our Markov chain Monte Carlo (MCMC) algorithm is the product $\mathrm{P}(\mathbf{f} \mid \mathbf{c}) \propto \mathrm{P}(\mathbf{c} \mid \mathbf{f})\ \mathrm{p}(\mathbf{f})$ Designing an efficient sampling strategy is difficult in many Gaussian process applications because the posterior describes a highly correlated high-dimensional variable (Titsias, Lawrence, & Rattray, 2008). The MCMC technique we used to draw from the posterior is called *elliptical slice sampling* (I. Murray, Adams, & MacKay, 2010). Elliptical slice sampling is a black-box technique that mixes quickly for models with a complicated likelihood function and a latent multivariate normal variable. For more details, see I. Murray et al. (2010).

We have also explored a more general framework that relaxes the relationship between chance-guessing and test performance and allows for multiple policies to be evaluated per subject. With regard to the latter, subjects may undergo multiple randomly ordered blocks of trials where in each block b a subject s is trained under a policy $f_{\mathbf{x}_s^b}$ and then tested. The observation model is altered so that the score in a block is given by $c_s^b \sim \text{Binomial}(\mu_s^b; n_s^b)$ where $\mu_s^b \equiv \text{logistic}(o' + \alpha_s + f_{\mathbf{x}_s^b})$, the factor $\alpha_s \sim \text{Normal}(0, \tau_\alpha^{-1})$ represents the ability of subject s across blocks, and the constant o' subsumes the role of o and g from the original model. In the spirit of item-response theory (Drasgow & Hulin, 1990), the model could be extended further to include factors that represent the difficulty

of individual test questions and interactions between subject ability and question difficulty.

5.2.3 Active selection

GP optimization requires a strategy for actively selecting the next experiment. (We refer to this as a 'strategy' instead of as a 'policy' to avoid confusion with instructional policies.) Many heuristic strategies have been proposed (Forrester & Keane, 2009), including: **grid sampling** over the policy space; expanding or contracting a **trust region**; and **goal-setting** approaches that identify regions of policy space where performance is likely to attain some target level or beat out the current best experiment result. In addition, greedy versus k-step predictive planning has been considered (Osborne, Garnett, & Roberts, 2009).

Every strategy faces an exploration/exploitation trade off. Exploration involves searching regions of the function with the maximum uncertainty; exploitation involves concentrating on the regions of the function that currently appear to be most promising. Each has a cost. A focus on exploration rapidly exhausts the subject budget for subjects. A focus on exploitation leads to selection of local optima.

The upper-confidence bound (UCB) strategy (Forrester & Keane, 2009; Srinivas, Krause, Kakade, & Seeger, 2010) attempts to avoid these two costs by starting in an exploratory mode and shifting to exploitation. This strategy chooses the most-promising experiment from an upper-confidence bound on the GPR: $\mathbf{x}_t = \text{argmax}_{\mathbf{x}} \ \hat{\mu}_{t-1}(\mathbf{x}) + \eta_t \hat{\sigma}_{t-1}(\mathbf{x})$, where t is a time index, $\hat{\mu}$ and $\hat{\sigma}$ are the mean and standard deviation of the GPR, and η_t controls the exploration/exploitation trade off. Large η_t focus on regions with the greatest uncertainty, but as $\eta_t \to 0$, the focus shifts to exploitation in the neighborhood of the current best policy. Annealing η_t as a function of t will yield exploration initially shifting toward exploitation.

We adapt the UCB strategy by transforming the UCB based on the GPR to an expression based on the the population accuracy (proportion correct) via $\mathbf{x}_t = \text{argmax}_{\mathbf{x}} P(\frac{c_s}{n_s} > \nu_t \mid f_{\mathbf{x}})$, where ν_t is an accuracy level determining the exploration/exploitation trade off. In simulations, we found that setting $\nu_t = .999$ was effective. Note that in applying the UCB selection strategy, we must

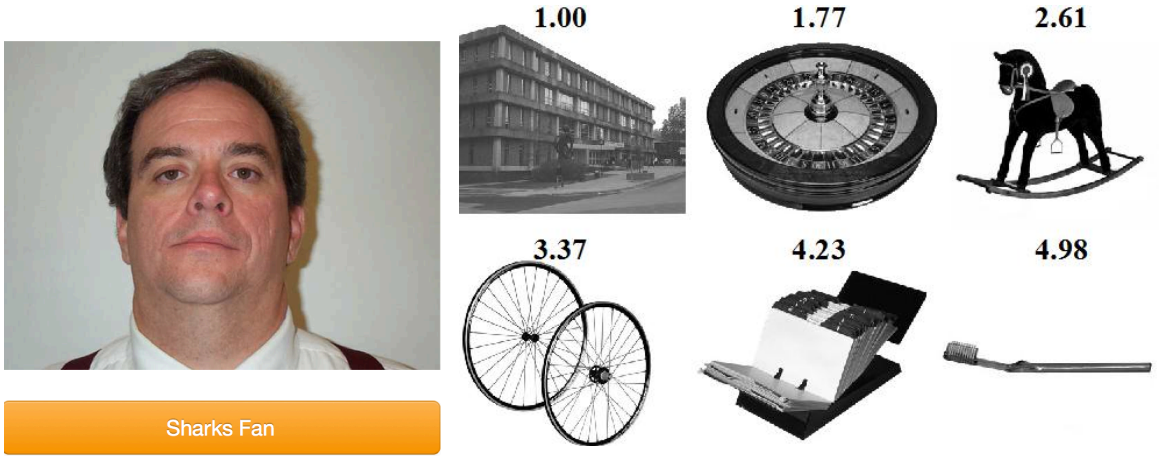


Figure 5.3: (left) Experiment 1 training display; (right) Example stimuli used in Experiment 2, along with their graspability ratings: 1 means not graspable and 5 means highly graspable.

search over a set of candidate policies. We applied a fine uniform grid search over policy space to perform this selection.

5.2.4 Experiment 1: Presentation rate optimization

De Jonge, Tabbers, Pecher, and Zeelenberg (2012) studied the effect of presentation rate on word-pair learning. During training, each pair was viewed for a total of 16 sec. Viewing was divided into $16/d$ trials each with a duration of d sec, where d ranged from 1 sec (viewing the pair 16 times) to 16 sec (viewing the pair once). de Jong et al. found that an intermediate duration yielded better cued recall performance both immediately and following a delay.

We explored a variant of this experiment in which subjects were asked to learn the favorite sporting team of six individuals. During training, each individual's face was shown along with their favorite team—either Jets or Sharks (Figure 5.3, left). The training policy specifies the duration d of each face-team pair. Training was over a 30 second period, with a total of $30/d$ trials and an average of $5/d$ presentations per face-team pair. Presentation sequences were blocked, where a block consists of all six individuals in random order. Immediately following training, subjects were tested on each of the six faces in random order and were asked to select the corresponding team. The training/testing procedure was repeated for eight rounds each using different faces. In

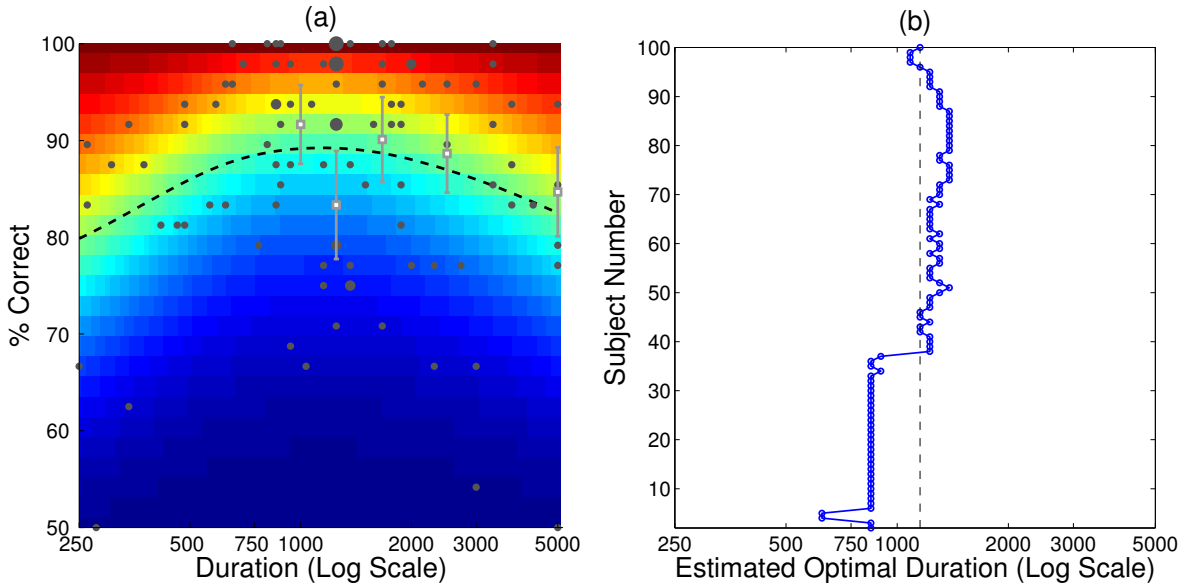


Figure 5.4: Experiment 1 results. (a) Posterior density of the PPF with 100 subjects. Light grey squares with error bars indicate the results of a traditional comparison among conditions. (b) Prediction of optimum presentation duration as more subjects are run; dashed line is asymptotic value.

total, each subject responded to 48 faces. The faces were balanced across ethnicity, age, and gender (provided by Minear & Park, 2004).

Using Mechanical Turk, we recruited 100 subjects who were paid \$0.30 for their participation. The policy space was defined to be in the logarithm of the duration, i.e., $d = \mathrm{e}^x$, where $x \in$ [ln(.25) ln(5)]. The space included only values of x such that $30/d$ is an integer; i.e., we ensured that no trials were cut short by the 30 second time limit. Subject 1's training policy, x_1, was set to the median of the range of admissible values (857 ms). After each subject t completed the experiment, the PPF posterior was reestimated, and the upper-confidence bound strategy was used to select the policy for subject $t+1$, x_{t+1}.

Figure 5.4a shows the PPF posterior based on 100 subjects. The diameter of the grey disks indicate the number of data points observed at that location in the space. The optimum of the PPF mean is at 1.15 sec, at which duration each face-team pair will be shown on expectation 4.33 times during training. Though the result seems intuitive, we have polled colleagues, and predictions

for the peak ranged from below 1 sec to 2.5 sec. Figure 5.4b uses the PPF mean to estimate the optimum duration, and this duration is plotted against the number of subjects. Our procedure yields an estimate for the optimum duration that is quite stable after about 40 subjects.

Ideally, one would like to compare the PPF posterior to ground truth. However, obtaining ground truth requires a massive data collection effort. We provide alternative evidence of two forms. First, to verify that the PPF posterior is sensitive to the data collected, we created a synthetic data set by randomly re-pairing policies and scores from the actual data set. This synthetic data set produced flat PPFs, quite different in shape than the unimodal PPF in Figure 5.4. Second, as an alternative, we contrast our result with a more traditional experimental study based on the same number of subjects. We ran 100 additional subjects in a standard experimental design involving evaluation of five alternative policies, $d \in \{1, 1.25, 1.667, 2.5, 5\}$, 20 subjects per policy. (These durations correspond to 1-5 presentations of each face-team pair during training.) The mean score for each policy is plotted in Figure 5.4a as light grey squares with bars indicating ± 2 standard errors of the mean. The result of the traditional experiment is coarsely consistent with the PPF posterior, but the budget of 100 subjects places a limitation on the interpretability of the results. When matched on budget, the optimization procedure appears to produce results that are more interpretable and less sensitive to noise in the data. Note that we have biased this comparison in favor of the traditional design by restricting the exploration of the policy space to the region 1 sec $\leq d \leq$ 5 sec. Nonetheless, no clear pattern emerges in the shape of the PPF based on the outcome of the traditional design.

5.2.5 Experiment 2: Training sequence optimization

In Experiment 2, we study concept learning from examples. Subjects are told that Martians will teach them the meaning of a Martian adjective, GLOPNOR, by presenting a series of example objects, some of which have the property GLOPNOR and others do not. During a training phase, objects are presented one at a time and subjects must classify the object as GLOPNOR or NOT-GLOPNOR. They then receive feedback as to the correctness of their response. On each trial,

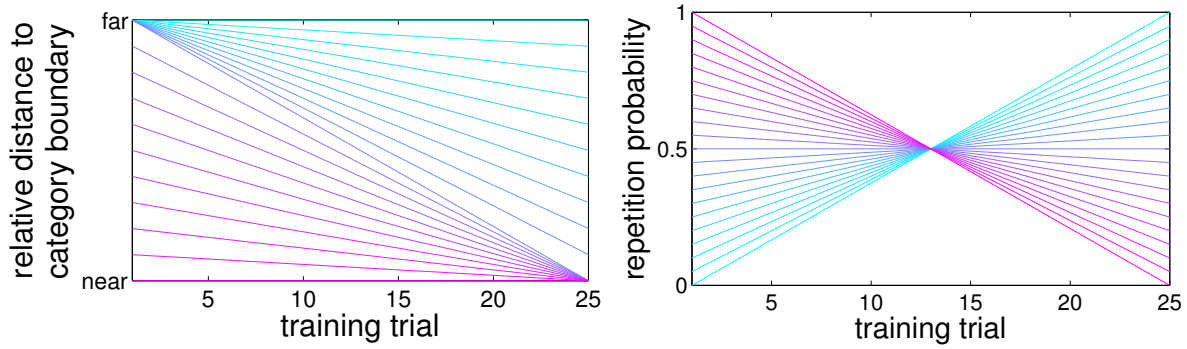


Figure 5.5: Experiment 2, trial dependent fading and repetition policies (left and right, respectively). Colored lines represent specific policies.

the object from the previous trial is shown in the corner of the display along with its correct classification, the reason for which is to facilitate comparison and contrasting of objects. Following 25 training trials, 24 test trials are administered in which the subject makes a classification but receives no feedback. The training and test trials are roughly balanced in number of positive and negative examples.

The stimuli in this experiment are drawn from a set of 320 objects normed by Salmon, McMullen, and Filliter (2010) for **graspability**, i.e., how manipulable an object is according to how easy it is to grasp and use the object **with one hand**. They polled 57 individuals, each of whom rated each of the objects multiple times using a 1–5 scale, where 1 means not graspable and 5 means highly graspable. Figure 5.3 shows several objects and their ratings. We divided the objects into two groups by their mean rating, with the NOT-GLOPNOR group having ratings in [1, 2.75] and the GLOPNOR group having ratings in [3.25, 5]. (We discarded objects with ratings in [2.75, 3.25] because they are too difficult even if one knows the concept). The classification task is easy if one knows that the concept is graspability. However, the challenge of inferring the concept is extremely difficult because there are many dimensions along which these objects vary and any one—or more—could be the classification dimension(s).

We defined an instructional policy space characterized by two dimensions: **fading** and **blocking**. Fading refers to the notion from the animal learning literature that learning is facilitated by

presenting exemplars far from the category boundary initially, and gradually transitioning toward more difficult exemplars over time. Exemplars far from the boundary may help individuals to attend to the dimension of interest; exemplars near the boundary may help individuals determine where the boundary lies (Pashler & Mozer, 2013). Theorists have also made computational arguments for the benefit of fading (Bengio, Louradour, Collobert, & Weston, 2009; Khan et al., 2011). Blocking refers to the issue discussed in the Introduction concerning the sequence of category labels: Should training exemplars be blocked or interleaved? That is, should the category label on one trial tend to be the same as or different than the label on the previous trial?

For fading, we considered a family of trial-dependent functions that specify the distance of the chosen exemplar to the category boundary (left panel of Figure 5.5). This family is parameterized by a single policy variable x_2, $0 \leq x_2 \leq 1$ that relates to the distance of an exemplar to the category boundary, d, as follows: $d(t, x_2) = \min(1, 2x_2) - (1 - |2x_2 - 1|)\frac{t-1}{T-1}$, where T is the total number of training trials and t is the current trial. For blocking, we also considered a family of trial-dependent functions that vary the probability of a category label repetition over trials (right panel of Figure 5.5). This family is parameterized by the policy variable x_1, $0 \leq x_1 \leq 1$, that relates to the probability of repeating the category label of the previous trial, r, as follows: $r(t, x_1) = x_1 + (1 - 2x_1)\frac{t-1}{T-1}$.

Figure 5.6a provides a visualization of sample training trial sequences for different points in the 2D policy space. Each graph represents an instance of a specific (probabilistic) policy. The abscissa of each graph is an index over the 25 training trials; the ordinate represents the category label and its distance from the category boundary. Policies in the top and bottom rows show sequences of all-easy and all-hard examples, respectively; intermediate rows achieve fading in various forms. Policies in the left-most column begin training with many repetitions and end training with many alternations; policies in the right-most column begin with alternations and end with repetitions; policies in the middle column have a time-invariant repetition probability of 0.5.

Regardless of the training sequence, the set of test objects was the same for all subjects. The test objects spanned the spectrum of distances from the category boundary. During test, subjects

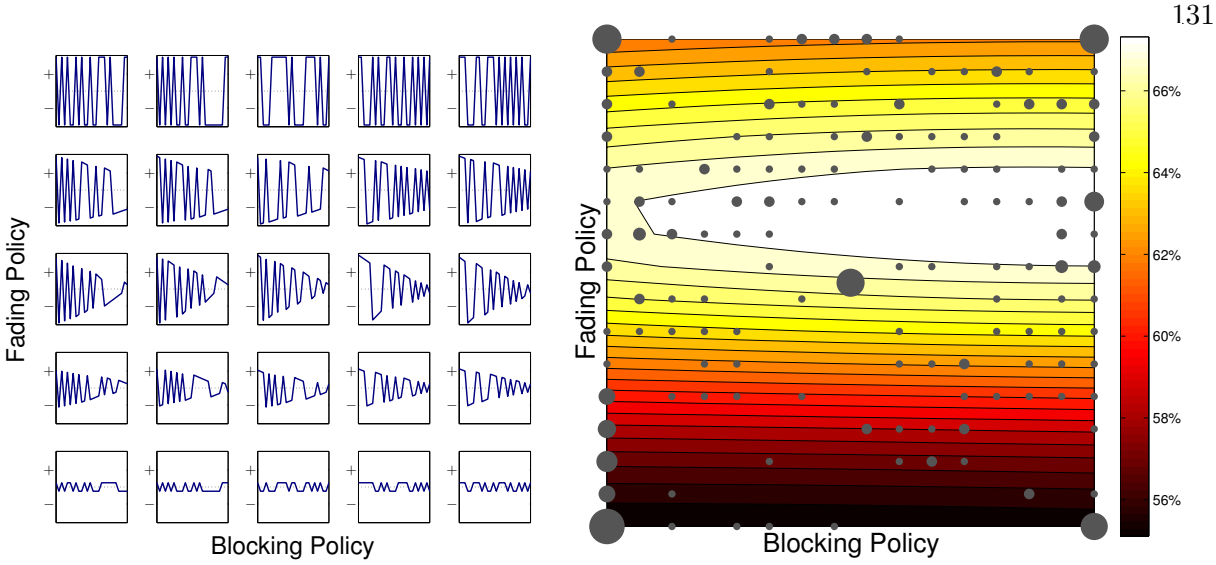


Figure 5.6: Experiment 2 (a) policy space and (b) policy performance function at 200 subjects

were required to make a forced choice GLOPNOR/NOT-GLOPNOR judgment.

We seeded the optimization process by running 10 subjects in each of four corners of policy space as well as in the center point of the space. We then ran 150 additional subjects using GP-based optimization. Figure 5.6 shows the PPF posterior mean over the 2D policy space, along with the selection in policy space of the 200 subjects. Contour map colors indicate the expected accuracy of the corresponding policy (in contrast to the earlier colored graphs in which the coloring indicates the cdf). The optimal policy is located at $\mathbf{x}^* = (1, .66)$.

To validate the outcome of this exploration, we ran 50 subjects at $\mathbf{x}^*$ as well as policies in the upper corners and the center of Figure 5.6. Consistent with the prediction of the PPF posterior, mean accuracy at $\mathbf{x}^*$ is 68.6%, compared to 60.9% for $(0, 1)$, 65.7% for $(1, 0)$, and 66.6% for $(.5, .5)$. Unfortunately, only one of the paired comparisons was statistically reliable by a two-tailed Bonferroni corrected t-test: $(0, 1)$ versus $\mathbf{x}^*$ ($p = .027$). However, post-hoc power computation revealed that with 50 subjects and the variability inherent in the data, the odds of observing a reliable 2% difference in the mean is only .10. Running an additional 50 subjects would raise the power to only .17. Thus, although we did not observe a statistically significant improvement at

the inferred optimum compared to sensible alternative policies, the results are consistent with our inferred optimum being an improvement over the type of policies one might have proposed a priori.

5.2.6 Discussion

The traditional experimental paradigm in psychology involves comparing a few alternative conditions by testing a large number of subjects in each condition. We have described a novel paradigm in which a large number of conditions are evaluated, each with only one or a few subjects. Our approach achieves an understanding of the functional relationship between conditions and performance, and it lends itself to discovering the conditions that attain optimal performance.

Experiments 1 and 2 focused on the problem of optimizing instruction, but the method described here has broad applicability across issues in the behavioral sciences. For example, one might attempt to maximize a worker's motivation by manipulating rewards, task difficulty, or time pressure. Motivation might be studied in an experimental context with voluntary time on task as a measure of intrinsic interest level.

Consider problems in a quite different domain, human vision. Optimization approaches might be used to determine optimal color combinations in a manner more efficient and feasible than exhaustive search (Schloss & Palmer, 2011). Also in the vision domain, one might search for optimal sequences and parameterizations of image transformations that would support complex visual tasks performed by experts (e.g., x-ray mammography screening) or ordinary visual tasks performed by the visually impaired.

From a more applied angle, A-B testing has become an extremely popular technique for fine tuning web site layout, marketing, and sales (Christian, 2012). With a large web population, two competing alternatives can quickly be evaluated. Our approach offers a more systematic alternative in which a space of alternatives can be explored efficiently, leading to discovery of solutions that might not have been conceived of as candidates a priori.

The present work did not address individual differences or high-dimensional policy spaces, but our framework can readily be extended. Individual differences can be accommodated via policies

that are parameterized by individual variables (e.g., age, education level, performance on related tasks, recent performance on the present task). For example, one might adopt a fading policy in which the rate of fading depends in a parametric manner on a running average of performance. High dimensional spaces are in principle no challenge for GPR, given a sensible distance metric. The challenge of high-dimensional spaces comes primarily from computational overhead in selecting the next policy to evaluate. However, this computational burden can be greatly relaxed by switching from a global optimization perspective to a local perspective: instead of considering candidate policies in the entire space, active selection might consider only policies in the neighborhood of previously explored policies.

5.3 Other human optimization tasks

The previous section focused on using our experimental paradigm to find optimal instructional strategies. In this section, we present experiments demonstrating the paradigm's applicability to an optimization task involving human decision-making, and we also provide evidence for its usefulness in modeling aesthetic judgements.

5.3.1 Experiment 3: Donation optimization

Mechanical Turk subjects from the United States participated in Experiment 3 under the pretense of answering a question about soft drink preferences—whether they prefer Coca-Cola or Pepsi—in return for a payment of 2 cents. After indicating their preference, subjects were offered an unanticipated bonus payment of 10 cents (see Figure 5.7). Subjects were given the option of donating some of their bonus payment to a charity. They could either select one of three suggested donation amounts or enter a "custom" donation amount ranging from 0 to 10 cents into a text box.

The policy space for this experiment consists of the three suggested donation amounts, and the goal of the optimization search is to efficiently discover the policy that maximizes the population-wide expected amount of money donated to charity. If the suggested donations are small, many subjects may make a donation, but each donation may tend to be small. Similarly, if the suggested

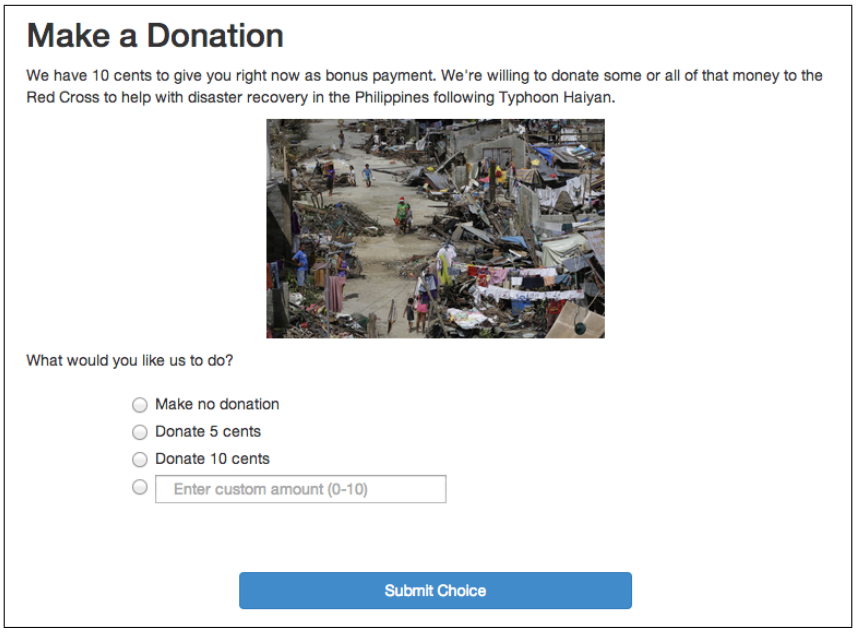


Figure 5.7: In Experiment 3, subjects were lured in on the pretense of answering a survey question about soft drink preferences. After answering the survey question, they were presented with the above dialogue which offered a 10 cent bonus and gave the option to forgo some or all of the bonus by making a donation to charity. Our technique iteratively searched over the space of all possible suggested donation amounts with the goal of finding the suggestions that maximize the expected amount of money donated.

donations are large, few subjects may make a donation, but each donation may tend to be large.

The likelihood model used for this experiment differs from that of Experiments 1 and 2 because the observations are qualitatively different. Let $\mathbf{x}$ denote a policy, a three dimensional vector of suggested donation amounts. Let $d_s \in 0, 1, \ldots, d_{\max}$ be the number of cents subject s donates, where $d_{\max}$ is the maxium possible donation amount. We use an ordered probit observation model,

$$d_s \mid f(\mathbf{x}_s), \epsilon_s = \begin{cases} 0 & \text{if } f(\mathbf{x}_s) + \epsilon_s < .5 \\ 1 & \text{if } .5 \leq f(\mathbf{x}_s) + \epsilon_s < 1.5 \\ \vdots & \vdots \\ d_{\max} - 1 & \text{if } d_{\max} - 1.5 \leq f(\mathbf{x}_s) + \epsilon_s < d_{\max} - .5 \\ d_{\max} & \text{if } d_{\max} - .5 \leq f(\mathbf{x}_s) + \epsilon_s \end{cases} \tag{5.10}$$

where $\epsilon_s \sim \text{Normal}(0, \sigma^2)$ is intersubject noise and f is the Gaussian process distributed prior. For

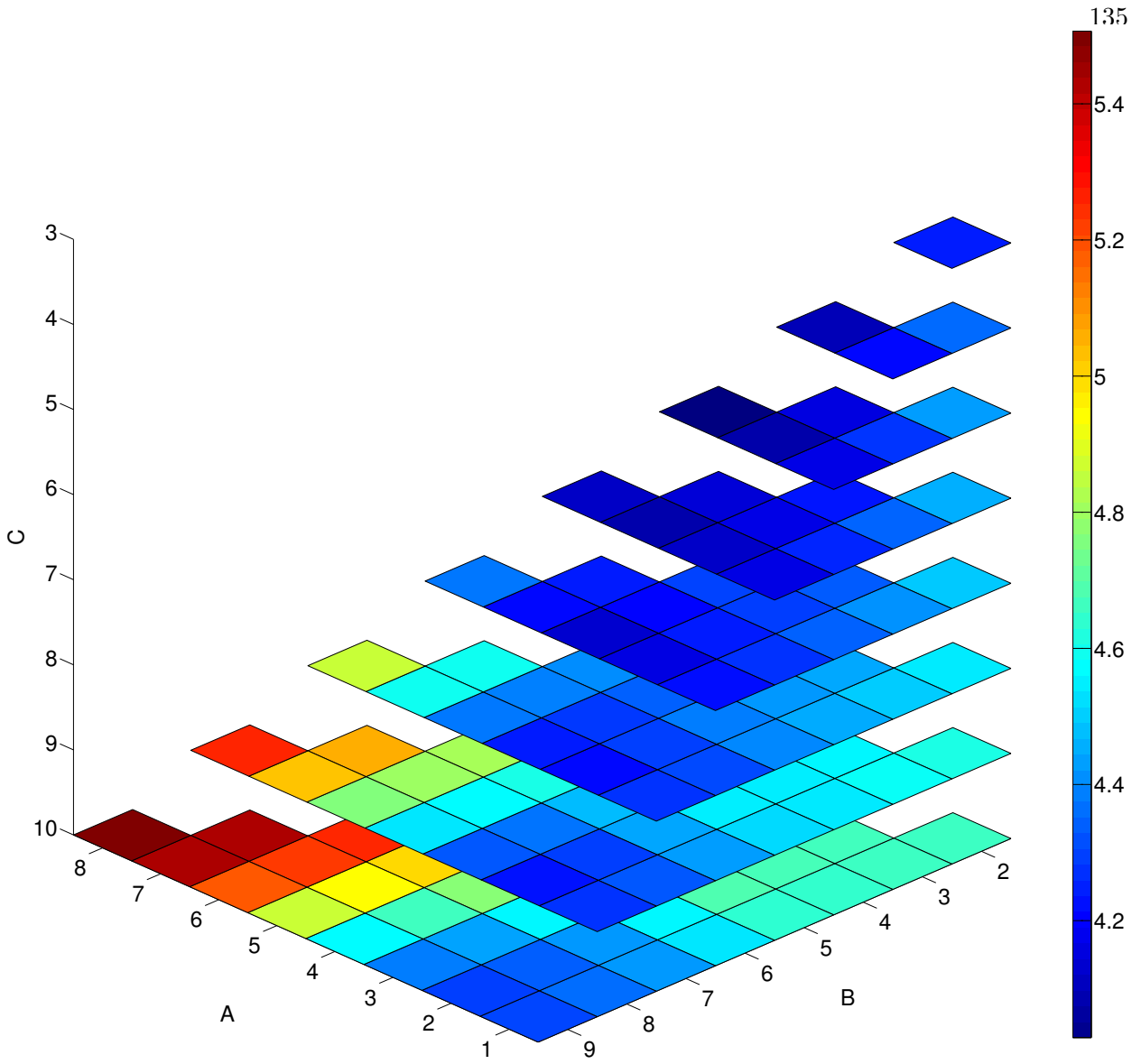


Figure 5.8: A visualization of the estimated policy performance function from Experiment 3 after 200 subjects. The axes A, B, and C correspond to the first, second, and third suggested donation amounts, respectively. Because $A < B < C$ and the suggested amounts are natural numbers, the policy space forms a pyramidal structure. The coloring of each location indicates the expected average number of cents a population of subjects will donate when presented with the corresponding policy. The optimal policy is to suggest that subjects donate 8, 9, or 10 cents.

posterior inference, we again used elliptical slice sampling.

The experiment's procedure was the same as that of Experiments 1 and 2. One subject was run at a time. Each subject's policy was chosen through an upper confidence bound selection rule using predictions conditioned on the observed donation amounts from the subjects run to date.

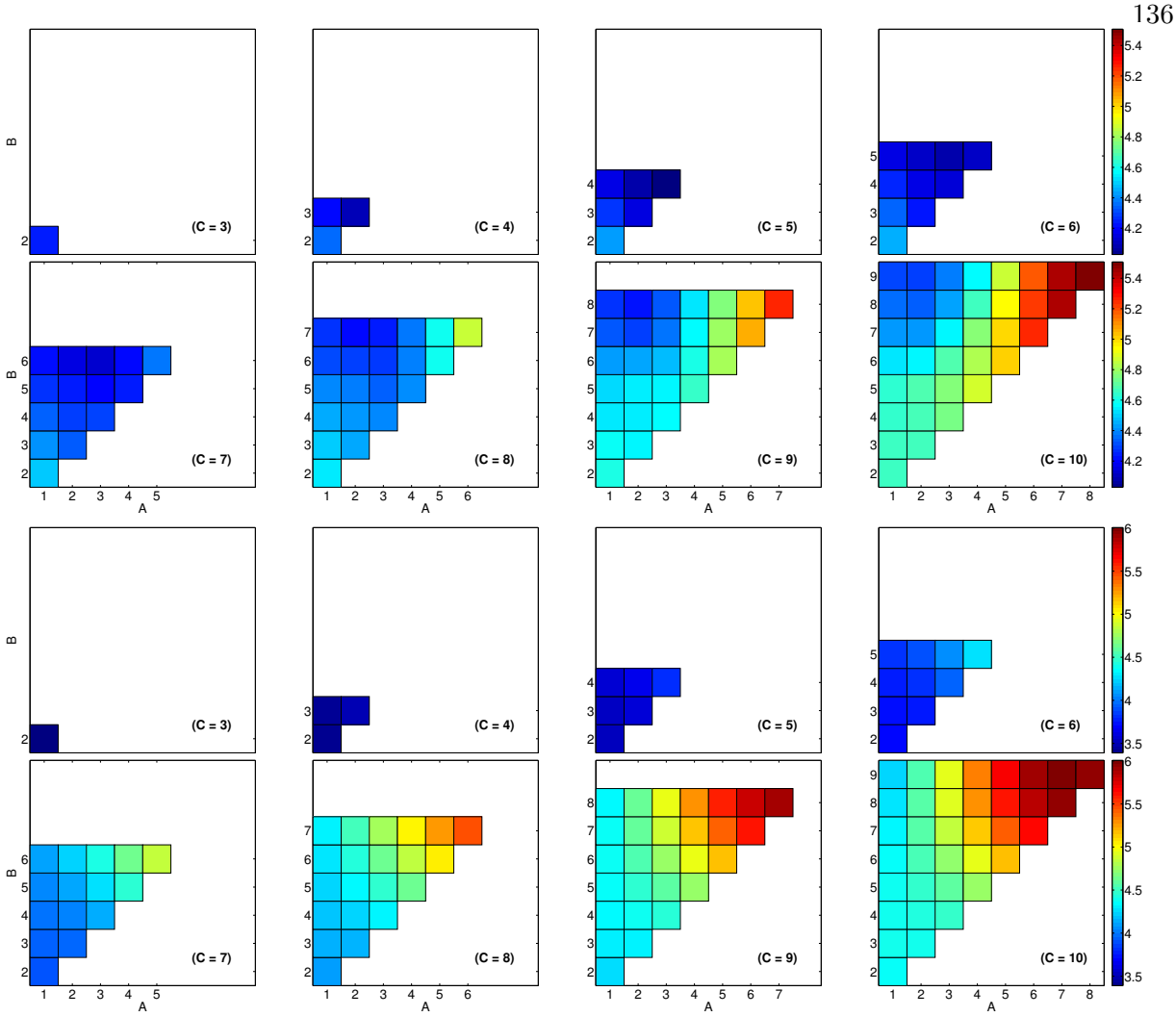


Figure 5.9: A comparison of results from Experiment 3 (rows 1 and 2) and a replication of the experiment involving non-U.S. citizens only (rows 3 and 4). Each graph shows the expected donation amount as a function of A and B for a fixed value of C. The optimal policy appears to have been unaffected by the change in demographics.

Results from the experiment after 200 subjects are presented in Figure 5.8. The three axes—labeled A, B, and C—indicate the first, second, and third suggested donation amounts, respectively. Each grid square corresponds to a particular policy. The coloring of each location indicates the expected average number of cents the population will donate when presented with the corresponding policy.

Despite the subtleness of the experimental manipulation, there are large differences in the effectiveness of the policies. The optimal policy is to suggest that subjects donate 8, 9, or 10 cents—the maximum possible suggestions in this task. The worst policies are generally those that

have a small minimum suggestion amount, even though such suggestions increase the probability that an individual subject will make a non-zero donation.

The subject population used in Experiment 3 was restricted to the United States. The motivation for this choice was to remove a potential source of noise from the experiment: we hypothesized that the degree to which subjects value 10 cents depends on their standard of living, and that most non-U.S. Mechanical Turk users are from countries with lower standards of living than that in the United States. We later repeated Experiment 3, but with a subject population restricted to non-U.S. citizens. Results from this replication study are shown in Figure 5.9 after 93 subjects. For this experiment, the optimal policy does not appear to be sensitive to the country of origin of the subjects.

5.3.2 Vision

Another potential application domain of our paradigm involves human vision. Psychologists who investigate the desirability of color combinations may do so by exhaustive search (Schloss & Palmer, 2011). For instance, the experimental paradigm for evaluating preferences for color pairs involves first defining a set of colors, then asking subjects to rate their preference for all possible color pairs in the Cartesian product of that set. This exhaustive procedure is very time consuming and limits the number of colors that can be considered. It also practically limits the experimenter to considering only color pairs, since the space of all possible color triplets or quadruplets is vast even with a relatively small set of colors.

Formally, we are interested in using preference judgements to infer the desirability of color combinations and to be able to interpolate to new, unjudged color combinations. We are presenting a sequence of color stimuli $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ and collecting preference judgements $y_1, y_2, \ldots, y_N$ (Figure 5.10). Each judgement is on a bounded rating scale, $y_i \in [a, b]$ (e.g., 0-5 stars). We assume that each stimulus x_i has an associated unobservable latent affinity or quality $f(\mathbf{x}_i) \in \mathbb{R}$, and preference judgements are a noisy mapping from the Gaussian process distributed $f(\mathbf{x}_i)$ to the rating scale. The rating scale limits the expressivity of the judgements: if a subject has a preference for a pair,

he or she may give it a rating at the maximum; if the subject has a very strong preference for the pair, he or she will be forced to give the same maximal rating. The rating scale "clips" y to the range $[a, b]$. Thus, the likelihood we use is a normal distribution bounded to $[a, b]$ so that all the probability mass above b gets moved to b and all the probability mass below a gets put at a. This gives the mixed distribution

$$p(y_i|f, x_i) = \Phi\left(\frac{a - f(x_i)}{\sigma}\right)\mathbb{1}_{y_i=a} + \sigma^{-1}\mathcal{N}\left(\frac{y_i - f(x_i)}{\sigma}\right)\mathbb{1}_{a<y_i<b} + \Phi\left(\frac{f(x_i) - b}{\sigma}\right)\mathbb{1}_{y_i=b} \quad (5.11)$$

where $\mathcal{N}$ denotes the standard normal probability density function, Φ denotes the standard normal cumulative density function, σ^2 is the noise variance, and $\mathbb{1}$ is the indicator function. By assumption, observations are conditionally independent given f. Thus, the full data likelihood factorizes as $p(\mathbf{y}|\mathbf{f}, X) = \prod_i p(y_i|f(x_i))$. For posterior inference, we use Laplace's method, which utilizes a Gaussian approximation to the intractable posterior $p(\mathbf{f}|X, \mathbf{y})$ through a second-order Taylor approximation.

Using this model on a dataset from Schloss and Palmer (2011) wherein subjects rate their affinity for pairs of colors, we can make predictions about the optimality of color combinations, even if the colors involved have not been tested. For example, in Figure 5.11, we systematically varied one of the colors—this is shown in the background. For each background color, we used the model to predict what the corresponding best and worst matching color would be. The predictions are shown as the smaller squares. This method for interpolating across subjects' preferences to new, unseen colors shows promise in allowing researchers to systematically explore larger color spaces.

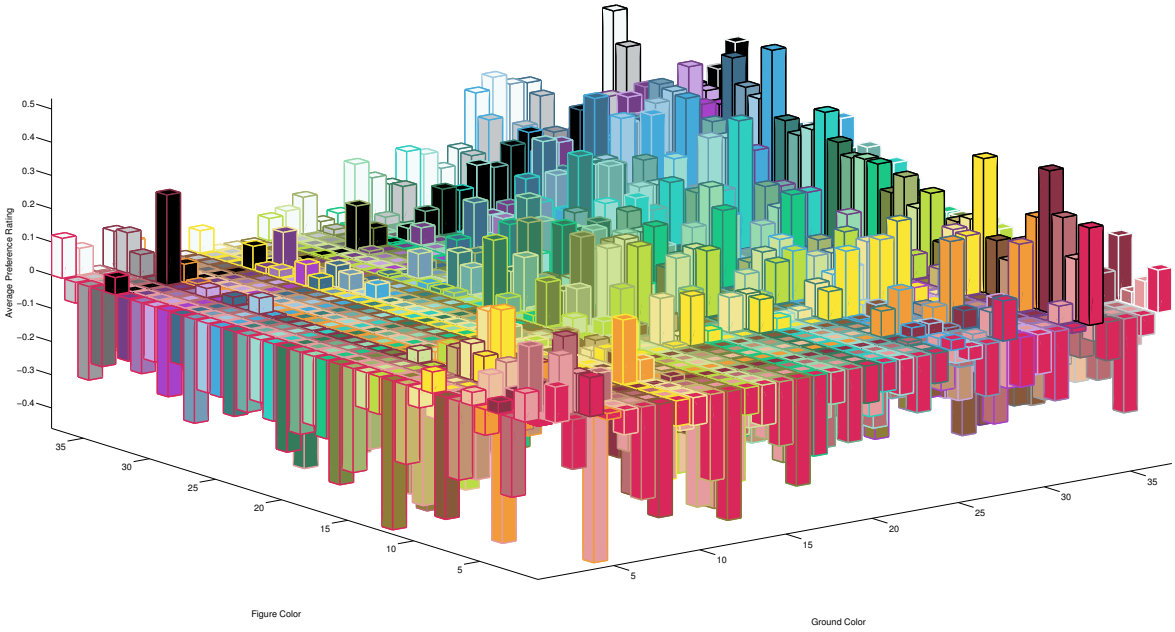


Figure 5.10: A visualization of the color preference ratings dataset. Each bar represents a particular color pair. The edges of a bar represent one color from the pair, and the interior color represents the other color from the pair. Each subject rated his or her preference for every color pair shown. The height of each bar represents the across-subject average preference.

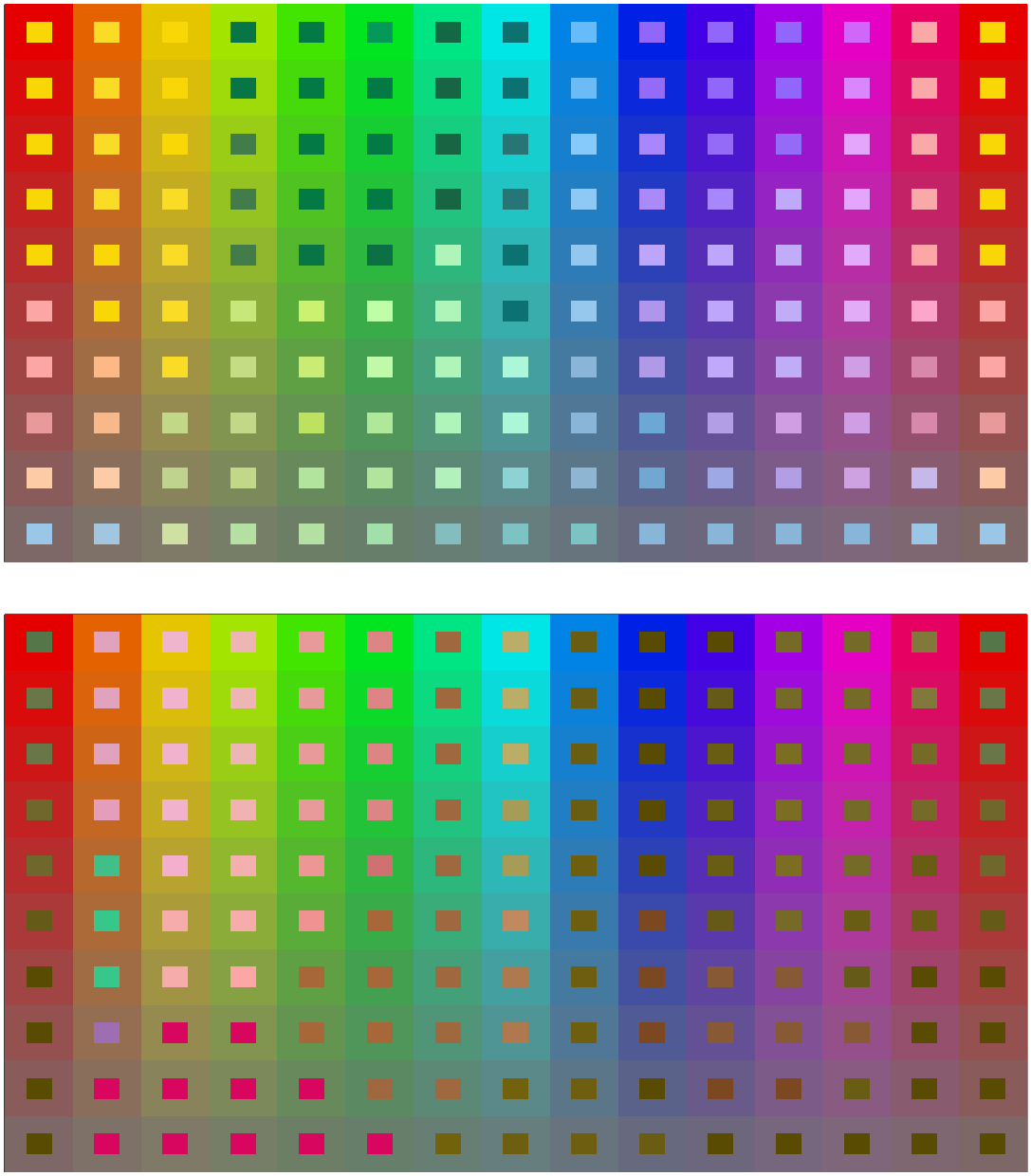


Figure 5.11: Predicted most and least *preferred* color pairings for a fixed ground lightness level with varying hue and saturation levels.

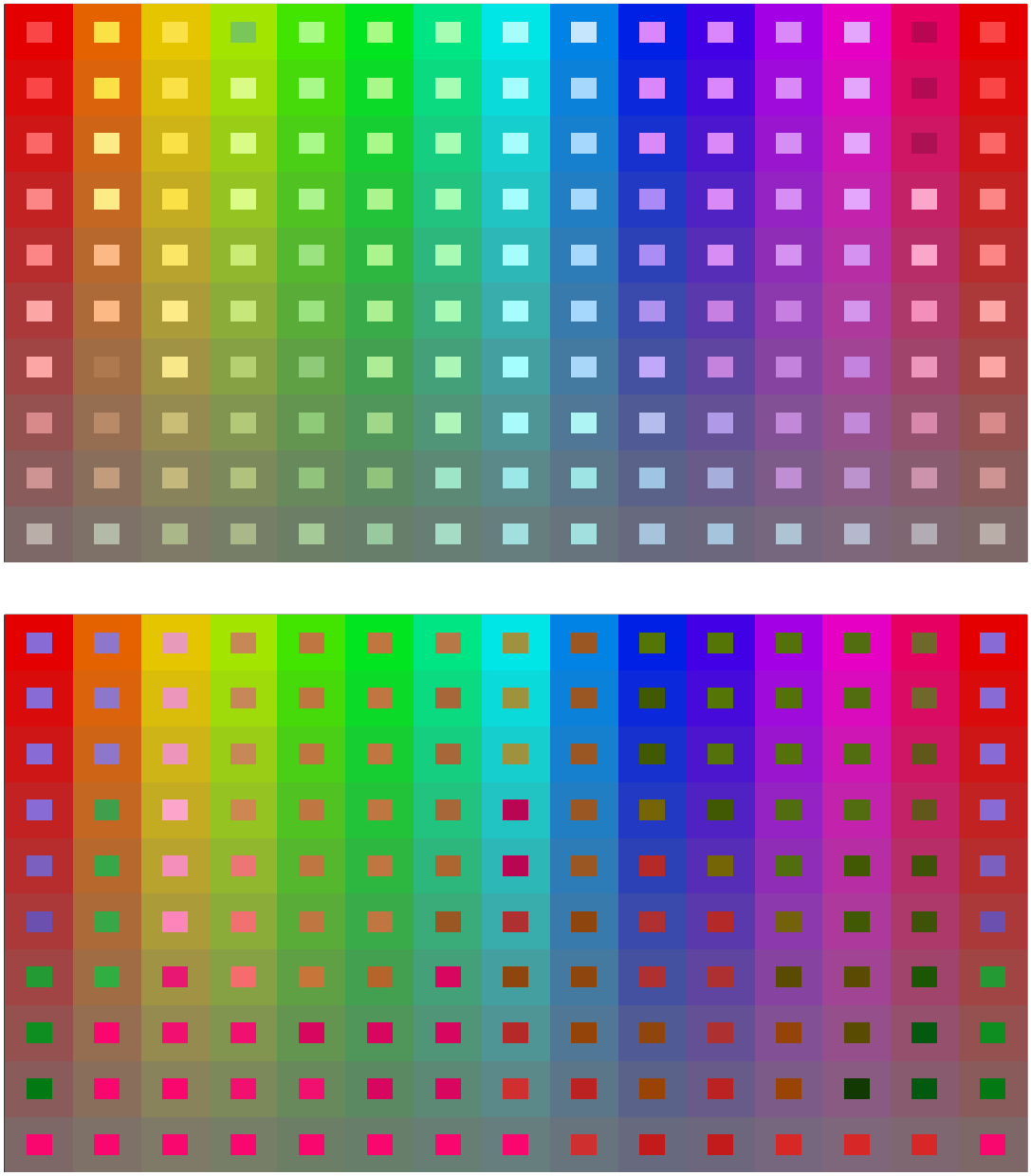


Figure 5.12: Predicted most and least *harmonious* color pairings for a fixed ground lightness level with varying hue and saturation levels.

Chapter 6

Effectiveness of different study formats

Retrieval practice study—study which involves both quizzing and reviewing—results in stronger and more durable memories than reviewing alone (H. Roediger & Karpicke, 2006a). However, incorporating quizzing into electronic tutoring systems is impractical for many common types of study materials; quiz answers that are visual, auditory, or procedural in nature cannot readily be entered into or assessed by a computer. A leading theoretical account of the mnemonic benefits of testing holds that the benefits are a result of memory traces being strengthened by the act of memory retrieval (Bjork, 1975). In this chapter, we investigate an important practical implication of this theory: after memory retrieval has occurred, it should not be necessary to physically enter the response into a computer to reap the benefits of retrieval practice.

Most studies of retrieval practice effects have required subjects to make overt responses during study, wherein subjects produce a response by writing, typing, or speaking (Smith, Roediger, & Karpicke, 2013). Some studies suggest that covert retrieval—where subjects mentally rehearse their response without physically producing it—is more beneficial than simply restudying (Izawa, 1976; Carpenter & Pashler, 2007; S. Kang, 2010; Putnam & Roediger, 2013). However, there are few studies that directly compare the effectiveness of overt and covert retrieval practice. Smith et al. (2013) compared overt and covert retrieval practice and found no difference in recall levels on a test shortly following study. In this chapter, we provide empirical evidence that a covert response modality can be more effective than the overt response modality on tests shortly after study, and that the apparent equivalence of the two is an artifact caused by controlling time per

trial. Though individual covert retrieval practice trials may be less effective than individual overt retrieval practice trials, they are faster and hence students can undergo substantially more trials in any fixed time window. This yields higher recall for short retention intervals and equivalent recall at longer retention intervals.

6.1 Experiment 1: Constant time per trial

Experiment 1 was a two-session experiment that used a between-subjects design to compare the efficacy of covert and overt retrieval practice study on foreign language vocabulary when time per trial is held constant. A within-subject condition varied the heuristic for determining which vocabulary item to present next to a subject during study. The first session of the experiment was divided into two blocks, one per scheduling heuristic. Within each block, students had an initial presentation of the material followed by 10 minutes of retrieval practice study. They then had a test on a random subset of the material after a 10 minute retention interval filled with a distracting task. The second session of the experiment occurred 48 hours later and tested students on all the remaining material.

6.1.1 Participants

48 undergraduates from the University of California, San Diego, Psychology Subject Pool participated for partial course credit.

6.1.2 Materials

The study material was 60 Swahili-English word pairs (Taken, Nelson, & Dunlosky, 1994). Students were cued with Swahili and trained to produce the corresponding English. For each subject, 30 pairs were randomly assigned to a *round robin* scheduling heuristic and 30 to a *best last* scheduling heuristic.

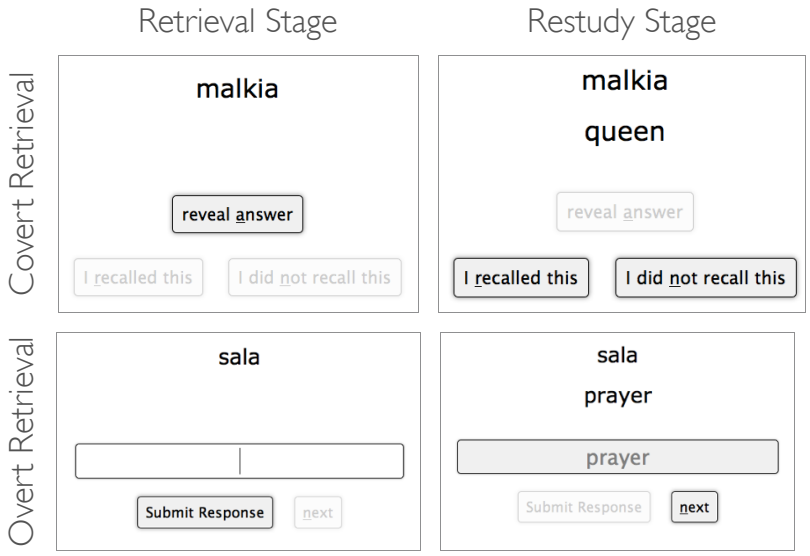


Figure 6.1: Screen captures of the self-paced covert and overt retrieval practice formats. Students were presented with the cue in the retrieval stage and responded either by typing in a response (overt retrieval condition) or by clicking a reveal button (covert retrieval condition). They then viewed the target in the restudy stage till they clicked a button or pushed the appropriate key.

6.1.3 Procedure

Day 1: The first session was divided into two consecutive blocks and each block was assigned a scheduling condition (round robin or best last) and given the condition's 30 vocabulary pairs. At the beginning of a study block, students underwent a study-only pass through the material in which each of the 30 vocabulary pairs was presented for 5 seconds, with a 250ms second blank screen between presentations. After all pairs had been presented, students began 10 minutes of retrieval practice study trials in the format dictated by their assigned condition. Regardless of the study format, each retrieval practice trial gave 6 seconds for retrieval and 4 seconds for review.

In the overt retrieval practice format, trials began with the presentation of the cue. Students had 6 seconds to type in a text box what they thought the target was. They were instructed to guess if they were unsure of the answer. After the 6 seconds had elapsed, regardless of the response's accuracy, the target was displayed along with the cue and the response. Following 4 seconds of review, the experiment proceeded immediately to the next trial.

In the covert retrieval practice format, trials also began with the presentation of a cue for 6

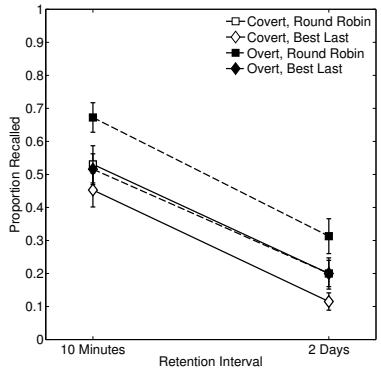
seconds. Students were instructed to recall the target from memory during this time. They were not required to type anything. After the 6 seconds had elapsed, students were presented with the cue and target and asked to click on either the button labeled *I recalled this* and *I did not recall this*. Students were asked to click the appropriate button (or push i or n on the keyboard). The buttons were always present on the screen, but were semi-transparent during the retrieval phase.

The within-subject item scheduling condition compared the effectiveness of two heuristics for determining which item should be selected for study at any given practice trial. The first, *round-robin* scheduling, had students just cycle through the items in a first-in-first-out order. Round-robin scheduling can be seen as taking advantage of the spacing effect by maximizing the amount of time between consecutive presentations of a vocabulary pair. The other scheduling condition, *best-last*, took advantage of the feedback available during training. The vocabulary pair it presented at any given practice trial was the pair that had been correctly recalled the least by the subject in all the preceding trials[1] . Ties were broken by choosing the pair with the highest normed difficulty (Taken et al., 1994). Thus, to elucidate, under best-last scheduling any pair that has been succesfully recalled n times will not be presented again for study till all the pairs that have been recalled $n-1$ times are presented and succesfully recalled. This scheduling method gives extra practice to vocabulary pairs a subject finds difficult.

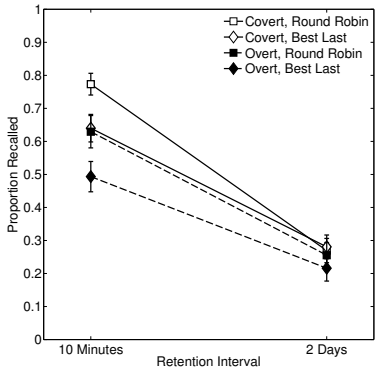
Following 10 minutes of retrieval practice study where item order was determined by the scheduling condition, students watched a television program for 10 minutes and then underwent a test on 15 randomly selected items from each scheduling condition. In each test trial, the Swahili cue was presented and a subject could take as much time as needed to type in what he or she thought was the English equivalent. No feedback was provided and trial order was randomized. Students were required to make a response in each trial.

Day 3: Students returned to the laboratory 48 hours ($\pm$15 minutes) after the first session to be tested on the 15 pairs from the round robin condition and the 15 pairs from the best-last condition that had not been tested on Day 1.

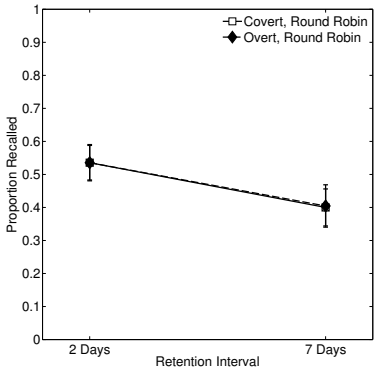
[1] With the constraint that the same pair could not be selected twice in a row.



(a) Experiment 1: 10 Minutes of Time-Locked Study



(b) Experiment 2: 10 Minutes of Self-Paced Study



(c) Experiment 3: 15 Minutes of Self-Paced Study

6.1.4 Results and discussion

Results from this experiment are shown in Figure 6.2a. In this experiment where time per trial is controlled, the overt response modality yielded significantly better performance than the covert response modality on both the initial test and delayed test, regardless of the scheduling condition.

6.2 Experiment 2: Self-paced trials

The design of experiment 2 is identical to Experiment 1 except that time per retrieval practice trial is not held constant (but total study time remains fixed at 10 minutes). Instead, students could advance to new trials at their own pace.

6.2.1 Subjects

Fifty undergraduates from the University of California, San Diego, Psychology Subject Pool participated in this experiment for course credit.

6.2.2 Procedure

Experiment 2's procedure is identical to Experiment 1 except that, rather than lasting a prescribed 10 seconds, students could advance through trials at their own pace. Pictures of the interface students used are shown in Figure 6.1.

In the overt retrieval practice format, the cue was presented and then the subject typed in what he or she thought the target was. The subject then clicked **Submit** or hit **Enter** on the keyboard. Next, regardless of the response's accuracy, the target was displayed along with the cue and the response, and a **Proceed** button appeared. The subject could hit the **Spacebar** or click **Proceed** to move on immediately to the next trial. Students were instructed to guess if they were unsure of the answer.

In the covert retrieval practice format, students were instructed to recall the target when presented with a cue. When a cue was presented to a subject, he or she clicked a **Reveal Answer** button after nominally retrieving the target from memory (or attempting to and failing). Then, the target was displayed alongside the cue, and two buttons appeared: a **Correct** buton and an **Incorrect** button. Students were instructed to click **Correct** (or push **c**) if they had been able to remember the target, and to click **Incorrect** (or push **i**) otherwise. The next trial began immediately after the subject responded.

6.2.3 Results

We conducted an ANOVA with three factors (study format, scheduler, and retention interval). Test performance is better at the shorter retention inteval (63.8% vs. 25.9%, $F(1, 32) = 272.3$, $p < .001$). Test performance is also better with the round-robin scheduler than with the best-last scheduler (49.1% vs. 40.6%, $F(1, 32) = 11.66$, $p = .002$), although scheduler interacted with retention interval ($F(1, 32) = 16.01$, $p < .001$), reflecting the fact that forgetting attenuated the differences between the conditions.

There is no main effect of study format ($F(1, 32) = 2.02$, $p = .164$). However, study format interacts with retention interval ($F(1, 32) = 7.75$, $p = .009$). Like the scheduler - retention interval interaction, this interaction is due to attenuated effects with forgetting: study format matters at the short RI (71.1% for covert, versus 56.5% for overt, $t(34) = 2.24$, $p = .031$) but not at the long RI (26.8% for covert, 25.0% overt, $t(34) = .33$).

When block order (RR-BL versus BL-RR) is included as a factor in the ANOVA, there is an

interaction of scheduler with block order ($F(1, 32) = 12.71$, $p = .001$) and a three-way interaction involving scheduler, order, and retention interval ($F(1, 32) = 8.93$, $p = .005$), simply indicating that participants did better in the second block than in the first.

6.3 Experiment 3: Self-paced trials, long retention intervals

The results of Experiment 2 suggest that the covert response condition, though superior for short retention intervals, may induce more rapid forgetting than the overt response condition. Thus, we hypothesized that for longer retention intervals, the covert response format may be inferior. This final experiment was designed to test that hypothesis.

We used a within-subject design to measure the retention of foreign language vocabulary two and seven days after it was studied via covert or overt retrieval practice. Day 1 of the experiment was divided into two randomly ordered study blocks: one block for covert study and one for overt study. Each study block began with an initial presentation of the material which was followed by 15 minutes of self-paced retrieval practice study trials. A randomly selected subset of vocabulary pairs from both conditions was tested in a cued free-response test two days later. The remaining pairs were tested one week after the initial session.

6.3.1 Participants

Students were drawn from our online research subject pool which consists of people of various ages and countries who have been screened for English proficiency, attentiveness to directions, and conscientious participation in prior studies. The experiment was conducted via the internet and was accessible to students through any standard web browser. Students who completed the experiment received $13 in payment via an Amazon.com gift certificate. We report data from students who completed all three sessions of the experiment ($n = 30$). The mean age of students who completed the experiment was 33.7 (SD = 11.9, range = 19 - 69). Eight students were male.

6.3.2 Procedure

Day 1: The first session was divided into two study blocks and each block was assigned a scheduling condition (round robin or best last) and 30 vocabulary pairs. Which scheduling condition came first was manipulated between students. At the beginning of a study block, students underwent a study-only pass through the material where each of the 30 pairs was presented for 8 seconds with a 1-second blank screen between presentations. After all pairs had been presented, students began 10 minutes of retrieval practice study trials in the format dictated by their condition. The timing of individual trials was determined by the subject.

Days 3 and 8: Students were reminded by email to complete the Day 3 and Day 8 test sessions. They were given a 14 hour window in which to participate (starting 7 hours before the appointed time and ending 7 hours after the appointed time). Thus, for example, students were allowed to start the second session between 41 and 55 hours after the start of their initial session on Day 1. Students who missed their time window were discontinued from the experiment and received no payment. Before registering for the experiment, students were shown a schedule and told that they would have to strictly adhere to it in order to receive compensation. For each subject, 15 vocabulary pairs from each condition were randomly selected for testing on Day 3. The pairs were presented in randomly ordered self-paced test trials. Day 8 followed the same procedure. The vocabulary pairs tested on Day 8 were all the vocabulary pairs that had not been tested on Day 3.

6.3.3 Results

Results from this experiment are shown in Figure 6.2c. We observed no significant differences in test performance on either exam.

6.4 Discussion

This line of research has practical implications for the design of tutoring systems. The results suggest that educators can use a covert response modality in tutoring systems—which is

more convenient—without harming efficacy. The covert condition was not inferior to the overt condition after longer retention intervals, and it was superior to the overt condition after short retention intervals.

Chapter 7

Long term recency is nothing more than ordinary forgetting

When tested on a list of items, individuals show a recency effect: the more recently a list item was presented, the more likely it is to be recalled. For short interpresentation intervals (IPIs) and retention intervals (RIs), this effect may be attributable to working memory. However, recency effects also occur over long timescales where IPIs and RIs stretch into the weeks and months. These **long-term recency** (LTR) effects have intrigued researchers because of their scale-invariant properties and the sense that understanding the mechanisms of LTR will provide insights into the fundamental nature of memory. An early explanation of LTR posited that it is a consequence of memory trace decay, but this *decay hypothesis* was discarded in part because LTR was not observed in continuous distractor recognition memory tasks (Glenberg & Kraus, 1981; Bjork & Whitten, 1974; Poltrock & MacLeod, 1977). Since then, a diverse collection of elaborate mechanistic accounts of LTR have been proposed. In this chapter, we revive the decay hypothesis. Based on the uncontroversial assumption that forgetting occurs according to a power-law function of time, we argue that not only is the decay hypothesis a sufficient qualitative explanation of LTR, but also that it yields excellent quantitative predictions of LTR strength as a function of list size, test type, IPI, and RI. Through fits to a simple model, this chapter aims to bring resolution to the subject of LTR by arguing that LTR is nothing more than ordinary forgetting.

7.1 Introduction

When subjects are studying a list of to-be-remembered items over a period of time, their recall accuracy at a subsequent test is greater for items at the end of the list than those in the middle. Studies of this phenomenon, the **recency effect**, date back to before the time of Ebbinghaus (Stigler, 1978), and in the past 135 years many experimental and theoretical papers have been published on the topic. Recency effects were initially attributed to residual information in working memory (R. Atkinson & Shiffrin, 1968). However, recency effects can occur when working memory is disrupted via a distractor task during the retention period (e.g., Nairne, Neath, Serra, & Byun, 1997). Surprisingly, recency effects also occur when list items are presented days or weeks apart (Baddeley & Hitch, 1977; Glenberg, Bradley, Kraus, & Renzaglia, 1983). For example, Glenberg et al. (1983) found a large recency effect for items spaced a full week apart, as shown in Figure 7.1a. They observed an astonishing 65% difference in the level of recall between items at the end of the list and items in the middle of the list.

Studies of such **long term** recency (LTR) effects (Baddeley & Hitch, 1977; Bjork & Whitten, 1974; Glenberg et al., 1980; Glenberg & Kraus, 1981; Glenberg et al., 1983; Greene, 1986; Nairne, 1991; Neath, 1993; Neath & Crowder, 1990, 1996) reveal a form of scale invariance. When recall is tested following a retention interval (**RI**) on the order of seconds, LTR will be observed if the time between items (the interpresentation interval or **IPI**) is on the order of seconds. When the RI is on the order of days, LTR will be observed only if the IPI is on that scale as well. This scale invariance leads one to wonder whether LTR might serve as a window into the operation of memory systems at many different timescales, and therefore might be a phenomenon whose mechanistic understanding will reveal deep insights into the nature of memory. Nonetheless, no consensus on the nature of the phenomenon has been reached.

This chapter argues that an obvious and parsimonious — but long discarded — account of LTR effects is fully consistent with the literature. This hypothesis, the **decay hypothesis**, posits that recency effects are due to the decay of memory trace (Glenberg et al., 1983). Simply put,

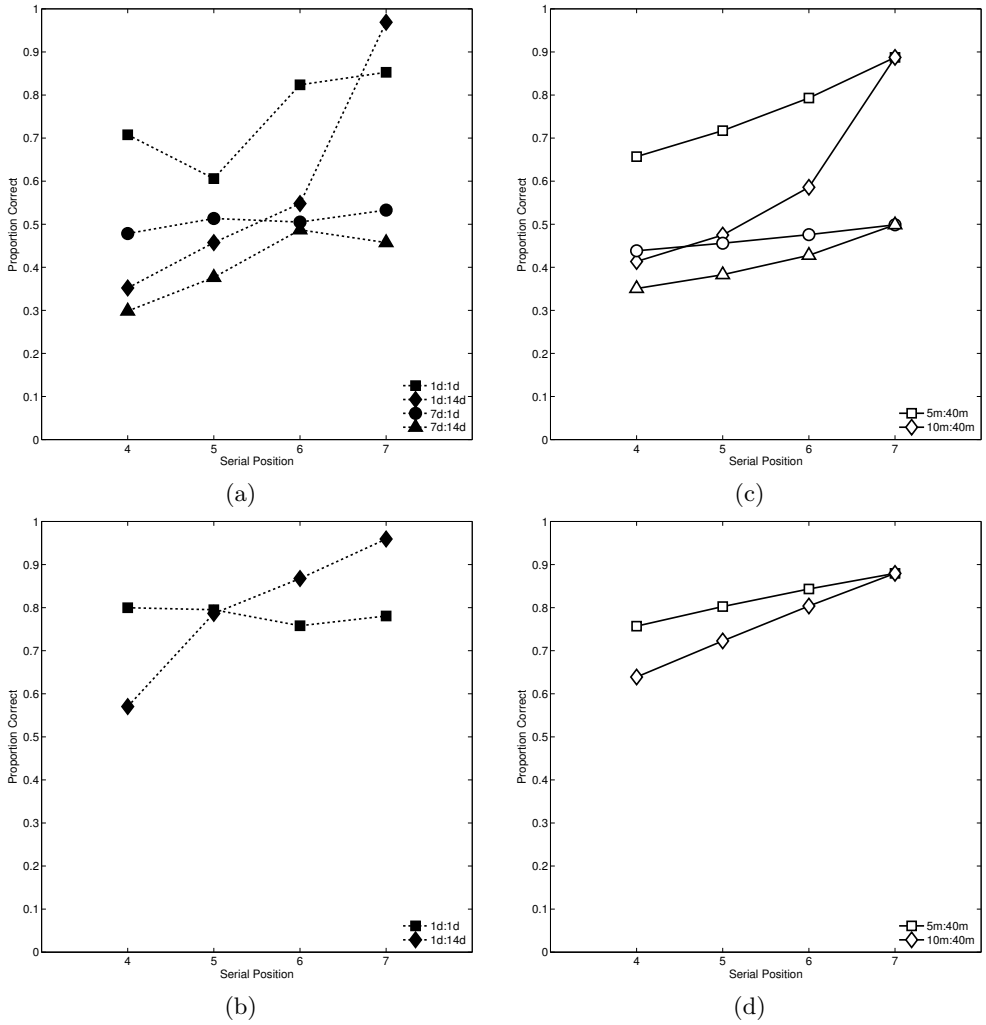


Figure 7.1: Glenberg et al. (1983) Experiment 5 (a) and 6 (b) empirical data, and Experiment 5 (c) and 6 (d) simulation. Here and throughout the chapter, we have excluded the first few serial positions because they evidence primacy, which is a separate phenomenon from recency and is not our focus.

people gradually forget things. Typically, when little time has elapsed since study, relatively little forgetting will have occurred; thus, items studied toward the end of a list are most easily recalled because they were studied most recently. Hence, long term recency (LTR) is a direct consequence of ordinary forgetting. This explanation was abandoned in favor of alternative theories because, in several key early studies, long-term recency effects were not observed in recognition tasks (in particular, see Glenberg & Kraus, 1981; Bjork & Whitten, 1974; Poltrock & MacLeod, 1977). This finding appeared to be decisive evidence against the decay hypothesis: if the decay of memory

strength is solely responsible for LTR, the manner in which subjects respond should be irrelevant and there should be LTR in recognition tasks just as there is in free recall tasks.

We show that the phenomenon of LTR, including the lack of statistically significant LTR effects in recognition memory, can be fully predicted by what is independently believed about forgetting: that recall probability follows a power-law function of time.

7.2 Formalization of the decay hypothesis

Our model rests on the relatively uncontroversial assumption that recall probability of an item following a single study presentation decays according to a power-law function (J. Anderson & Schooler, 1991; Wickelgren, 1974; Wixted & Carpenter, 2007; Wixted & Ebbesen, 1991). The recall probability following an elapsed time t since study, $p(t)$, is defined as $p(t) = (1 + \alpha t)^{-\beta}$, where α is a time-scaling parameter ($\alpha > 0$) and β is the decay rate ($\beta > 0$).[1] This equation is an instance of the Wickelgren power-law forgetting curve $\gamma(1 + \alpha t)^{-\beta}$ (Wickelgren, 1974), where γ represents initial recall probability or the effectiveness of study. Without loss of explanatory power, in this chapter we assume that initial encoding is certain ($\gamma = 1$).

In free recall, subjects determine the order of report. Consequently, the **effective** retention interval of an item depends not only on its serial position in the initial list, but also on its recall output position. If forgetting follows a power law, slight variability in retention interval should not matter for material held in memory for hours or days, but due to the steepness of the forgetting curve shortly after study, variability in retention interval can have noticeable effects on recall accuracy for material held in memory just seconds or minutes. Studies of long-term recency do not necessarily involve long retention intervals; for instance, in Nairne et al. (1997), IPIs and RIs were as short as one second and responses extended over a twelve-second recall window.

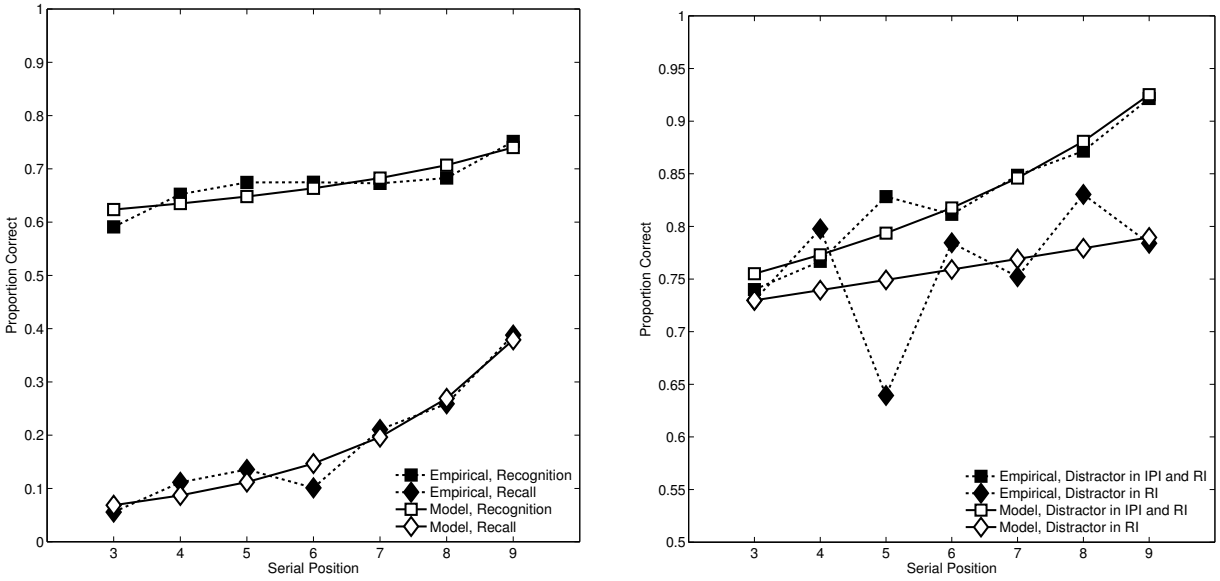
Because small variability in the RI can have a large effect on recall probability in such a situation, we found it necessary to make an additional assumption about free report in order to

[1] This account is noncommittal as to whether t refers to the mere passage of time, to a measure of the number of intervening events, or to a combination thereof. LTR appears to be due to both passage of time and interference (da Costa Pinto & Baddeley, 1991).

determine the effective retention interval. We assume that items presented late in a list are more likely to be recalled first because their memory traces are strongest among the list items and they will out-compete older items in the list. Empirical support for this assumption comes from Nilsson, Wright, and Murdock (1975), who found that in a free recall test given immediately after a sequence of visually presented stimuli, later items in the list tend to be recalled before earlier items. (While this assumption may be incorrect for the initial items in a list due to primacy, those items are irrelevant for the purpose of determining LTR effects.) The consequence of this assumption is that the last items will have shorter effective RIs than earlier items, and increasing the effective retention interval of the earlier items by a measurable percentage will amplify recency effects. This amplification is noticeable only when RIs and IPIs are brief.

We characterize recall output order in terms of a probabilistic generative process having the property that if items in serial positions i and j are both reported, item i will be reported after j if and only if $i < j$. In the generative process, the time at which a memory retrieval attempt for serial position i occurs depends on which later items j, i.e., $j > i$, were correctly recalled. Let $R_i \in \{0, 1\}$ denote whether the ith serial position is recalled during the test and let T_i be the time at which the memory retrieval was attempted. We assume that $T_{i-1} = T_i + R_i \, \mathcal{L}(p(T_i))$, where $\mathcal{L}$ is the response latency (described in the next paragraph) and p is the power-law function already described. Whether or not an item is recalled is determined by a biased coin flip: $R_i \sim \text{Bernoulli}(p(T_i))$. Recall does not necessarily begin at the last serial position or proceed consecutively, but it does always proceed from high to low serial positions. The model's recall probability for serial position i is the expectation $\mathbb{E}[R_i]$.

To estimate response latencies, we leverage ACT-R (J. Anderson et al., 2004), which is perhaps the best accepted model of long-term memory. Based on ACT-R's declarative memory module, we adopt the assumption that when successful recall of an item occurs, the time to recall it depends on its memory strength. In ACT-R, this strength also determines recall probability. Response latency $\mathcal{L}$ in ACT-R can, under simplifying assumptions, be solved algebraically in terms of recall probability and written as $\mathcal{L}(p) = \psi \frac{1-p}{p}$, where ψ scales how much response latency



(a) (b)

Figure 7.2: Serial position curves and model fits for (a) Glenberg & Kraus (1981) and (b) Talmi & Goshen-Gottstein (2006). Because of the design of both experiments, the model fits shown are simply the two-parameter power law forgetting curve; no adjustments for response times or recall order were made.

increases with decreasing odds of successful recall.[2]

In summary, our model embodies well-accepted characteristics of memory recall and includes a simple generative process to describe free recall. It has three parameters: α, β, and ψ. The parameters are constrained post hoc to describe the material, population, and testing procedure of a study.

7.3 Empirical phenomena associated with LTR

Recency is evident in serial position curves by a characteristic upward bend for the final serial positions (e.g., Figure 7.2a). The **strength** of LTR can be characterized by the steepness of the upward bend, which Glenberg et al. (1983) and subsequent authors quantified in terms of the slope

[2] Recall probability p in ACT-R as given in terms of memory strength m and free parameters τ and θ is $p(m) = (1+\exp(\frac{\tau-m}{\theta}))^{-1}$. Response latency $\mathcal{L}$ is given in terms of m and free parameters ω and ϕ by $\mathcal{L}(m) = \omega\exp(-m)+\phi$. Assuming ϕ, a fixed time cost associated with perceptual motor encoding, is negligible, $\mathcal{L}(p) \approx \psi(\frac{1-p}{p})$ where $\psi \equiv \theta\omega e^{-\tau}$.

of the line fit by least-squares to the last three serial positions.[3] In this section, we present evidence that the decay hypothesis explains the key phenomena associated with LTR and obtains excellent quantitative fits to various experimental outcomes as demonstrated by serial position curves and the associated LTR strength.

7.3.1 Absence of LTR in recognition tasks

A study by Glenberg and Kraus (1981), titled *Long-term recency is not found on a recognition test*, contributed to the abandonment of the decay hypothesis. LTR was assessed in two testing formats: free recall and recognition. The dotted lines in Figure 7.2a represent the serial position curves for recognition (squares) and recall (diamonds). Glenberg and Kraus performed several analyses, including an ANOVA testing for a main effect of serial position across the final 3 positions of each curve. Finding a reliable effect in recall but not recognition performance, the authors rejected the decay hypothesis. Their reasoning was that if LTR was a consequence of memory trace decay over time, testing format should not matter. Because testing format matters, the decay hypothesis seemed implausible. In other early studies, LTR was not detected in recognition either (Bjork & Whitten, 1974; Poltrock & MacLeod, 1977).

In our model, the power-law forgetting curves do not directly represent the strength of memory; rather, they indicate memory strength **as reflected in a particular read-out task**. The same memory state may yield poor performance in a challenging task like free recall, where veridical recall requires reconstruction of the specific items studied, but good performance in an easy task like recognition, where the memory trace must merely be strong enough to support a reliable old vs. new discrimination. Thus, distinct forgetting curves are warranted for recall and recognition.

The solid lines in Figure 7.2a show independent least-squares fits of the two-parameter forgetting curve $p(t)$ to the two serial position curves. The forgetting curves, reflecting proportion correct as a function of time, are obtained by flipping the solid lines from right to left. (For the recognition

[3] With straightforward algebra, the slope of the least-squares fit can be shown to be half the difference between the score at the last and third-to-last serial positions, $\frac{1}{2}(\mathbb{E}[R_n] - \mathbb{E}[R_{n-2}])$.

condition, the ψ parameter—determining free recall order—was not used, because testing was cued and randomized.) The model's forgetting curves are good matches to the serial position curves. Forgetting, as reflected in the drop in performance from serial position 9 to position 3, is shallower for recognition. Consequently, if the model predictions are correct, the experiment may not have had sufficient power to detect a difference in recognition accuracy across serial positions.

With the model's forgetting curve in the recognition condition, we can perform a power analysis to determine how likely an LTR effect is to be detected by Glenberg and Kraus (1981) at the 95% significance level. The experiment included 54 subjects, and each was tested on 3 lists in each testing condition. Assuming (a) model estimates of recall and recognition probability are accurate for serial positions 7-9, (b) probability is the same across subjects and lists tested, and (c) items within a list are independent of one another, we used the model to simulate experimental outcomes and tested for a main effect of serial position. Although according to the model there is a true LTR effect for both recognition and recall, the simulated experiment had only a 14% chance of detecting the effect in recognition, but a 98% chance in recall. To meet the convention of 80% statistical power (J. Cohen, 1992), Glenberg and Kraus would have needed to run approximately 400 subjects.

Talmi and Goshen-Gottstein (2006) critiqued the multi-probe testing procedure used in earlier recognition experiments, noting that the procedure likely attenuated or eliminated LTR. Instead, they probed only one serial position per trial in recognition testing. Their study included two presentation conditions: in one condition, subjects performed a distracting task during the IPI and RI; in the other, subjects performed a distracting task only during the RI. (We omit a third condition in the experiment because it did not test LTR.) The serial position curves obtained in the study, along with the model fits, are shown in Figure 7.2b.[4] Talmi and Goshen-Gottstein (2006) reported reliable LTR in the condition with a distractor in the IPI and RI but not in the condition with a distractor only in the RI. These findings are consistent with a statistical power analysis we

[4] Because Talmi and Goshen-Gottstein tested only one serial position per list, there was no uncertainty in the effective retention interval, and we again fit a two-parameter model which did not make use of the read-out order assumptions or parameter ψ.

performed based on the model, which reveals a 93% chance of observing an extant LTR effect in the former condition, but only an 8% chance of observing extant LTR in the latter condition.

In summary, the failure to detect LTR in some recognition studies does not disconfirm the decay hypothesis because those studies lacked the statistical power necessary to reach this conclusion: forgetting rates in recognition are slow, and consequently differences across serial positions are so small that experimental noise can mask them. Previous studies had no reasonable expectation of observing an extant LTR effect given their inadequate power. Although the power could be increased by running more subjects, Talmi and Goshen-Gottstein (2006) employed experimental manipulations that helped increase the power by increasing the magnitude of forgetting (though they did so for reasons unrelated to power).

7.3.2 Effect of list length

Increasing the length of the list of to-be-remembered items has little effect on the recall accuracy of the last few serial positions but lowers recall accuracy for earlier serial positions (Greene, 1986; Murdock, 1962). For example, Greene (1986) performed an LTR study in which list length was manipulated within-subject so that the lists were either 6 or 10 items long. With IPIs and RIs of 20s filled with a distracting task, list length did not affect recall accuracy for any serial position relative to the end of the list (dotted lines in Figure 7.3).

Our simple model makes the strong assumption that each list item decays independently. Because there are no interactions among items, the number of items preceding a serial position is irrelevant and consequently the model predicts that the recall accuracies of the final serial positions are unaffected by an increase in list length. The model predicts that recall accuracies for early serial positions are lowered because these items' effective RIs increase when list length is increased. A model fit to the data is shown as the solid lines in Figure 7.3 and describes the empirical data well.

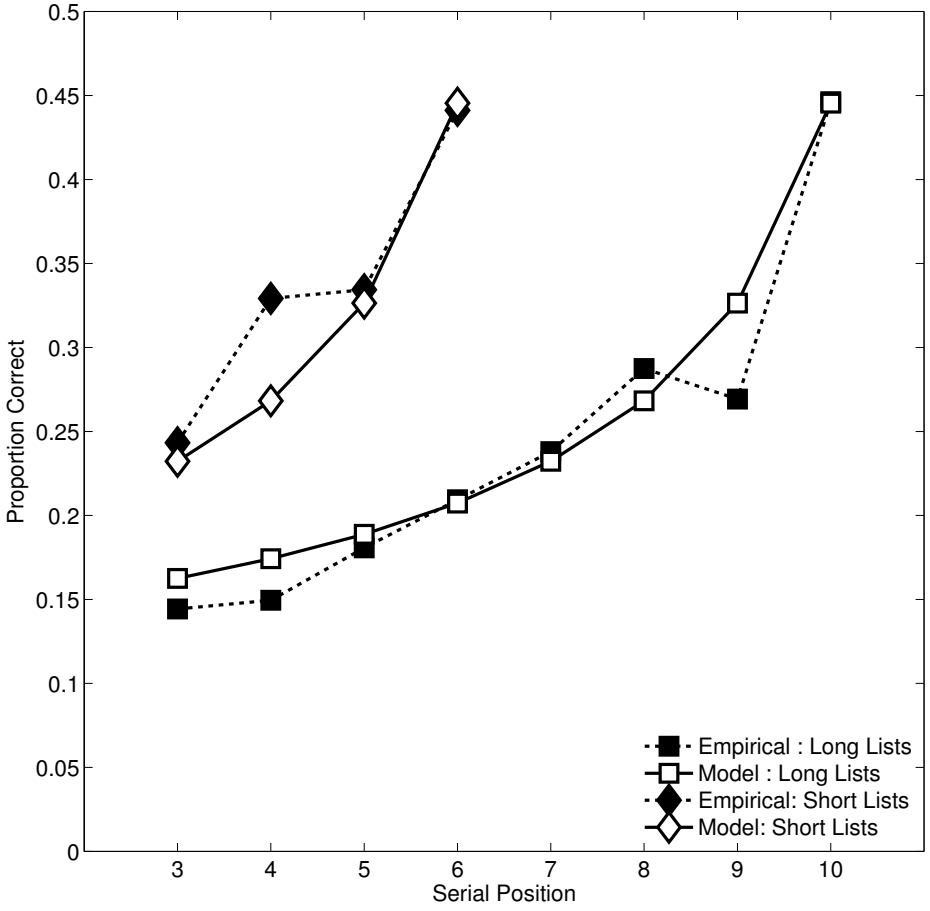


Figure 7.3: Serial position curves from Greene (1986) and a single model parameterization obtained by a least-squares fit to both serial position curves. The strength of LTR, the steepness of the upward bend in the curves on the last few serial positions, is invariant to list size. For early serial positions, recall accuracy is decreased by an increase in list length.

7.3.3 Ratio rule

Various authors have noted what appears to be a form of scale invariance of LTR wherein the strength of LTR depends only on the ratio of IPI to RI (Baddeley & Hitch, 1977; Bjork & Whitten, 1974; Glenberg et al., 1983, 1980; Nairne et al., 1997). Further, as the IPI:RI ratio increases, so does the strength of LTR. Thus, LTR is stronger if the IPI is increased for a fixed RI or if the RI is decreased for a fixed IPI. This dependence of LTR solely on the IPI:RI ratio has been dubbed the **ratio rule**.

Glenberg et al. (1983) conducted a series of experiments exploring the ratio rule, two of

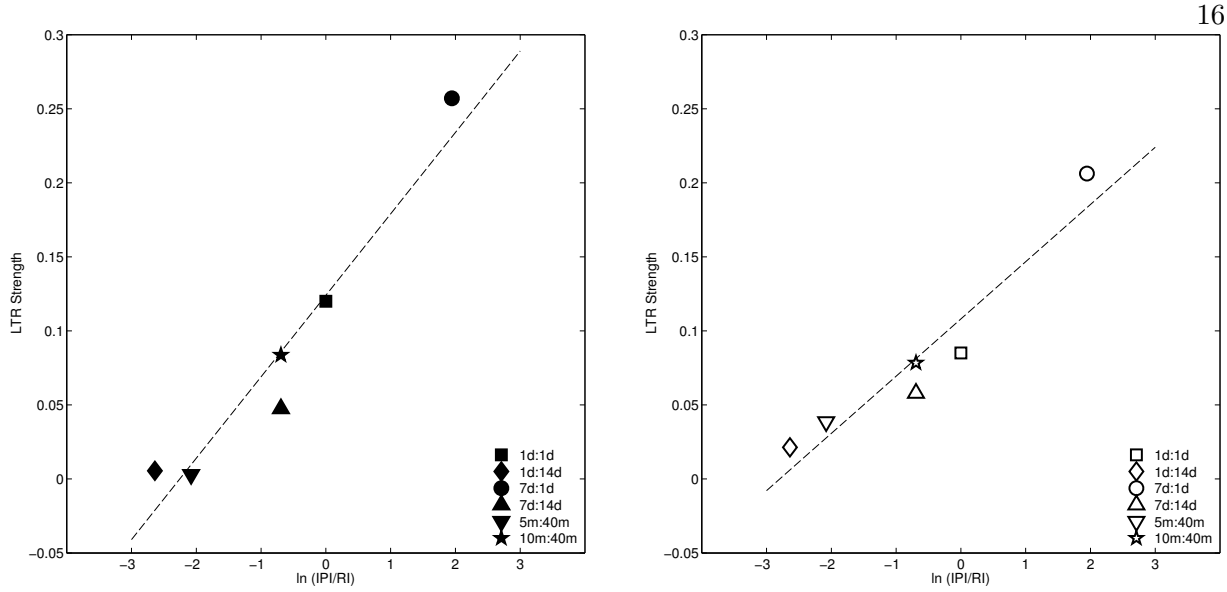


Figure 7.4: (a) Empirical and (b) simulated LTR strength for Glenberg et al. (1983). The simulation used the model fits shown in Figures 7.1c and 7.1d.

which examined scale invariance of the ratio rule by varying the IPI and RI over several orders of magnitude in a free recall task: In Experiment 5, each subject participated in 7 study sessions separated by an IPI of 1 or 7 days and was then tested following an RI of 1 or 14 days. In Experiment 6, IPIs were 5 or 20 minutes, the RI was 40 minutes, and the IPI and RI were filled with a distracting task (television) to prevent rehearsal. The serial position curves reported from these two experiments are shown in Figures 7.1a,b. Figure 7.4a shows the LTR strength, the slope measure defined earlier, across a variety of IPIs and RIs combined from the two experiments. The abscissa expresses the IPI:RI ratio on a logarithmic scale. The dashed regression line suggests a log-linear trend: the LTR strength is proportional to the logarithm of the IPI:RI ratio. The figure also offers some direct support for the ratio rule via two points, the star and upward-facing triangle, with the same IPI:RI ratio having roughly the same LTR strength.

Figures 7.1c,d show least-squares fits of the model to the empirical serial position curves (Figures 7.1a,b). Figure 7.4b shows the fitted model's prediction of the empirical LTR strengths (Figure 7.4a). The model's predicted LTR strength shows a close qualitative correspondence to the

empirical LTR strengths and provides further support for the decay hypothesis.

Nairne et al. (1997) includes a single session LTR experiment that, like Glenberg et al. (1983), explores the scale invariance of the ratio rule over a wide range of IPIs and RIs. Subjects were presented with 6-item lists of letters, and the test session's format was free recall. During the IPI and RI, subjects were presented with a randomly selected digit every 500ms to disrupt short term memory. The serial position curves reported from this experiment are shown in Figures 7.5a,b. (They are divided into two figures for visual clarity.) The dotted line in Figure 7.6 shows LTR strength as a function of the log IPI:RI ratio. As with Glenberg et al. (1983), the observed LTR strength exhibits a log-linear trend and is supportive of the ratio rule.

Figures 7.5c,d show a single least-squares fit of the model to all of the empirical serial position curves (Figures 7.5a,b). The solid line in Figure 7.6 shows the fitted model's prediction of the empirical LTR strengths. The model's predicted LTR strength shows a close quantitative correspondence to the empirical LTR strength.

7.3.4 Systematic deviations from the ratio rule

Nairne et al. (1997) conducted an experiment in which they kept the IPI:RI ratio constant while varying the IPI and RI. They used a multiple choice test format in which subjects were presented with 16 letters and were asked to click on the six that appeared in the list. Figure 7.7a shows serial position curves from this experiment. The dotted line shows the variation in LTR strength as a functon of the IPI and RI. If the ratio rule is strictly correct, then LTR strength should be constant along the abscissa. In actuality, LTR systematically decreased as the duration of the IPI and RI increased. Thus, the ratio rule does not always hold: as the timescale of an experiment increases, LTR effects decrease.

Figure 7.7b shows a least-squares fit of the model to the serial position curves (Figure 7.7a). The solid line in Figure 7.8 shows the fitted model's LTR strength as a function of the IPI and RI. It demonstrates that the decay hypothesis can account for the observed deviations from the ratio rule. When the IPI increases, the effective RI of individual items also increases, which shifts items

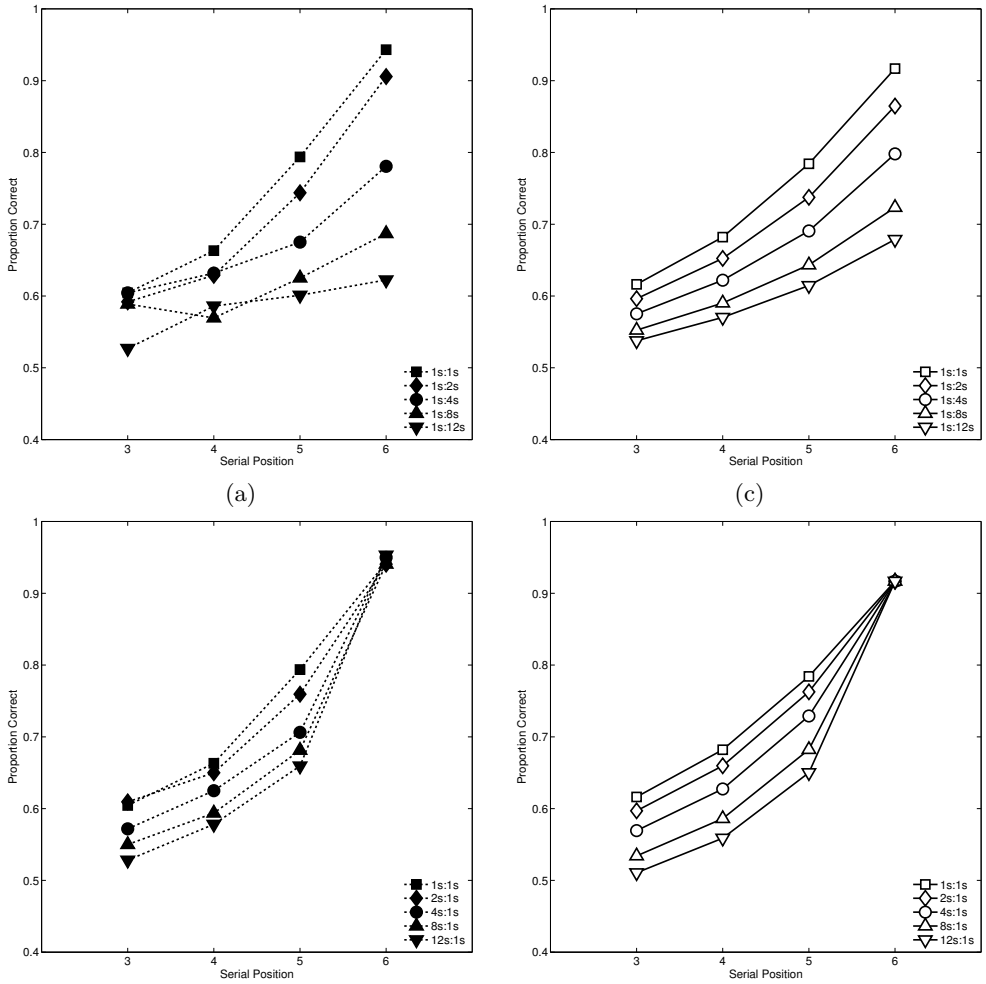


Figure 7.5: (a,b) Serial position curves for Nairne et al. (1997) Experiment 1 and (c,d) the model fit, a model parameterization obtained by least-squares.

toward the relatively flat portion of the power-law forgetting curve. The plateauing of forgetting as the timescale increases reduces LTR strength by reducing the differences in recall probabilities among different serial positions.

7.4 Conclusion

LTR and its associated phenomena have long appeared enigmatic. Why does LTR have an apparent scale invariance? Why is it more readily observed in free recall than in recognition? Why is it invariant to list length? Why does LTR strength have a systematic relationship with the IPI:RI

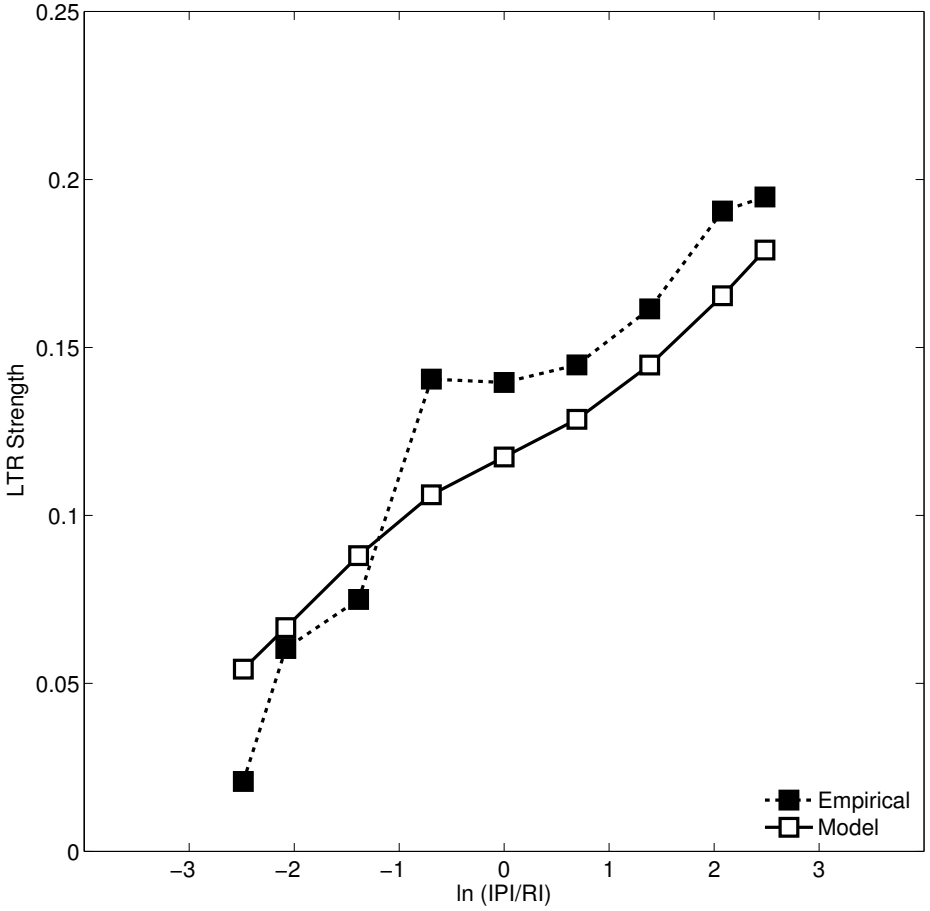


Figure 7.6: Empirical and simulated LTR strength for Nairne et al. (1997) Experiment 1. The simulation used the fit shown in Figures 7.5c,d.

ratio, yet sometimes it changes even when the ratio is held constant? Our simple model, based on the notion that LTR is nothing more than ordinary forgetting, answers all of these questions and provides quantitative fits to experimental data. Though separate qualitative arguments about how the decay hypothesis accounts for each of these could be made, the single quantitative account embodied in our model represents the most rigorous and unified treatment of the decay hypothesis to date. On grounds of parsimony, an explanation of LTR distinct from ordinary forgetting does not seem to be warranted.

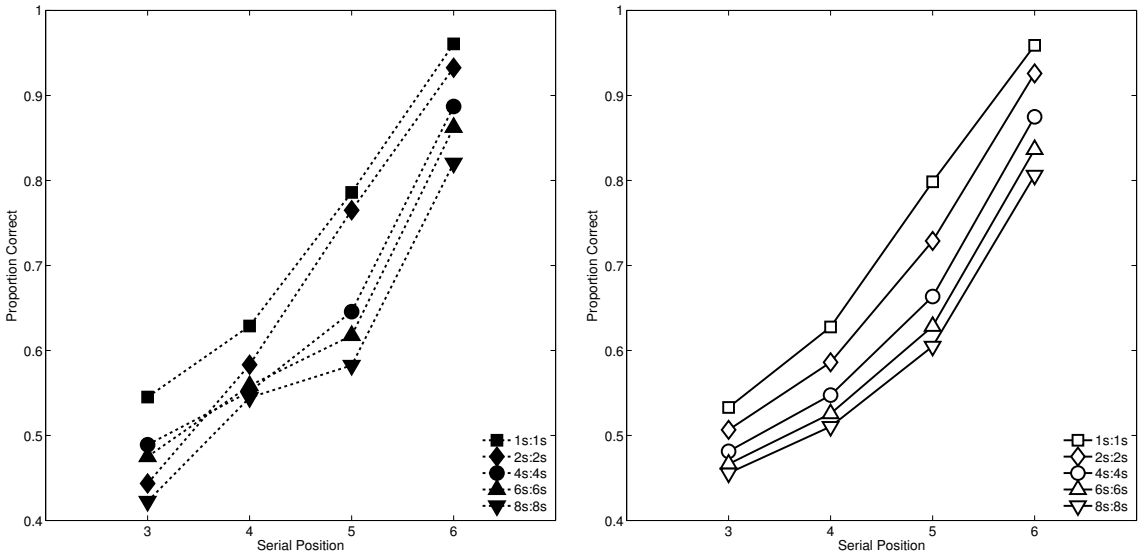


Figure 7.7: (a) Serial position curves for Nairne et al. (1997) Experiment 3 and (b) the least-squares model fit.

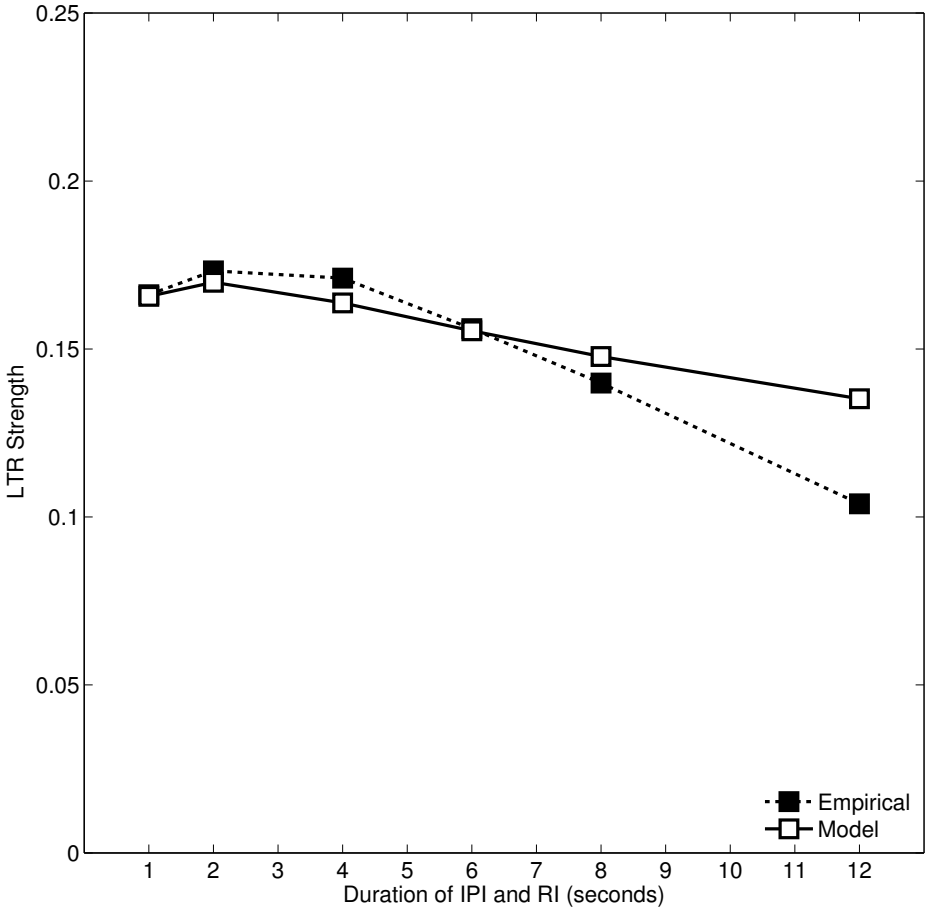


Figure 7.8: Empirical and simulated LTR strength for Nairne et al. (1997) Experiment 3. The simulation used the model fit shown in Figure 7.7b.

Chapter 8

Major Contributions

The primary contribution of this work is twofold. On the theoretical side, we developed novel statistical models of learning and forgetting which combine Bayesian methods for inferring individual differences with a psychological theory of memory, and we used a model of forgetting to provide a parsimonious theoretical account of long-term recency effects. On the practical side, we created and evaluated model-based approaches to optimizing human learning for both individual students and populations of students.

Although scientists have been researching human learning and forgetting since the nineteenth century (Stigler, 1978), surprisingly little of the research has translated into improved educational practices (e.g., Dempster, 1988). We suggest that the lack of translation is due to the qualitative nature of the advice that cognitive psychologists have traditionally been able to give educators. It may not be sufficient, for example, for educators to be told that temporally spaced study is generally better than temporally massed study. Generic advice is unlikely to be well tailored to any individual student because different students have different needs. Furthermore, moving abstract advice into concrete, classroom practice is challenging—choosing too much or too little spacing is bad, but educators have no way to systematically select the best middle-ground spacing via rule-of-thumb heuristics. This thesis is premised on the idea that educators often need quantitative, prescriptive guidance about what instructional strategies they should employ, including specific guidance regarding how they should distribute their students' study for their material and regarding what exactly an individual student should study next. The statistical approaches developed in this thesis

for delivering optimized instruction take an important step in this direction: they treat education as a probabilistic modeling and control problem, one constrained by known characteristics of human learning and forgetting, and in this way they can provide educators the kind of quantitative guidance they need. The series of longitudinal experiments we presented involving middle school students demonstrates how incorporating such systematic instruction into classrooms can yield large improvements in the retention of course material over educationally relevant timescales.

References

Agarwal, D., & Merugu, S. (2007). Predictive discrete latent factor models for large scale dyadic data. In Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining.

Anderson, E. J., & Ferris, M. C. (2001). A direct search algorithm for optimization with noisy function evaluations. SIAM Journal of Optimization, 11, 837–857.

Anderson, J. (1976). Language, memory, and thought. Hillsdale, NJ: Erlbaum.

Anderson, J., Bothell, D., Byrne, M., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. Psychological Review, 111, 1036–1060.

Anderson, J., & Milson, R. (1989). Human memory: An adaptive perspective. Psychological Review, 96, 703–719.

Anderson, J., & Schooler, L. (1991). Reflections of the environment in memory. Psychological Science, 2(6).

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

Anderson, R., & Tweney, R. (1997). Artifactual power curves in forgetting. Memory & Cognition, 25(5), 724–730.

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

Atkinson, R., & Shiffrin, R. (1968). Human memory: a proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), The psychology of learning and motivation: Advances in research and theory. New York: Academic Press.

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

Baddeley, A., & Hitch, G. (1977). Attention and performance VI. London: Academic Press.

Baker, R., Corbett, A., & Aleven, V. (2008). More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing. In Intelligent Tutoring Systems (pp. 406–415).

Balota, D., Duchek, J., Sergent-Marshall, S., & Roediger, H. (2006). Does expanded retrieval produce benefits over equal interval spacing? Explorations of spacing effects in healthy aging and early stage alzheimer's disease. Psychology & Aging, 21, 19–31.

Beck, J. (2007). Difficulties in inferring student knowledge from observations (and why you should care). In Educational data mining: Supplementary proceedings of the 13th international conference on artificial intelligence in education (pp. 21–30).

Beck, J., & Chang, K. (2007). Identifiability: A fundamental problem of student modeling. In Proceedings of the 11th international conference on user modeling (pp. 137–146). Berlin, Heidelberg: Springer-Verlag.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In L. Bottou & M. Littman (Eds.), Proceedings of the 26th international conference on machine learning (pp. 41–48). Montreal: Omnipress.

Bjork, R. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), Information processing and cognition: The Loyola Symposium. Hillsdale, NJ: Erlbaum.

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

Bjork, R., & Bjork, E. (1992). From learning processes to cognitive processes: Essays in honor of William K. Estes. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.

Bjork, R., & Whitten, W. (1974). Recency-sensitive retrieval process in long-term free recall. Cognitive Psychology, 6, 173 - 189.

Brown, G., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. Psychological Review, 114, 539 - 576.

Budé, L., Imbos, T., van de Wiel, M., & Berger, M. (2011). The effect of distributed practice on students' conceptual understanding of statistics. Higher Education, 62, 69–79.

Burrell, Q. (1980). A simple stochastic model for library loans. Journal of Documentation, 36, 115–132.

Burrell, Q., & Cane, V. (1982). The analysis of library data. Journal of the Royal Statistical Society, 145, 439–471.

Camp, C., Bird, M., & Cherry, K. (2000). Cognitive rehabilitation in old age. In R. Hill, L. Backman, & A. N. Stigsdotter (Eds.), (pp. 224–248). Oxford, UK: Oxford University Press.

Carpenter, S., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. Psychonomic Bulletin & Review, 14, 474–478.

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

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. Memory & Cognition, 20, 633–642.

Carvalho, P. F., & Goldstone, R. L. (2011). Stimulus similarity relations modulate benefits for blocking versus interleaving during category learning. (Presentation at the 52nd Annual Meeting of the Psychonomics Society, Seattle, WA)

Cen, H., Koedinger, K., & Junker, B. (2007). Is over practice necessary? Improving learning efficiency with the cognitive tutor through educational data mining. Frontiers in artificial intelligence and applications, 158.

Cen, H., Koedinger, K., & Junker, B. (2008). Comparing two IRT models for conjunctive skills. In B. W. et al. (Ed.), Proceedings of the 9th international conference on intelligent tutoring systems.

Cen, H., Koedinger, K. R., & Junker, B. (2006). Learning factors analysis - a general method for cognitive model evaluation and improvement. In M. Ikeda, K. D. Ashley, & T. W. Chan (Eds.), Intelligent tutoring systems (pp. 164–175). Springer.

Cepeda, N., Coburn, N., Rohrer, D., Wixted, J., Mozer, M., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. Experimental Psychology.

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

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

Cetintas, S., Si, L., Xin, Y., & Hord, C. (2010). Predicting correctness of problem solving in ITS with a temporal collaborative filtering approach. Proceedings of the 10th international conference on Intelligent Tutoring Systems, 15–24.

Chang, K., Beck, J., Mostow, J., & Corbett, A. (2006). A Bayes net toolkit for student modeling in intelligent tutoring systems. Proceedings of the 8th International Conference on Intelligent Tutoring Systems.

Chi, M., Koedinger, K., Gordon, G., Jordan, P., & van Lehn, K. (2011). Instructional factors analysis: A cognitive model for multiple instructional interventions. In C. Conati & S. Ventura (Eds.), Proceedings of the 4th international conference on educational data mining (pp. 61–70).

Chi, M., van Lehn, K., Litman, D., & Jordan, P. (2011). Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. User Modeling and User-Adapted Interaction. Special Issue on Data Mining for Personalized Educational Systems, 21, 137–180.

Christian, B. (2012). The A/B test: Inside the technology that's changing the rules of business. Wired, 20(4).

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155 - 159.

Cohen, M. S., Yan, V. X., Halamish, V., & Bjork, R. A. (2013). Do students think that difficult or valuable materials should be restudied sooner rather than later? Journal of Experimental Psychology: Learning, Memory, and Cognition, 39(6), 1682–1696.

Conati, C., Gertner, A., & van Lehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. Journal of User Modeling and User-Adapted Interaction, 12, 371–417.

Conati, C., Gertner, A. S., van Lehn, K., & Druzdzel, M. J. (1997). On-line student modeling for coached problem solving using Bayesian networks. In Proceedings of the sixth international conference on user modeling (pp. 231–242). Springer.

Conati, C., & Muldner, K. (2007). Evaluating a decision-theoretic approach to tailored example selection. Proceedings of the 20th International Joint Conference on Artificial Intelligence.

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

Corbett, A., & Bhatnagar, A. (1997). Student modeling in the ACT programming tutor: Adjusting a procedural learning model with declarative knowledge. User Modeling: Proceedings of the Sixth International Conference.

Cull, W. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. Applied Cognitive Psychology, 14, 215–235.

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.

da Costa Pinto, A., & Baddeley, A. (1991, Sep). Where did you park your car? analysis of a naturalistic long-term recency effect. European Journal of Cognitive Psychology, 3(3), 297 - 313.

Davelaar, E., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H., & Usher, M. (2005). The demise of short-term memory revisited: Empirical and computational investigations of recency effects. Psychological Review, 112(1), 3 - 42.

De Jonge, M., Tabbers, H. K., Pecher, D., & Zeelenberg, R. (2012). The effect of study time distribution on learning and retention: A goldilocks principle for presentation rate. Journal of Experimental Psychology: Learning, Memory, and Cognition, 38, 405–412.

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

Dempster, F. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. American Psychologist, 43(8), 627–634.

Dempster, F. (1991). Synthesis of research on reviews and tests. Educational Leadership, 48, 71–76.

Desmarais, M., & Baker, R. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. User Modeling and User-Adapted Interaction, 22, 9–38.

Draney, K., Pirolli, P., & Wilson, M. (1995). Cognitively diagnostic assessment (P. Nichols, S. Chipman, & R. Brennan, Eds.). Lawrence Erlbaum Associates.

Drasgow, F., & Hulin, C. (1990). Item response theory (Vol. 1). Consulting Psychological Press.

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.

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

Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve. Journal of the American Statistical Association, 83(402), 414–425.

Engelkamp, J., Zimmer, H., Mohr, G., & Sellen, O. (1994). Memory of self-performed tasks: Self-performing during recognition. Memory & Cognition, 22(1), 34–39.

Fischer, G. (1973). The linear logistic test model as an instrument in educational research. Acta Psychologica, 37(6), 359–374.

Fischer, G. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), Rasch models: Foundations, recent developments, and applications (pp. 131–155). New York: Springer-Verlag.

Foos, P., & Smith, K. (1974). Effects of spacing and spacing patterns in free recall. Journal of Experimental Psychology, 103, 112–116.

Forrester, A. I. J., & Keane, A. J. (2009). Recent advances in surrogate-based optimization. Progress in Aerospace Sciences, 45, 50–79.

Fox, J. (2010). Bayesian item response theory. New York: Springer.

Gavrilov, L. A., & Gavrilova, N. S. (2001). The reliability theory of aging and longevity. Journal of Theoretical Biology, 213(4), 527–545.

Ghahramani, Z. (1995). Factorial learning and the EM algorithm. In Advances in neural information processing systems (pp. 617–624).

Glenberg, A. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. Journal of Verbal Learning and Verbal Behavior, 15, 1–16.

Glenberg, A., Bradley, M., Kraus, T., & Renzaglia, G. (1983). Studies of the long-term recency effect: Support for a contextually guided retrieval hypothesis. Journal of Experimental Psychology: Learning, Memory, and Cognition, 9(2), 231 - 255.

Glenberg, A., Bradley, M., Stevenson, J., Kraus, T., Tkachuk, M., Gretz, A., ... Turpin, B. (1980). A two-process account of long-term serial position effects. Journal of Experimental Psychology: Human Learning and Memory, 6(4), 355 - 369.

Glenberg, A., & Kraus, T. (1981, Nov). Long-term recency is not found on a recognition test. Journal of Experimental Psychology: Human Learning and Memory, 7(6), 475 - 479.

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. Journal of Experimental Psychology: General, 130, 116–139.

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

Greene, R. (1986). A common basis for recency effects in immediate and delayed recall. Journal of Experimental Psychology: Learning, Memory, and Cognition, 12, 413 - 418.

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.

Hartigan, J. A. (1972). Direct clustering of a data matrix. Journal of the american statistical association, 67(337), 123–129.

Hofmann, T., Puzicha, J., & Jordan, M. (1999). Learning from dyadic data. In Advances in Neural Information Processing Systems 11. MIT Press.

Hser, Y., & Wickens, T. (1989). The effects of the spacing of test trials and study trials in paired-association learning. Educational Psychology, 9, 99–120.

Huguenard, B. R., Lerch, F. J., Junker, B. W., Patz, R. J., & Kass, R. E. (1997). Working memory failure in phone-based interaction. ACM Transactions on Computer-Human Interaction, 4, 67–102.

Izawa, C. (1976). Vocalized and silent tests in paired-associate learning. The American Journal of Psychology, 89, 681–693.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36, 1441-1451.

Jameson, A. (1996). Numerical uncertainty management in user and student modeling: An overview of systems and issues. User Modeling and User-Adapted Interaction, 5(3), 193–251.

Jost, A. (1897). Die assoziationsfestigkeit in ihrer abhangigkeit von der verteilung der wiederholungen [The strength of associations in their dependence on the distribution of repetitions]. Zeitschrift fur Psychologie und Physiologie der Sinnesorgane, 14, 436–472.

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

Kalman, R. (1960). A new approach to linear filtering and prediction problems. Transactions of the ASME—Journal of Basic Engineering, 35–45.

Kang, S. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. Memory & Cognition, 38, 1009–1017.

Kang, S., McDermott, K., & Roediger, H. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. European Journal of Cognitive Psychology, 19, 528–558.

Kang, S. H. K., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term: Should spacing be expanding or equal-interval. Psychonomic Bulletin & Review.

Kang, S. H. K., & Pashler, H. (2011). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. Applied Cognitive Psychology, 26, 97–103.

Karpicke, J., & Roediger, H. (2007a). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, 704–719.

Karpicke, J., & Roediger, H. (2010). Is expanding retrieval a superior method for learning text materials? Memory & Cognition, 38, 116–124.

Karpicke, J., & Roediger, H., III. (2007b). Repeated retrieval during learning is the key to long-term retention. Journal of Memory and Language, 57, 151–162.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In Proceedings of the 21st national conference on artificial intelligence.

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

Khajah, M., Lindsey, R., & Mozer, M. (2014). Maximizing students' retention via spaced review: Practical guidance from computational models of memory. Topics in Cognitive Science, 6, 157–169.

Khan, F., Zhu, X. J., & Mutlu, B. (2011). How do humans teach: On curriculum learning and teaching dimension. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Weinberger (Eds.), Advances in Neural Information Processing Systems 24 (pp. 1449–1457). La Jolla, CA: NIPS Foundation.

Koedinger, K., & MacLaren, B. (1997). Implicit strategies and errors in an improved model of early algebra problem solving. In Proc. of the 19th ann. conf. of the cog. sci. soc. (pp. 382–387). Hillsdale, NJ: Erlbaum.

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.

Koppenaal, L., & Glanzer, M. (1990). An examination of the continuous distractor task and the 'long term recency effect'. Memory & Cognition, 18, 183 - 195.

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

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? Psychological Science, 19, 585–592.

Kumar, D., & Klefsjo, B. (1994). Proportional hazards model: A review. Reliability Engineering and Systems Safety, 44, 177–188.

Küpper-Tetzel, C., & Erdfelder, E. (2012). Encoding, maintenance, and retrieval processes in the lag effect: A multinomial processing tree analysis. Memory, 20, 37–47.

Landauer, T., & Bjork, R. (1978). Practical aspects of memory. In M. Gruneberg, P. Morris, & R. Sykes (Eds.), (pp. 625–632). London: Academic Press.

Landauer, T., & Eldridge, L. (1967). Effect of tests without feedback and presentation-test interval in paired-associate learning. Journal of Experimental Psychology, 75, 290–298.

Lee, J., & Brunskill, E. (2012). The impact on individualizing student models on necessary practice opportunities. In Proceedings of the 5th international conference on educational data mining.

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

Lewis-Smith, M. (1975). Short-term memory as a processing deficit. American Journal of Psychology, 88, 605 - 606.

Lindsey, R., Lewis, O., Pashler, H., & Mozer, M. (2010). Predicting students' retention of facts from feedback during training. Proceedings of the 32nd Annual Conference of the Cognitive Science Society.

Lindsey, R., Mozer, M., Cepeda, N., & Pashler, H. (2009). Optimizing memory retention with cognitive models. In Proceedings of the 9th international conference on cognitive modeling.

Lindsey, R., Mozer, M., Huggins, W., & Pashler, H. (2013). Optimizing instructional strategies. In Neural information processing systems 26. La Jolla, CA: NIPS Foundation.

Lindsey, R., Shroyer, J., Pashler, H., & Mozer, M. (2014). Improving students' long-term knowledge retention through personalized review. Psychological Science, 25(3), 639–647.

Logan, J., & Balota, D. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. Aging, Neuropsychology, and Cognition, 15, 257–280.

Love, B. C., & Giguère, G. (2011). Idealized training in noisy situations improves generalization. (Presentation at the 52nd Annual Meeting of the Psychonomics Society, Seattle, WA)

Maddox, G., Balota, D., Coane, J., & Duchek, J. (2011). The role of forgetting rate in producing a benefit of expanded over equal spaced retrieval in young and older adults. Psychology & Aging, 26, 661–670.

Marr, D. (1982). Vision: A computational approach. San Francisco: Freeman & Co.

Martin, J., & van Lehn, 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.

Meeds, E., Ghahramani, Z., Neal, R. M., & Roweis, S. T. (2007). Modeling dyadic data with binary latent factors. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), Advances in Neural Information Processing Systems 19 (pp. 977–984). Cambridge, MA: MIT Press.

Melton, A. (1970). The situation with respect to the spacing of repetitions and memory. Journal of Verbal Learning and Verbal Behavior, 9, 956–606.

Menon, A., & Elkan, C. (n.d.). Dyadic prediction using a latent feature log-linear model. (Unpublished)

Menon, A., & Elkan, C. (2011). Link prediction via matrix factorization. Proceedings of the 2011 European conference on machine learning and knowledge discovery in databases, 437–452.

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

Miller, K., Jordan, M. I., & Griffiths, T. L. (2009). Nonparametric latent feature models for link prediction. In Advances in neural information processing systems (pp. 1276–1284).

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. Behavior Research Methods, Instruments, and Computers, 36, 630–633.

Mislevy, R. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. Applied Psychological Measurement, 11(1), 81–91.

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

Murdock, B. B., Jr. (1962). The serial position effect of free recall. Journal of Experimental Psychology, 64, 482 - 488.

Murray, I., & Adams, R. P. (2010). Slice sampling covariance hyperparameters of latent Gaussian models. In J. Lafferty, C. K. I. Williams, R. Zemel, J. Shawe-Taylor, & A. Culotta (Eds.), Advances in neural information processing systems 23 (pp. 1723–1731).

Murray, I., Adams, R. P., & MacKay, D. J. (2010). Elliptical slice sampling. Journal of Machine Learning Research, 9, 541–548.

Murray, R., van Lehn, K., & Mostow, J. (2004). Looking ahead to select tutorial actions: A decision-theoretic approach. International Journal of Artificial Intelligence in Education, 14(3).

Myung, J., Kim, C., & Pitt, M. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. Memory & Cognition, 28(5), 832–840.

Nairne, J. (1991). Positional uncertainty in long-term memory. Memory and Cogition, 19(4), 332 - 340.

Nairne, J., Neath, I., Serra, M., & Byun, E. (1997, Aug). Positional distinctiveness and the ratio rule in free recall. Journal of Memory and Language, 37(2), 155-166.

Navarro, D., Griffiths, T., Steyvers, M., & Lee, M. (2006). Modeling individual differences with Dirichlet processes. Journal of Mathematical Psychology, 50, 101–122.

Neath, I. (1993). Contextual and distinctive processes and the serial position function. Journal of Memory and Language, 32, 820 - 840.

Neath, I., & Crowder, R. (1990). Schedules of presentation and temporal distinctiveness in human memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16(2), 316 - 327.

Neath, I., & Crowder, R. (1996). Distinctiveness and very short-term serial position effects. Memory, 4, 225 - 242.

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

Newell, A., & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), Cognitive skills and their acquisition (p. 1-55). Hillsdale, NJ: Erlbaum.

Nilsson, L.-G., Wright, E., & Murdock, B. B. (1975). The effects of visual presentation method on single-trial free recall. Memory & Cognition, 3(4), 427 - 433.

Osborne, M. A., Garnett, R., & Roberts, S. J. (2009). Gaussian processes for global optimization. In International conference on learning and intelligent optimization.

Pardos, Z., & Heffernan, N. (2010). Modeling individualization in a bayesian networks implementation of knowledge tracing. Proceedings of the 18th International Conference on User Modeling, Adaptation and Personalization, 255–266.

Pardos, Z., & Heffernan, N. (in press). Using HMMs and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset. Journal of Machine Learning Research W&CP.

Pashler, H., Mozer, M., & Wixted, J. D. (unpublished). Metrics of forgetting: weakening of associations versus skills.

Pashler, H., & Mozer, M. C. (2013). Enhancing perceptual category learning through fading: When does it help? Journal of Experimental Psychology: Learning, Memory, and Cognition.

Pashler, H., Rohrer, D., Cepeda, N., & Carpenter, S. (2007). Enhancing learning and retarding forgetting: Choices and consequences. Psychonomic Bulletin & Review, 14(2), 187–193.

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

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

Pavlik, P., Cen, H., & Koedinger, K. (2009). Performance factors analysis—a new alternative to knowledge tracing. In V. Dimitrova & R. Mizoguchi (Eds.), Proceedings of the 14th International Conference on Artificial Intelligence in Education. Brighton, England.

Pavlik, P., Presson, N., & Koedinger, K. (2007). Optimizing knowledge component learning using a dynamic structural model of practice. Proceedings of the International Conference on Cognitive Modeling.

Poltrock, S., & MacLeod, C. (1977). Primacy and recency in the continuous distractor paradigm. Journal of Experimental Psychology: Human Learning and Memory, 3(5), 560 - 571.

Postman, L., & Phillips, L. (1965). Short-term temporal changes in free recall. The Quarterly Journal of Experimental Psychology, 17, 132 - 138.

Putnam, A., & Roediger, H., III. (2013). Does response mode affect amount recalled or the magnitude of the testing effect? Memory & Cognition, 41(1), 36–48.

Pyc, M., & Rawon, K. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory. Journal of Memory and Language, 60, 437–447.

Qiu, Y., Qi, Y., Lu, H., Pardos, Z., & Heffernan, N. (2011). Does time matter? Modeling the effect of time with Bayesian knowledge tracing. The 4th International Conference on Educational Data Mining, 139–148.

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

Rafferty, A., Brunskill, E., Griffiths, T., & Shafto, P. (2011). Faster teaching by POMDP planning. Proceedings of The 15th International Conference on Artificial Intelligence in Education, 280–287.

Rafferty, A., LaMar, M., & Griffiths, T. (2012). Inferring learners knowledge from observed actions. The 5th International Conference on Educational Data Mining.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In Proceedings of the 4th Berkeley Symposium on Mathematical Statistics & Probability (pp. 321–333).

Rasmussen, C. E., & Williams, C. K. I. (2006). Gaussian processes for machine learning. Cambridge, MA: MIT press.

Reye, J. (2004). Student modelling based on belief networks. International Journal of Artificial Intelligence in Education, 14, 1–33.

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

Ritter, S., Anderson, J., Koedinger, K., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. Psychonomic Bulletin and Review, 14(2), 249–255.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. Psychological Review, 107(2), 358–367.

Roediger, H., & Karpicke, J. (2006a). The power of testing memory: Basic research and implications for educational practice. Perspectives on Psychological Science, 1, 181–210.

Roediger, H., & Karpicke, J. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17(3), 249–255.

Roediger, H., III, & Crowder, R. (1976). A serial position effect in recall of united states presidents. Bulletin of the Psychonomic Society, 8, 275 - 278.

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.

Rubin, D., Hinton, S., & Wenzel, A. (1999). The precise time course of retention. Journal of Experimental Psychology: Learning, Memory, and Cognition, 25, 1161 - 1176.

Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. (1989). Design and analysis of computer experiments. Statistical Science, 4, 409–435.

Salakhutdinov, R., Mnih, A., & Hinton, G. (2007). Restricted boltzmann machines for collaborative filtering. International Conference on Machine Learning, 791–798.

Salmon, J. P., McMullen, P. A., & Filliter, J. H. (2010). Norms for two types of manipulability (graspability and functional usage), familiarity, and age of acquisition for 320 photographs of objects. Behavioral Research Methods, 42, 82–95.

Schloss, K. B., & Palmer, S. E. (2011). Aesthetic response to color combinations: preference, harmony, and similarity. Attention, Perception, & Psychophysics, 73, 551–571.

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

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

Sederberg, P., Howard, M., & Kahana, M. (2008). A context-based theory of recency and contiguity in free recall. Psychological Review, 115(4), 893 - 912.

Shan, H., & Banerjee, A. (2008). Bayesian co-clustering. In Eighth IEEE international conference on data mining (pp. 530–539).

Smith, M., Roediger, H., III, & Karpicke, J. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. Journal of Experimental Psychology: Learning, Memory, and Cognition, 39(6), 1712–1725.

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

Srinivas, N., Krause, A., Kakade, S., & Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In Proceedings of the 27th international conference on machine learning. Haifa, Israel.

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

Stewart, T., & West, R. (2007). Deconstructing and reconstructing ACT-R: Exploring the architectural space. Cognitive Systems Research, 8(3), 227–236.

Stigler, S. M. (1978, Jan.). Some forgotten work on memory. Journal of Experimental Psychology: Human Learning and Memory, 4(1), 1–4.

Storm, B., Bjork, R., & Storm, J. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. Memory & Cognition, 38, 244–253.

Talmi, D., & Goshen-Gottstein, Y. (2006). The long-term recency effect in recognition memory. Memory, 14(4), 424 - 436.

Thai-Nghe, N., Drumond, L., Horváth, T., Krohn-Grimberghe, A., Nanopoulos, A., & Schmidt-Thieme, L. (2011). Factorization techniques for predicting student performance. In O. C. Santos & J. G. Boticario (Eds.), Educational recommender systems and technologies: Practices and challenges.

Thai-Nghe, N., Drumond, L., Horváth, T., Nanopoulos, A., & Schmidt-Thieme, L. (2011). Matrix and tensor factorization for predicting student performance. In Proceedings of the 3rd international conference on computer supported education.

Thai-Nghe, N., Horváth, T., & Schmidt-Thieme, L. (2011). Factorization models for forecasting student performance. In Proceedings of the 4th international conference on educational data mining.

Thapar, A., & Greene, R. (1993). Evidence against a short-term store account of long-term recency effects. Memory and Cognition, 21, 329 - 337.

Thios, S., & D'Agostino, P. (1976). Journal of verbal learning and verbal behavior. Effects of repetition as a function of study-phase retrieval, 15, 529–536.

Titsias, M. K., Lawrence, N. D., & Rattray, M. (2008). Efficient sampling for gaussian process inference using control variables. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), Advances in Neural Information Processing Systems 21 (pp. 1681–1688). MIT Press.

Todorov, E. (2006). Optimal control theory. In K. Doya (Ed.), Bayesian brain. MIT Press.

Toscher, A., & Jahrer, M. (2010). Collaborative filtering applied to educational data mining. In KDD cup 2010: Improving cognitive models with educational data mining.

Tsai, L. (1927). The relation of retention to the distribution of relearning. Journal of Experimental Psychology, 10, 30–39.

Tzeng, O. (1973). Positive recency effect in delayed free recall. Journal of Verbal Learning and Verbal Behavior, 12(4), 436 - 439.

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 (pp. 17–20).

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

Villano, M. (1992). Probabilistic student models: Bayesian belief networks and knowledge space theory. In Proceedings of the second international conference on intelligent tutoring systems.

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.

Wahlheim, C., Maddox, G., & Jacoby, L. (2014). The role of reminding in the effects of spaced repetitions on cued recall: Sufficient but not necessary. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40, 94–105.

Welch, G. (2002). An introduction to Kalman filtering. Technical Report, Department of Computer Science and Engineering, University of North Carolina at Chapel Hill.

Whitehill, J. (2013). Understanding act-r: An outsider's perspective. (Unpublished)

Whitehill, J., & Movellan, J. R. (2010). Optimal teaching machines (Tech. Rep.). La Jolla, CA: Department of Computer Science, UCSD.

Wickelgren, W. (1974). Single-trace fragility theory of memory dynamics. Memory and Cognition, 2, 775 - 780.

Wickelgren, W. (1976). Handbook of learning and cognitive processes (Vol. 6). John Wiley & Sons Inc.

Wixted, J. (2004a). On common ground: Jost's (1897) law of forgetting and ribot's (1881) law of retrograde amnesia. Psychological Review, 111, 864 - 879.

Wixted, J. (2004b). The psychology and neuroscience of forgetting. Annual Review of Psychology, 55, 235–269.

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

Wixted, J., & Ebbesen, E. (1991). On the form of forgetting. Psychological Science, 2, 409 - 415.

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

Zechmeister, E., & Shaughnessy, J. (1980). When you know that you know and when you think that you know but you don't. Bulletin of the Psychonomic Society, 15, 41–44.