Predicting Individual Differences in Student Learning via Collaborative Filtering

Robert V. Lindsey
Department of Computer Science
University of Colorado
Boulder, CO USA
robert.lindsey@colorado.edu

Michael C. Mozer
Department of Computer Science
and Institute of Cognitive Science
University of Colorado
Boulder, CO USA
mozer@colorado.edu

Abstract

Effective teaching requires an understanding of a student's knowledge state—what material the student has and has not mastered and what material is fragile and easily lost. To facilitate automated teaching, our goal is to construct models that infer the knowledge state of individual students for specific elements of knowledge. The challenge to inference is that the available evidence is quite weak. 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? To overcome the sparsity of observations, we use a collaborative filtering approach that leverages information about a population of students studying a population of items (elements of knowledge) to infer how well a specific student has learned a specific item. We extend *item-response theory*, a traditional class of models that recover latent traits of students and items, to address the facts that knowledge state is nonstationary and that both observations and predictions may span a broad range of time. This extension is based on a psychological model of memory that can take into account dynamic information about study history. We evaluate three alternative models whose latent variables are determined either via maximum likelihood estimation or a hierarchical Bayesian approach. We show for two different student-learning data sets that, when we combine multiple weak sources of information from the population, we can make strong inferences about an individual student's knowledge and performance.

1 Introduction

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., [1–3]) 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. Applicable domains range from the concrete to the abstract, and from the perceptual to the cognitive, and 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).

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.

Inferring a student's knowledge state from behavioral evidence is a daunting challenge because behavioral evidence is fairly weak. It can trivially include whether or not students correctly answered specific questions in the past, but can include subtler forms of evidence as well. For correct answers, the time to respond might be diagnostic of the knowledge state; for erroneous answers, the response itself might be diagnostic. Valuable information also comes from the *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 particularly fragile when the material being learned is unfamiliar [4, 5]. Further, the temporal distribution of practice has an impact on the durability of learning for various types of material [6, 7].

Consider declarative (fact) learning, the domain we will use as an illustration throughout this paper. If a student studies via cued retrieval practice, as when flashcards are used for drilling, one bit of information is obtained about the student's memory state for a fact: either the fact is available or it is not. From this meager information, we hope to then predict whether the fact will be accessible in an hour, a week, or a month.

Complicating the prediction problem is the ubiquity of individual differences in every form of learning. Taking an example from fact learning, Figure 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, inter-item variability is a consideration. Learning a foreign vocabulary word may be easy if it is similar to its English equivalent, but hard if it is similar to a different English word. Figure 1b shows the distribution of recall accuracy for 120 Lithuanian-English vocabulary items averaged over a set of students. With a single round of study, an exam administered several minutes later suggests that items show a tremendous range in difficulty (*krantas→shore* was learned by only 3% of students; *lova→bed* was learned by 76% of students). Although some of this variability is due to measurement error, the importance of inter-item differences is acknowledged by psychologists, and in fact, these data were collected in order to determine difficulty norms for individual items.

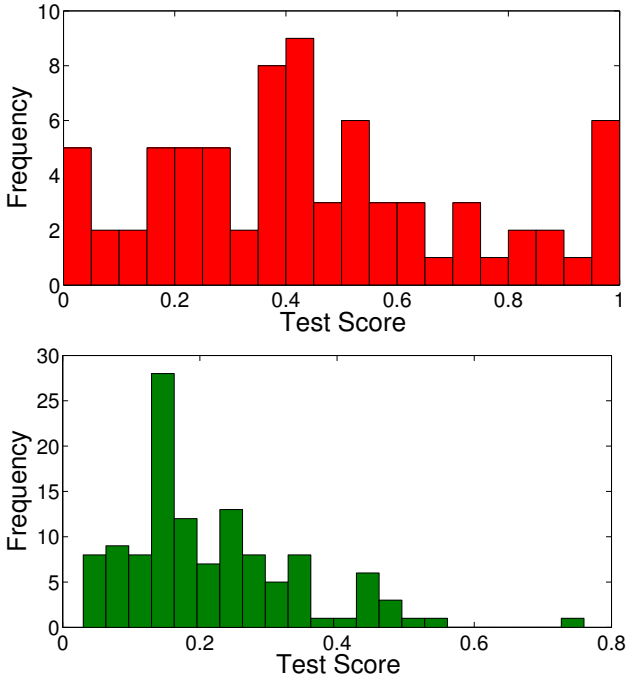

Figure 1: (upper) Histogram of proportion of items reported correctly on a cued recall task for a population of 60 students learning 32 Japanese-English vocabulary pairs [8]; (lower) Histogram of proportion of subjects correctly reporting an item on a cued recall task for a population of 120 Lithuanian-English vocabulary pairs being learned by roughly 80 students [9]

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.

2 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 [10]. 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.

2.1 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:

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

(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 [11].

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 [12]. Extensions have been proposed to allow for a student to have a different ability at different times [13], 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.

2.2 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,

$$Pr(\text{recall}) = m(1 + ht)^{-f}, \quad (2)$$

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 [14].

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 [15], we could individuate forgetting for each student-item by setting the decay exponent based on latent IRT-like traits:

$$Pr(R_{si} = 1) = m(1 + ht_{si})^{-\exp(\tilde{a}_s - \tilde{d}_i)}, \qquad (3)$$

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 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

$$Pr(R_{si} = 1) = \frac{(1 + ht_{si})^{-f}}{1 + \exp(d_i - a_s)}. \qquad (4)$$

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

$$Pr(R_{si} = 1) = \frac{(1 + ht_{si})^{-\exp(\tilde{a}_s - \tilde{d}_i)}}{1 + \exp(d_i - a_s)}. \qquad (5)$$

Both this hybrid model and Equation 4 simplify to 1PL (Equation 1) at $t = 0$. For $t > 0$, recall probability decays as a power-law function of time.

2.3 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 1);
- MEMORY: a power-law forgetting model with population-wide parameters (Equation 2);
- HYBRID DECAY: a power-law forgetting model with decay rates based on latent student and ite traits (Equation 3);
- HYBRID SCALE: a power-law forgetting model with the degree-of-learning based on latent student and item traits (Equation 4); and
- HYBRID BOTH: a power-law forgetting model that individuates both the decay rate and degree-of-learning (Equation 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 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.

2.4 Simulation methodology

We employed Markov chain Monte Carlo techniques for posterior inference in the Bayesian models presented in Table 1. Gibbs sampling isn't feasible in our models, but we can use Metropolis-within-Gibbs [16], an extension of Gibbs sampling wherein each 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

$$p(a_s \mid \mathbf{a}_{\neg \mathbf{s}}, \mathbf{d}, h, m, R) \propto \prod_i P(r_{si} \mid a_s, d_i, h, m) \times \left(1 + \frac{a_s^2}{2(\psi_2 + \frac{1}{2}\sum_{j \neq s} a_j)}\right)^{\psi_1 + \frac{n_S - 1}{2}} \qquad (6)$$

where the first term is given by Equation 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

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 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	Anonymous (2012)	Anonymous (2008)
Materials	Japanese-English vocabulary	Interesting but obscure facts
# Students	32	1354
# Items	60	32
Rounds of Practice	3	1
Retention Intervals	3 min–27 days	7 sec–53 min

Table 2: Experimental data used for simulations

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 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 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 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, 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 [17].

Figure 2 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 summarizes the AUC values for Study $\mathcal{S}_1$. The baseline MEMORY model is trounced

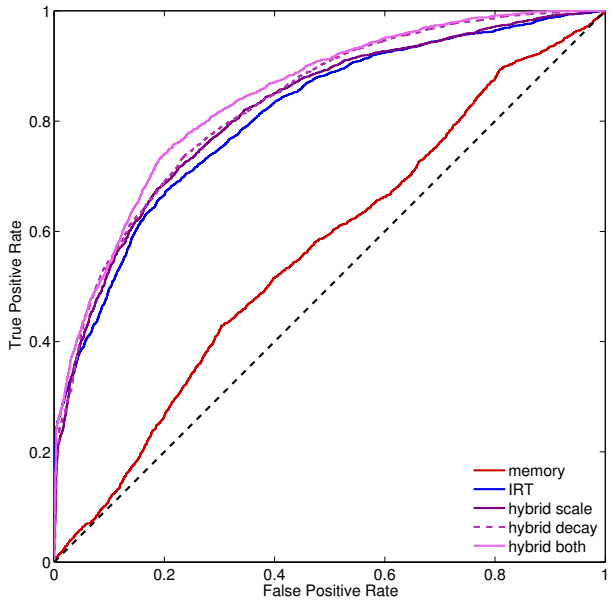

Figure 2: Mean ROC curves for the Bayesian models on held out data from Study $\mathcal{S}_1$.

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 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 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.

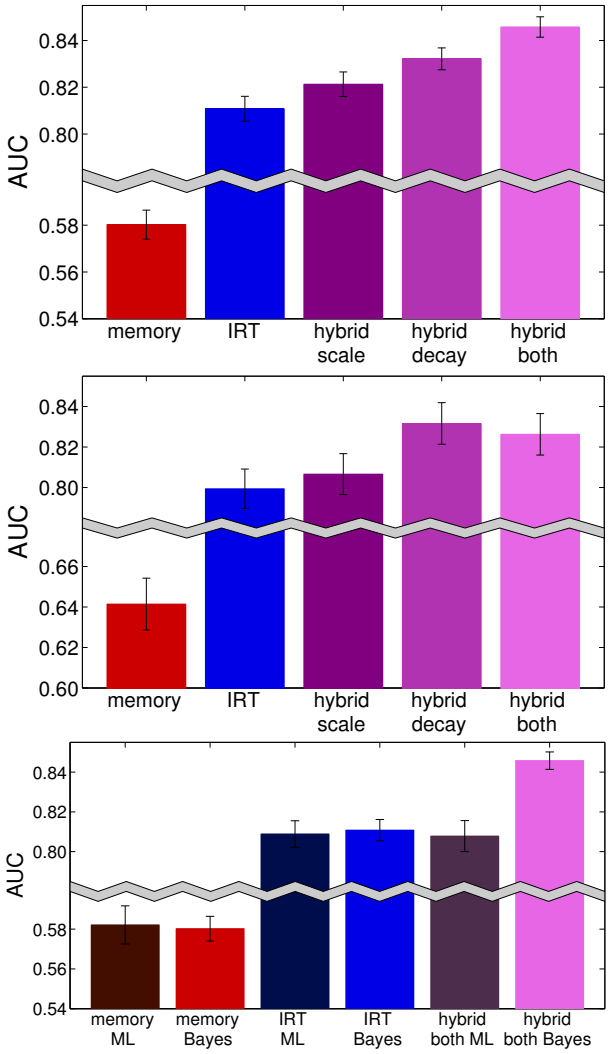

Figure 3: The top and middle graphs show mean AUC values on the five BAYES models trained and evaluated on Studies $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. The bottom graph compares BAYES and ML versions of three models on Study $\mathcal{S}_1$. The error bars indicate a 95% confidence interval 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.

3.1 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 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 4 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.

4 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 [15]. 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.

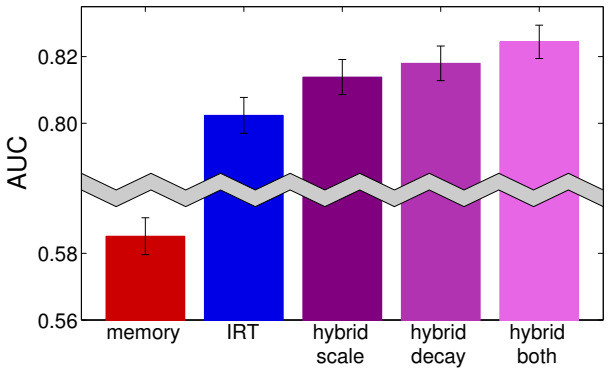

Figure 4: Mean AUC values when random items are heldout during validation folds, Study $\mathcal{S}_1$

Acknowledgment

This research was supported by NSF grants BCS-0339103 and BCS-720375. The first author is supported by an NSF Graduate Research Fellowship.

References

[1] J. R. Anderson, F. G. Conrad, and A. T. Corbett, "Skill acquisition and the LISP tutor," *Cognitive Science*, vol. 13, pp. 467–506, 1989.

[2] K. R. Koedinger and A. T. Corbett, *Cognitive Tutors: Technology bringing learning science to the classroom.* Cambridge UK: Cambridge University Press, 2006, pp. 61–78.

[3] J. Martin and K. VanLehn, "Student assessment using Bayesian nets," *International Journal of Human-Computer Studies*, vol. 42, pp. 575–591, 1995.

[4] D. Rohrer and K. Taylor, "The effects of overlearning and distributed practice on the retention of mathematics knowledge," *Applied Cognitive Psychology*, vol. 20, pp. 1209–1224, 2006.

[5] J. Wixted, "The psychology and neuroscience of forgetting," *Annual Review of Psychology*, vol. 55, pp. 235–269, 2004.

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

[7] T. Rickard, J. Lau, and H. Pashler, "Spacing and the transition from calculation to retrieval," *Psychonomic Bulletin & Review*, vol. 15, pp. 656–661, 2008.

[8] S. H. K. Kang, R. V. Lindsey, M. C. Mozer, and H. Pashler, "Retrieval practice over the long term: Expanding or equal-interval spacing?" 2012.

[9] P. J. Grimaldi, M. A. Pyc, and K. A. Rawson, "Normative multitrial recall performance, metacognitive judgments and retrieval latencies for Lithuanian-English paired associates," *Behavioral Research Methods*, vol. 42, pp. 634–642, 2010.

[10] P. DeBoek and M. Wilson, Eds., *Explanatory item response models. A generalized linear and nonlinear approach.* New York: Springer, 2004.

[11] J.-P. Fox, *Bayesian Item Response Theory.* New York: Springer, 2010.

[12] L. A. Roussos, J. L. Templin, and R. A. Henson, "Skills diagnosis using IRT-based latent class models," *Journal of Educational Measurement*, vol. 44, pp. 293–311, 2007.

[13] D. F. Andrade and H. R. Tavares, "Item response theory for longitudinal data: population parameter estimation," *Journal of Multivariate Analysis*, vol. 95, pp. 1–22, 2005.

[14] J. T. Wixted and S. K. Carpenter, "The Wickelgren power law and the Ebbinghaus savings function," *Psychological Science*, vol. 18, pp. 133–134, 2007.

[15] P. I. J. Pavlik and J. R. Anderson, "Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect," *Cognitive Science*, vol. 29, pp. 559–586, 2005.

[16] R. J. Patz and B. W. Junker, "A straightforward approach to Markov chain Monte Carlo methods for item response models," *Journal of Educational and Behavioral Statistics*, vol. 24, pp. 146–178, 1999.

[17] D. M. Green and J. A. Swets, *Signal detection theory and psychophysics.* New York: Wiley, 1966.