Move your lamp post: Recent data reflects learner knowledge better than older data

April Galyardt
University of Georgia
galyardt@uga.edu

Ilya Goldin
Pearson
ilya.goldin@pearson.com

In educational technology and learning sciences, there are multiple uses for a predictive model of whether a student will perform a task correctly or not. For example, an intelligent tutoring system may use such a model to estimate whether or not a student has mastered a skill. We analyze the significance of data recency in making such predictions, i.e., asking whether relatively more recent observations of a student's performance matter more than relatively older observations. We investigate several representations of recency, such as the count of prior practice in the AFM model, and the proportion of recent successes with exponential and box kernels. We find that an exponential decay of a proportion of successes provides the summary of recent practice with the highest predictive accuracy over alternative models. As a secondary contribution, we develop a new logistic regression model, Recent-Performance Factors Analysis, that leverages this representation of recent performance, and has higher predictive accuracy than existing logistic regression models.
Keywords & Phrases: Student modeling, Linear logistic test model, instance weighting, adaptive learning

1. INTRODUCTION

A central field of research in educational technology and assessment is concerned with modeling the probability that a student will respond correctly to some question. This modeling is used to analyze test answers, as with Item Response Theory; in adaptive learning technologies, such as the use of Bayesian Knowledge Tracing (Corbett and Anderson, 1995) in intelligent tutoring systems; to analyze the domains that students study, such as the study of transfer across tasks (Pavlik et al., 2015); and to understand behaviors such as gaming the system (Baker et al., 2004).

Our work examines several different representations of recency. The intuition is simple: as students practice a skill, we expect their understanding to increase and their performance to improve. Having recently succeeded at a task may be an indicator that learning has taken place, and such a moment of learning ought to contribute to our prediction of successful performance. This work is the first thorough investigation of recency effects in performance modeling.

We begin by describing a space of representations of student performance, including representations of the entirety of observed student practice and only a recent subset. We further incorporate these representations into the Linear Logistic Test Model framework. This framework subsumes many existing modeling efforts, including the Additive Factors Model (AFM) (Cen et al., 2006), Performance Factors Analysis (PFA) (Pavlik et al., 2009), and the recency-

weighted model by Gong and colleagues (Gong et al., 2011). We then propose the Recent-Performance Factors Analysis (R-PFA) in this framework; this model emphasizes recent over total practice. We evaluate the representations of practice on a real-world dataset of practice from the Assistments system (Heffernan and Heffernan, 2014). Finally, since real-world datasets exhibit certain data limitations including replicability, we further examine the properties of the new R-PFA model and several alternatives on a range of simulated datasets.

2. Related Work in Performance Modeling

Models of student performance represent the probability that a student will respond to a task in a particular way. For example, Bayesian Knowledge Tracing (BKT) (Corbett and Anderson, 1995) represents this as a first-order Hidden-Markov model (HMM). Models of student performance have many uses: to predict a student's response; to describe a latent property of the task or the student; to consider what information bears on the prediction, and the relationship among the sources of variance (Junker, 2011). By considering different representation of the history of practice, our work falls into this latter category. Thus, even though our investigation uses predictive models as a framework, the goal is not predictive accuracy *per se*.

As choices for a modeling framework, there are two prominent approaches in the literature: graphical models, including BKT, and logistic regression models. The original Corbett and Anderson BKT model describes students moving from a state where skills are unknown to a state where they have learned a skill. It assumes that all students learn at the same skill-specific rate, and that all items associated with a skill have the same difficulty. While the generative structure of BKT represents (to a degree) human learning, this HMM structure also leads to mathematical complexity and may lead to implausible parameter values (Beck and Chang, 2007). There have been many attempts to modify BKT structure (e.g., Baker et al. 2011; Beck et al. 2008; Falakmasir et al. 2015), and a variety of methods for estimating BKT models (e.g., Falakmasir et al. 2013; Boots et al. 2011). It is also possible to incorporate covariates such as student ability and item difficulty into Knowledge Tracing (González-Brenes et al., 2014; Khajah et al., 2014), although BKT with these extensions had lower predictive accuracy than the one-parameter item response theory (1PL-IRT) model on three out of four datasets from intelligent tutors (Khajah et al., 2014).

In comparison to BKT, logistic regression models are simple, numerically stable, and well understood. The linearly additive structure of logistic regression has facilitated investigations of the history of practice with hints across problems (Chi et al., 2011) and within the same problem (Goldin et al., 2012). In the same way, linearly additive models provide a flexible test bed for our investigation of recency.

Linear Logistic Test Models

The Linear Logistic Test Models (LLTM) (Fischer, 1973; de Boeck and Wilson, 2004) are logistic regression models with parameters for skill difficulty and student ability. The related Rasch and Item Response Theory (IRT) models include parameters for individual task difficulty, rather than grouping together tasks that share an underlying skill. Because task difficulty is estimated from data, replacing per-task parameters with per-skill parameters in LLTM reduces the number of model parameters, and leverages the power of task-level observations to provide a relatively more robust estimate of skill difficulty.

Other IRT models explore considerations such as partial credit modeling, complex parametrizations of items and students, responses to items with multiple correct answers, and interference

from distractor answer options (de Boeck and Wilson, 2004). However, IRT models tend to be applied to data from test administrations, not learning environments, and they do not account for change over time, let alone for recency. The MRCML model (Adams et al., 1997) can account for change in student ability between test administrations, but it does not consider recency. Similarly, variants of Diagnostic Classification Models (Rupp et al., 2010) target assessment, not data with evidence of learning.

The Additive Factors Model (AFM) and Performance Factors Analysis (PFA) reflect a student's history of practice (Cen et al., 2006; Pavlik et al., 2009). AFM includes a slope coefficient for the total number of prior opportunities a student has had to practice a skill. The claim is that the more practice a student has had, the more likely they should be to answer the next item correctly. PFA separates the number of total prior practice opportunities into counts of successes and failures, representing the claim that successful and unsuccessful practice may have differential value for student learning and thus for probability of correctness on the next task.

PFA-decay (Gong et al., 2011) uses the same counts of successes and counts of failures as PFA, and further down-weights older practice. Gong et al. (2011) found very modest improvements of PFA-decay over PFA. The representation of recent practice history that we develop in section 3. is based on non-parametric kernel weighting, and subsumes PFA-decay as a special case. We demonstrate in section 5.2. that the primary reason Gong et al. (2011) observed only modest improvement is that PFA-decay places too much weight on older attempts.

3. A Representation of Recent Practice

Informally, when a student is learning a new skill, we generally expect that initial attempts to apply the skill may be unsuccessful, and that as learning occurs, attempts will be predominantly successful. We can relate this idea to a 'moment-of-learning'. If a student has been successful in recent practice, then a moment-of-learning has likely already occurred. If recent attempts have not been successful, then the student has most likely not yet learned the skill. By contrast, evidence from older attempts may not be as informative as more recent evidence with regard to whether a moment of learning has taken place.

The question then is how to summarize a student's recent history. First, how many trials are sufficient to cover "recent"? Should all these "recent" trials be counted equally, or should the most recent evidence receive more weight? To resolve this, we use kernel weighting functions (Wasserman 2006, p.55). For example, a box kernel weights all of the most recent K attempts equally, while with an exponential kernel, attempts farther in the past receive exponentially smaller weights.

Second, the simplest representation would be either a count or proportion of successes out of the most recent several trials, but which is better? Because a proportion is bounded between 0 and 1, no matter the choice of kernel, or the length of the recent history being considered, coefficients in a model using proportions will have the same scale (e.g., the effect of ten percent of recent practice being successful, twenty percent, etc.). By contrast, counts are not bounded, and therefore their coefficients are more difficult to interpret. For example, looking at the most recent three practice opportunities vs. the most recent five opportunities, the count has a different maximum and the scale of the coefficients will change. (Proportions have a predictive advantage as well, which we will discuss later.)

Proportions reveal a complexity that is hidden when using a count of successes to summarize a student's history. At the beginning of practice, the count of prior successes is zero because

no data have been observed. The proportion of prior successes is (zero successes)/(zero opportunities), i.e., not mathematically defined. One way to resolve this is to make the assumption that if the student has had any unobserved prior practice, it has been unsuccessful. Notably, the count of prior successes implicitly contains the same assumption (zero prior successes). Computationally, we implement this using a trivial technique we call "ghost attempts": we prepend a small number of unsuccessful attempts to a student's practice history, forcing the proportion of prior successes on the first observed practice attempt to be zero. In this way, the intercept can be interpreted as the difficulty of the skill for a student whose recent proportion of successes is zero, and the slope of the proportion of successes can be interpreted as the effect of observing proportionally more successes in recent practice. As a side benefit, this reduces noise in the predictor, which in turn reduces the standard error of the coefficients.

Recent performance includes both successful and unsuccessful practice, and a representation of recency can include both. It may be that it is sufficient to consider total practice, without distinguishing the two; that successes and failures should be treated separately but analogously, or that distinct representations are appropriate. Below, we discuss recent history in terms of successes, but all of these ideas apply equally to the calculation of a recent proportion of failures.

3.1. Defining "Recent" Using a Kernel Function

To formally define the recent proportion of success, we first introduce the following notation:

j	skill index, $j = 1, \ldots, J$
i	student index, $i = 1, \ldots, N$
t	practice opportunity index, $t = 1, \ldots, O_{ij}$
X_{ijt}	response by student i, on opportunity t of skill j, $X_{ijt} = \begin{cases} 0 & \text{if incorrect} \\ 1 & \text{if correct} \end{cases}$
p_{ijt}	Probability of a correct response: $Pr(X_{ijt} = 1)$
T_{ijt}	count of past opportunities
S_{ijt}	recency-weighted count of previous successes, up to trial t
F_{ijt}	recency-weighted count of previous failures, up to trial t
R_{ijt}	recency-weighted proportion of past successes

The general formula for the proportion of recent success is

$$R_{ijt} = \sum_{\ell=(1-g)}^{t-1} w_{t\ell} X_{ij\ell} \tag{1}$$

where $w_{t\ell}$ is an appropriate weight, and $1-g$ is the start of practice including any ghost attempts. How much practice history is included in the proportion of recent success is controlled by the weights $w_{t\ell}$ in equation 1. A general kernel weighting framework allows for calculating $w_{t\ell}$ in a principled way that still allows the substitution of different kernels for different purposes; however, we note that the choice of kernel function is usually less important than the bandwidth of the kernel (Wasserman, 2006).

To calculate the proportion of success over the entire history of practice (including g ghost attempts), then

$$w_{t\ell} = \frac{1}{t - 1 + g}. \tag{2}$$

Whereas, for the proportion of success over the last K attempts, the weights are

$$w_{t\ell} = \begin{cases} \frac{1}{K} & \text{if } t - \ell < K \\ 0 & \text{otherwise} \end{cases} . \tag{3}$$

The box kernel in equation 3 gives equal weight to each attempt within the window. We shall refer to R_{ijt} calculated with these weights as R^{box}_{ijt}. By definition, the box kernel is not a smooth kernel and in most applications smooth kernels have slightly better performance (Wasserman, 2006).

An exponential kernel

$$w_{t\ell} = \frac{d^{t-\ell}}{\sum_{\ell=(1-g)}^{t-1} d^{t-\ell}} \tag{4}$$

is a smooth kernel that down-weights older attempts according to the decay parameter d. For simplicity, in general, when we say R_{ijt}, we are referring to R computed with this exponential kernel. As necessary to avoid confusion, we will denote it as R^{exp}_{ijt}. Different values of d control the 'smoothing' over the history of practice. If $d = 1$, then equation 4 simplifies to equation 2, and a student's entire history of practice gets equal weight. Alternatively, if $d = 0.1$, then 90% of the weight is on the single previous trial, and 9% is placed on the 2nd most recent attempt, so that effectively only the last attempt is counted in the recent history. For exponential decay, d ranges from 0 to 1, while for the box kernel, the window size K ranges from 1 to infinity, so that selecting the optimal d has a more tractable search space than K. Thus the exponential kernel has a computational advantage for tuning decay weight, the advantage of smoothness, and an interpretability advantage since older evidence is down-weighted.

Gong et al. (2011) uses an un-normalized exponential decay function in PFA-decay;

$$w_{t\ell} = d^{t-\ell}. \tag{5}$$

These un-normalized weights do not create a recent proportion of success, but a recent count of success. Note also that when the weights are un-normalized, including ghost attempts has no effect, since it just adds zeros to the total. From that perspective, S_{ijt} already includes an infinite number of ghost attempts.

$$S_{ijt} = \sum_{\ell=(1-g)}^{t-1} w_{t\ell} X_{ij\ell} = \sum_{\ell=0}^{t-1} d^{t-\ell} X_{ij\ell} \tag{6}$$

Another transformation that can be applied to count variables is a logarithmic transformation (e.g., Yudelson et al. 2014; Chi et al. 2011). The intuition behind it is that the probability of a correct should not increase infinitely with practice, and additional evidence should have diminishing returns. Computationally, the logarithmic transformation places more weight on early attempts, and diminishing weight on later attempts. This transformation is effectively the opposite of the exponential kernel.

KERNEL BANDWIDTH Kernel functions have a bandwidth tuning parameter that controls the "width" of the kernel: K for the box kernel, d for the exponential. Figure 1 illustrates the effect of bandwidth tuning for three different success representations over three different practice histories.

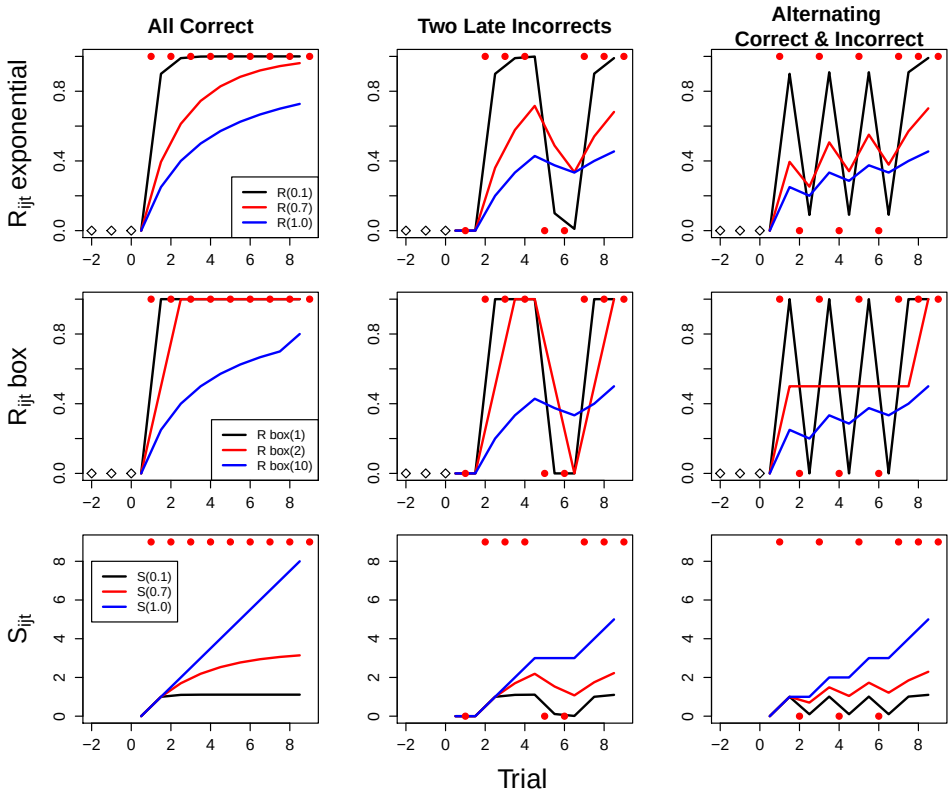

Figure 1: Effect of decay weighting on a (potentially decayed) count of successes S_{ijt} (bottom row), proportion of successes R_{ijt} with a box kernel (middle) and exponential kernel (top). Red circles indicate observed successful (top) and unsuccessful (bottom) practice. Black diamonds indicate ghost attempts. Blue line shows the tuning parameter with the widest bandwidth (longest memory), black line shows the narrowest bandwidth, red line shows the optimal bandwidth.

The simplest case is a count of successes S_{ijt} (Figure 1, bottom row). As the student accumulates correct answers (red circles at top of panel), S_{ijt} with no transformations (the blue line) grows indefinitely. Incorrect answers (in the "Two Late Incorrects" and "Alternating Correct & Incorrect" panels, indicated by red circles at bottom of panel) cause S_{ijt} with no transformation to plateau, and correct answers (red circles) make it grow again. The exponential decay of S_{ijt} in equation 6 gives more weight to more recent observations, and less to older observations. The black line shows the effect of a very steep decay rate (narrow bandwidth), and the red line shows an optimal decay rate according to the results in section 6.

The next case is a proportion of successes in a lookback window R^{box}_{ijt} (Figure 1, middle row). The box kernel discards all evidence outside the window, and gives equal weight to all evidence in the window (equation 3). The top row of figure 1 shows R^{exp}_{ijt}. The exponential kernel gives more weight to more recent evidence, and down-weights older evidence. The tuning parameter is the decay rate. Ghost attempts are included here in the calculation of R, and are discussed in section 3.2.

Notice that the student histories as summarized by R^{box} and R^{exp} are highly similar. When the bandwidth is small (black lines), there is very little difference between the two kernels. For either kernel, all or most of the weight is on the single last attempt. The effects of this are most visible on the pattern "Alternating Correct & Incorrect": at this small bandwidth $R \approx 1$ if the most recent attempt was a success, and $R \approx 0$ if the most recent attempt was a failure. At large bandwidths (blue lines) again both kernels generate similar summaries of a students practice (although with $d = 1$, R^{exp} is equally weighting each attempt in the entire history of practice, while with $K = 10$, R^{box} is equally weighting the most recent 10 attempts). It is at the middle bandwidth that the differences between the two kernels become most apparent. R^{box} equally weights the most recent K attempts, and can change rapidly, as happens in the "Two Late Incorrects" panel: R^{box} moves from 1.0 to 0.0 in only 2 attempts. In contrast, the exponential kernel smoothly down-weights the older evidence. In practice, this is why smoother kernels are generally preferred over the box kernel (Wasserman, 2006, p.56).

When bandwidth is large and maintains a longer memory over a student's practice history (blue), the proportion of recent success R may give too much weight to older evidence, i.e., fail to reflect recent learning. If bandwidth is small and the window over recent practice is narrow (black), then wide fluctuations in R may make the term useless as a predictor of future performance. We can see the effect of the different bandwidths most clearly in the pattern "Two Late Incorrects". This student has the attempt history $X_{ij} = (0, 1, 1, 1, 0, 0, 1, 1, 1)$, as indicated by the red circles. With decay $d = 1.0$ (blue line), each attempt in the entire history of practice receives equal weight, so that R decreases very little when the student misses two items in a row, but it also increases very little when the student responds correctly three times in a row. On the other hand, when $d = 0.1$ (black line), 90% of the weight is on the single most recent attempt, so that after three correct attempts in a row, $R \approx 1$, but after missing two attempts, $R \approx 0$.

The behavior of R^{exp}_{ijt} with $d = 0.7$ mirrors our intuition. If we were tutoring a student one-on-one, on the third correct attempt in a row, we might think 'Ok, the student seems to be learning the skill.' When the next attempt is incorrect, we might think 'That might have just been a slip.' On the second incorrect attempt, we might revise our assessment of the student's knowledge: 'Maybe the student doesn't know this.' But after 3 subsequent correct responses in a row, we might be fairly convinced the student has learned the skill. This parallels exactly the Bayesian updating of the probability that a student has learned a skill that takes place in a Bayesian Knowledge Tracing model. In this way, exponential decay weighting is capturing

student performance in a similar way to BKT, but without the complexity of a Hidden Markov Model.

Gong et al. (2011) fix d at 0.9, aiming not to "eliminate the effects of further practices too quickly." This is an overly simplistic choice, and as we shall demonstrate in section 5.2., tuning the decay parameter appropriately produces large improvements in predictive accuracy, and tuning the decay parameter for successes and failures separately is even more effective.

Here, we have illustrated the effects of the tuning parameter on two different kernels, the box kernel and the exponential kernel. It is certainly possible to use other kernel weighting functions (half-Gaussian, Epanechnikov, quadratic, etc.), but this illustration should emphasize why the choice of bandwidth is more important than the choice of kernel.

3.2. Introducing Prior Belief through Ghost Attempts

Ghost attempts have several important impacts within the model: they enable the computation of the predictor of the proportion of prior successes (with any choice of kernel), they introduce a sensible prior belief, and they reduce noise in the predictor.

Ghost attempts enable computation of the predictor because they resolve the issue of the proportion of recent success before a student's first attempt on a skill. As discussed above, on a first attempt, the proportion of recent success is $0/0$. For the intercept to be interpretable in a model with R_{ijt} as a predictor, the most sensible choice is to choose to set $R_{ij1} = 0$. This choice is equivalent to stipulating g unobserved attempts $X_{ij(1-g)}, \ldots, X_{ij0}$, all incorrect, so that $R_{ij1} = 0/g$.

Ghost attempts are one way to make explicit the assumption that at time 0, a student has not already learned the skill, which is very plausible in real-world learning datasets. (If a student is assigned to tutoring, it is likely that the student does not already know the material.) In this way, ghost attempts introduce a weakly informative prior belief into the R predictor, analogous to how Bayesian models incorporate prior beliefs.

Finally, ghost attempts reduce noise in the R predictor. Without ghost attempts, it is possible to stipulate that the proportion of prior successes on the first attempt $R_{ij1} = 0$. However, after just one observation, R_{ij2} may change dramatically, because it includes exactly one observation. By increasing the amount of information in R, even if that information comes from prior belief rather than data, ghost attempts cause R to grow more slowly. Noise in the R predictor would increase the standard error of both the intercept and the slope, and would drive the slope estimate closer to zero.

Figure 2 shows the effect of different values g of ghost attempts on two practice histories, $X_{ij} = (1, 0, 0, 0, 0)$, and $X_{i'j} = (0, 1, 1, 1, 1)$. Consider the summary of a student's practice history after the first attempt when no ghost attempts are included. Student i with a single success has a 100% success rate, and student i' with a single failure has a 0% success rate. Now consider the students' histories after attempt 5. Student i' now has four successes in a row, but they only have an recent success rate of 0.91, thus, they look less successful after 4 correct attempts than student i looked after a single correct attempt. This difference is *noise* in the proportion when the number of attempts is small. As we add ghost attempts, noise in the proportion of recent success is reduced.

As previously noted, the count of successes S_{ijt} implicitly includes an infinite number of such ghost attempts, but how many ghost attempts should be included in proportion of successes R? The value we choose ($g = 3$) is smaller than the length of the median practice history in the

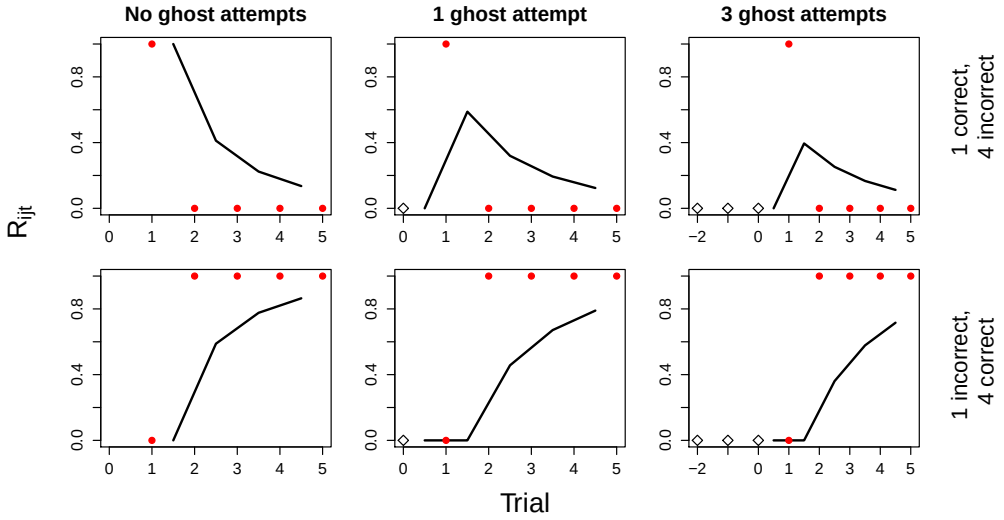

Figure 2: Effect of ghost attempts. Rows show two different patterns of student practice, while columns show different numbers of ghost attempts. Red circles indicate the student's practice history X_{ijt}, open black diamonds indicate the ghost attempts. R_{ijt} is computed using the exponential kernel with the optimal decay parameter $d = 0.7$. When no ghost attempts are included, it is impossible to calculate R_{ij1}.

dataset we study, so that the ghost attempts influence a relatively small fraction of the dataset. (Their influence is felt most keenly on the first attempt, and practically vanishes by the fourth.) It is an informed prior because we consider the dataset at hand, but we choose one value, so there is no overfit due to parameter search. It is reasonable to consider different values of g (different prior beliefs) for different applications, but note that g may affect the optimal bandwidth for a particular application.

4. Representations of History of Practice

All the models we examine are logistic regression models, with the general form

$$p_{ijt} = Pr(X_{ijt} = 1|Z = z) = \frac{\exp(z'\beta)}{1 + \exp(z'\beta)}. \tag{7}$$

The fixed structure of the logistic regression model allows us to compare the predictive utility of different representations of a student's prior practice. These terms, which replace the generic Z's in equation 7, are displayed for clear comparison in table 1.

In section 3., we defined the recent proportion of success R, and the same construction extends to a recent proportion of failure. Common, alternative representations of prior practice include the total length of practice T, the count of total successes S, and the count of total failures F. The total length of practice is the summary used in the Additive Factors Model (AFM) (Cen et al., 2006; Chi et al., 2011)

$$logit(p_{ijt}) = \theta_i + \beta_j + \gamma_j T_{ijt}. \tag{8}$$

Table 1: Terms in logistic model variants.

	Student ability	skill difficulty	Success count	Failure count	Total trials	Recent success rate
AFM	θ_i	β_j			$\gamma_j T_{ijt}$	
PFA	θ_i	β_j	$\alpha_j S_{ijt}$	$\rho_j F_{ijt}$		
S-only	θ_i	β_j	$\alpha_j S_{ijt}$			
R-only	θ_i	β_j				$\delta_j R_{ijt}$
R-AFM	θ_i	β_j			$\gamma_j T_{ijt}$	$\delta_j R_{ijt}$
R-PFA	θ_i	β_j		$\rho_j F_{ijt}$		$\delta_j R_{ijt}$

Performance Factors Analysis (PFA) (Pavlik et al., 2009; Chi et al., 2011) utilizes both the count of prior successes and prior failures:

$$logit(p_{ijt}) = \theta_i + \beta_j + \alpha_j S_{ijt} + \rho_j F_{ijt}. \tag{9}$$

The decayed count of successes defined in equation 6 is utilized by PFA-decay (Gong et al., 2011). Aside from the decay weight, PFA-decay uses the same predictors S and F as original PFA. In fact, when $d = 1$, PFA-decay and PFA are exactly the same. Thus, we refer to both these models that only include (possibly decayed) counts S and F as PFA.

To separate the effects of recent practice, total practice, and the differential predictive effects of recent success and failure, we compare three model variants that contain R:

$$\text{R-only} \quad logit(p_{ijt}) = \theta_i + \beta_j + \delta_j R_{ijt}, \tag{10}$$

$$\text{R-AFM} \quad logit(p_{ijt}) = \theta_i + \beta_j + \gamma_j T_{ijt} + \delta_j R_{ijt}, \tag{11}$$

$$\text{R-PFA} \quad logit(p_{ijt}) = \theta_i + \beta_j + \rho_j F_{ijt} + \delta_j R_{ijt}. \tag{12}$$

We compare these three recent-history models with the established AFM and PFA models, as well as an S-only baseline model that uses only the count of successes.[1] For PFA & R-PFA, which include two decay-weighted variables, we consider both the case where the tuning parameters are equal and the case where they are tuned separately. This allows for the potentially differential predictive power of recent successes vs. recent failures.

5. Model Application to Real-world Data

5.1. Methods

We evaluate the models described above in modeling student performance in the Assistments data used in the "moment of learning" work by Baker and colleagues (Baker et al., 2011). The data contain first attempts by 4,138 students on problem sets involving 54 skills, for a total of 187,309 first attempts (Table 2). Each problem is coded with a single skill. Each skill was

[1]The R-AFM model is what Galyardt and Goldin (2014) called R-PFA. We have adjusted the name to be more consistent with the names of other models. AFM uses total history, while PFA uses success and failures.

Table 2: Dimensionality of Assistments data.

	Full Data	CV Filter Rule	CV Data
Total Instances	187,309		33,431
Students	4,138		170
Skills	54		32
Overall % correct	65.7%		66.7%
Median students per skill	410	# of students > 10	51
Median skills per student	3	# of skills > 6	9
Median attempts per student per skill	4	# of attempts > 8	15
Maximum attempts per student per skill	141	First 39 attempts	39

attempted between 89 and 16,200 times, and had an overall percent correct between 23% and 95%. The data are from the mastery learning "Skill Builder" feature of Assistments, which allows teachers to set a threshold for the number of problems a student must correctly answer in a row to be considered proficient. For this data set, the threshold was set at either three or five.

This data set is sparse at the student level. First, the median number of skills seen by each student is three, 22% of students practiced only a single skill, and 75% of students practice 7 or fewer different skills. Second, the median number of attempts per skill by students is 4, and 435 students (11%) made 3 or fewer total problem-solving attempts. This sparsity of data at the student level means that any student effects in a model should be fit as random effects coming from a common distribution. In this way, we "pool" the data, so the student effects θ_i for students with less data shrink towards the mean student effect. The ghost attempts necessarily have the greatest influence on practice strings that are relatively short, i.e., they reduce the noise that would otherwise be present in R_{ijt} for these attempts.

For estimation purposes, a sufficient number of students practice each skill; of the 54 skills, only one is practiced by fewer than 25 students. However, since the standard error of a slope coefficient depends on the variance in its predictor, the small number of attempts per skill for the majority of students leads to higher uncertainty in the slope. For this reason, we also treat all skill intercepts and slopes as random effects. We did not include the covariance matrix for skill parameters in the model.

We used the `glmer` function in the R package `lme4` to fit all models listed in Table 1 (Bates et al., 2013). Counting the different kernels, and all of the different tunings of relevant decay weights, we fit a total of 242 models. Data[2] and analysis code[3] are posted online.

MODEL COMPARISON Due to the sparsity of the data, the Akaike information criterion (AIC) (Akaike, 1973; Akaike, 1985) is the best choice for a measure of model fit. AIC is a likelihood-based measure of model accuracy that incorporates a penalty for model complexity; and when the formula for the likelihood is correctly specified, AIC can easily accommodate hierarchical models. Minimizing AIC is equivalent to minimizing KL-divergence risk, and Stone (1977) proves that AIC is equivalent in general to cross-validation. Our simulation in section 6. reaffirms that AIC and cross-validation are equivalent for model selection in this application.

[2] https://sites.google.com/site/assistmentsdata/home/goldstein-baker-heffernan
[3] https://sites.google.com/site/aprilgalyardt/research

Student-stratified cross-validation is the commonly preferred method of model comparison within the educational data mining community: a model is trained on one set of students, and used to make predictions for a held-out set of students. In this way, one can claim to have a reasonable expectation of how well the model will perform on entirely new students. However, the sparsity at the student level, both in skills per student, and attempts per skill, makes student-stratified cross-validation on the full dataset unreliable, if not entirely untenable. Therefore, we also used 25-fold student-by-skill stratified cross-validated MAD scores (L_1 loss) on a reduced data set (Table 2) to compare AFM, PFA and R-PFA. We chose these because they use the three different representations of practice. For parameter estimates to be stable as segments of the data are left out, we only included students who practiced at least a minimal number of skills, and skills which were practiced by at least a minimal number of students. The thresholds we used are not severe, but because the full dataset is exceptionally sparse, these filters resulted in an 82% decrease in the size of the data. However, as discussed below, the CV-MAD model rankings on the reduced data set agree completely with the AIC rankings on the full data set.

5.2. Results and Discussion

We present results on the value of using only recent data rather than all observed data; on the relative utility of alternative data transformations; and on the optimal amount of recent data to retain. These results by necessity apply only to the Assistments data. The simulation in section 6. tests whether these results hold up in a variety of conditions. Work on other datasets provides evidence for the generalizability of these findings (Goldin and Galyardt, 2015b).

The Best-Performing Model The best overall model for predicting future success from a student's history is R-PFA with an exponential kernel and the optimal decay weight for successes is $d = 0.7$, and for failures is $d = 0.1$. For 49 of the 54 skills, the coefficients for the effect of recent successes are significantly positive. The remaining 5 skills have few associated attempts, very wide CI's, and are not significantly different than zero. In general, the more recent successes a student has had, the higher the probability of correctly responding to the next item, which corresponds to our intuitions about learning. It has been previously documented in PFA and AFM that the slopes for the effect of the count of past failures F_{ij} (and occasionally even for the count of past successes S_{ij}) are often negative, e.g., (Kaser et al., 2014). Such negative slopes signal an area of concern (with the performance model itself or with the skill decomposition), because more practice, successful or unsuccessful, should in general increase the probability of a correct response. The R-PFA result that recent success is predictive of future success counters the negative-slope phenomenon.

The Value of Recent History Figure 3 presents a heatmap AIC scores for all models. Note that any difference in AIC scores that is visible in the color gradient is a difference of a hundred or more, where an AIC difference of 3 is generally statistically significant.

The models containing any summary of recent success outperformed the models using total length of practice (AFM), and count of success and failures (PFA). CV-MAD scores for the reduced data set rank R-PFA (0.383), PFA (0.390), and AFM (0.395) in the same order. The fact that PFA outperforms AFM replicates prior research (Pavlik et al., 2009; Chi et al., 2011). The 1PL-IRT model, which lacks any summary of practice, is by far the worst performing model.

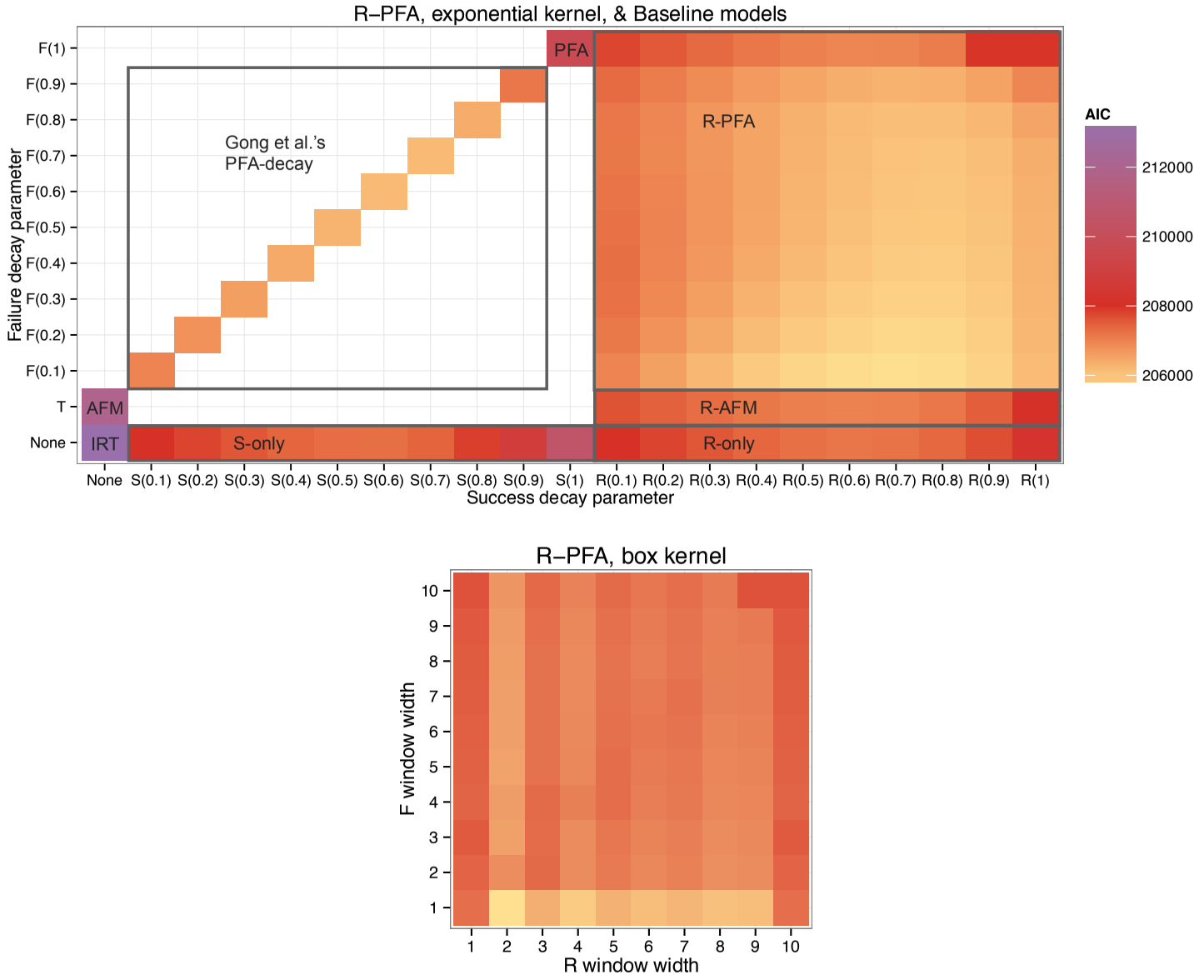

Figure 3: Each square on this heatmap shows the AIC score for one of the models compared on the Assistments dataset. Lower AIC is better. The models are arranged by the different representations of a student's history of success, and history of failure. On the top: Representations of success are on the x-axis, and arranged from none, to decayed counts of success (S), to recent proportions of success (R^{exp}). Representations of failure are on the y-axis, and arranged from none, to total amount of practice (T), to decayed counts of failure (F). On the bottom: The y-axis shows recent proportion of success (R^{box}), and the x-axis shows recent count of failure (F^{box}).

S-only with decay weight $d = 1$ outperforms AFM. With a decay parameter of 1, S_{ijt} is simply the total count of all prior successes for person i on skill j. Thus, a simple count of successes is a better predictor of future success than the total count of practice.

R-only with just the proportion of recent success, and no summary of failure, has better predictive performance than PFA. Even with a decay parameter of $d = 1.0$ which retains the entire history of practice, the proportion of success is a better predictor than the count of successes (S-only) or the separate counts of successes and failures (PFA). At the optimal bandwidth of $d = 0.6$ for R-only, most of the weight is on the last 3-5 attempts and the difference in AIC scores between R-only and PFA is 3000. R-PFA which treats the effects of recent success and recent failure separately, offers an even greater predictive advantage.

CHOICE OF KERNEL Consistent with the literature on nonparametric statistics, choice of kernel is less important than bandwidth, but smoother kernels are generally preferred (Wasserman, 2006). With the smooth exponential kernel, predictive performance degrades gracefully from optimal to sub-optimal bandwidths; with the non-smooth box kernel, performance does not degrade gracefully (Figure 3).

In addition, the normalized exponential kernel applied to the proportion of success R is a better predictor of student performance than the unnormalized exponential kernel that generates the decayed count of success S. At every fixed bandwidth d, R-only outperforms S-only.

OPTIMAL AMOUNT OF RECENT PRACTICE CONSIDERED The optimal tuning parameter for both S and R^{exp} is $d = 0.7$. This places 52% of the weight on the last 2 attempts, and 85% of weight on the last 5 attempts. For the box kernel, the optimal window for R^{box} is $K = 2$, so the effective size of the optimal bandwidth is similar for both kernels.[4]

In all cases, the smallest bandwidth investigated here for failures is found to be optimal, implying that only a failure on the single most recent attempt contains relevant information for predicting future performance, and prior failures are less informative. Tuning the success rate and the failure rate separately offers a distinct advantage. In the best-performing model, R acts like a running average over the last 2-5 actions, while F effectively indicates whether the last action was correct or incorrect.

The difference between the optimal tuning parameters for recent successes and recent failures may account for the difference in slips and guesses. If a student knows the skill and has been correctly responding, then $R \approx 1$ and $F \approx 0$. If this student then slips and responds incorrectly, with the optimal decay parameters, R will decrease to 0.7, and F jumps to 0.9. If the incorrect answer was truly a slip then the student will likely answer correctly on the next attempt, so that R increases towards 1 again, and F falls back toward zero. In this way, R is largely unaffected by slips, while F is an indicator that the last response may have been a slip. Now consider a student who does not know the skill, and has a history of incorrect practice attempts, so that $R \approx 0$ and $F \approx 1$. If this student then guesses correctly on an item, R only increases to 0.3, and F falls to 0.1. Here R is largely unaffected by the correct answer, while F is an indicator that the last response may have been a guess.

[4]For the box kernel, the AIC heatmap in figure 3 appears to have stripes indicating that even window widths are better than odd windows. This may be because even-width windows are functionally smoother than odd-width windows. Consider the history of practice $X_{ij} = \{0, 1, 0, 1, 0, 1, ...\}$. With an odd window size for the box kernel the recent proportion oscillates, e.g., K=3, $R_{ij} = \{..., \frac{1}{3}, \frac{2}{3}, \frac{1}{3}, \frac{2}{3}, ...\}$. If the window is even, whether $K = 2$ or 4, or 6, we get $R_{ij} = \{..., 0.5, 0.5, 0.5, 0.5, ...\}$.

5.3. UNDERSTANDING MODEL DIFFERENCES THROUGH VISUALIZATION

What is the source of the predictive advantage of a student's recent history of performance (R-PFA) compared to a student's entire history of performance (PFA)? Where do the predictions of the two models agree and where do they disagree? Given the sparsity in our dataset, it is inappropriate to examine errors in prediction on a held-out or cross-validation set; nonetheless it is instructive to compare predictive accuracy on the training set.

We present the difference in the predictions in figure 4, which we call the Viz-R visualization (Goldin and Galyardt, 2015a). To read the figure, consider that a conventional way to evaluate a single model is by using a confusion matrix that has four cells: true positive predictions, false positives, true negatives, and false negatives. In comparing two models, there are twice as many cells, because the models may agree or disagree. These 8 cells are spread over the the two rows of the figure, which correspond to actually incorrect (top) and actually correct (bottom) outcomes. Each row is divided into 4 facets according to the value of the R predictor:

- R in $[0, 0.3]$ indicates that the student has produced either 1 or fewer right answers in the last 4 attempts, or is at the very beginning of practice.
- R in $(0.3, 0.5]$ indicates 2 correct answers in the last 3-4 attempts.
- R in $(0.5, 0.7]$ implies that the most recent 2 answers were correct.
- R in $(0.7, 1]$ means that at least the last 3 answers were correct.

We use the standard labels (true positives, etc.) for cells where the models agree. With PFA as the baseline model and R-PFA as the comparison, "true positive wins" are where R-PFA gains true positive predictions over PFA, "false negative losses" are where R-PFA makes a false negative error that PFA does not commit, and so forth.

Viz-R is similar to a confusion matrix in that it shows the percentage of observations in each cell; but it also shows the distance between observations and the classification decision boundary for each model.[5] The X and Y axes indicate the predicted probabilities $\hat{p}$ from the PFA and R-PFA, respectively. This (x, y) position has a different meaning for the *actually correct* and *actually incorrect* outcomes. For example, the top-right quadrant for the *actually incorrect* outcomes indicates false positive values, and the top-right quadrant for the *actually correct* outcomes indicates true positives.

There is very little correlation between the predictions of PFA and R-PFA, but there are notable differences between the model predictions in two cases. First, when students have had more than a couple of attempts at a skill, but the student has had few recent successes, R-PFA is much better at predicting incorrect outcomes than PFA (top row, R in $[0, 0.3]$, TN Win, yellow-red color). This advantage is due directly to using recent history to make predictions.

To understand this, consider that the advantage is absent when there is little history (on the first or second attempt, R in $[0, 0.3]$, blue-green color): the only basis for a prediction then is

[5]The logistic models examined here all used a decision boundary of 0.5. Instances for which the predicted probability of a correct response was above the boundary ($\hat{p} > 0.5$) are treated as predicting a correct rather than an incorrect, and vice versa. The decision boundary has a margin of error; a model is not confident in a prediction that is close to the boundary. Thus, it is sensible to prefer models that not only make predictions on the right side of the decision boundary, but also close to the true class and far from the boundary. (In fact, this is reflected in metrics such as AIC, MAD, RMSE, and point-biserial correlation, but not in metrics that use 0-1 loss such as precision, recall, and AUC.)

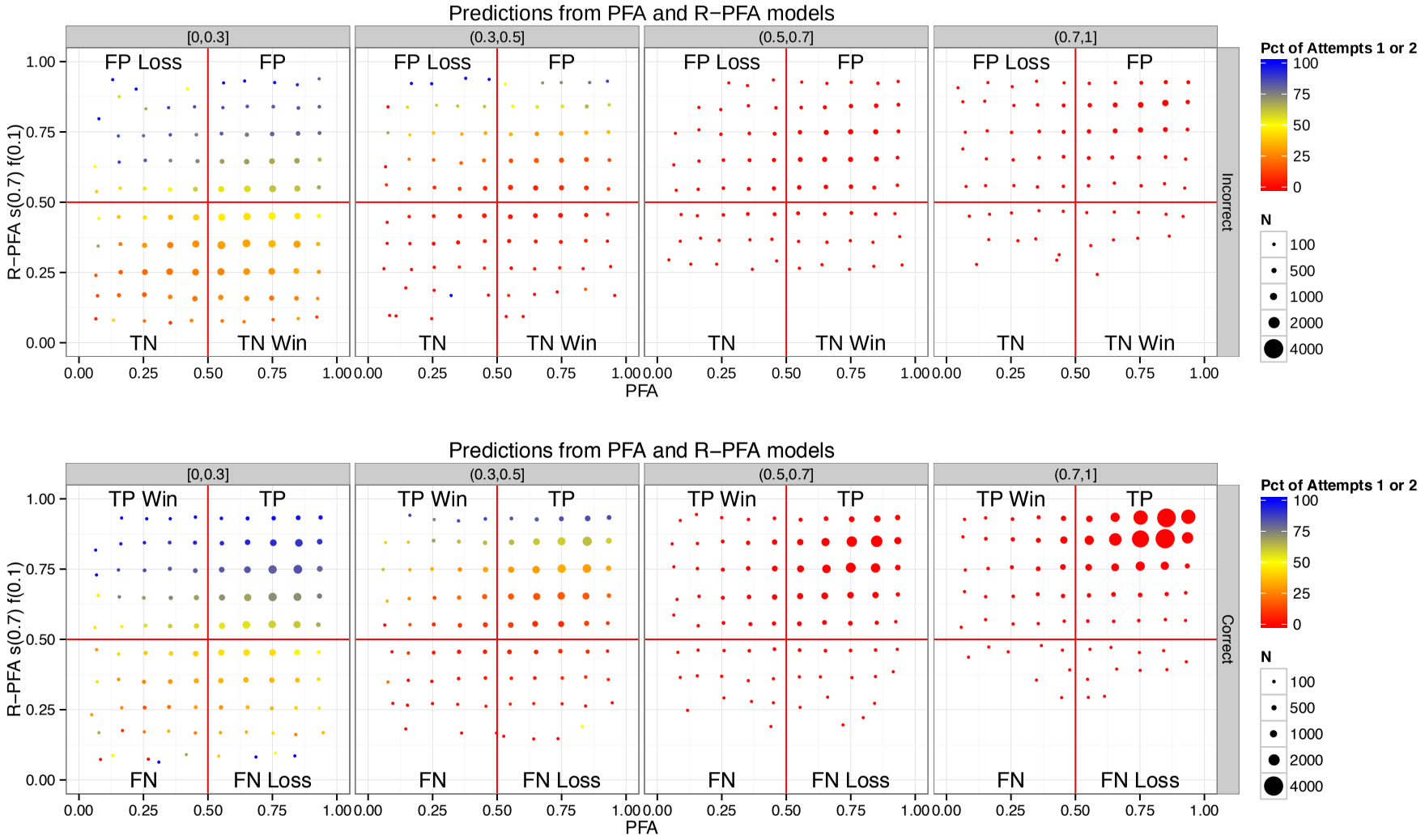

Figure 4: PFA vs. R-PFA predictions in terms of the $\hat{p}$ value from the logistic model. The bubbles display $\hat{p}$ values from the two logistic models; the more data points there are near that (x, y) position, the bigger the bubble. The color of the bubble indicates the percentage attempts in that position that are 1st or 2nd attempts by a student on a particular skill. The two rows of the figure correspond to actually incorrect (top) and actually correct (bottom) outcomes. Each row is divided into 4 facets according to the value of the R predictor. The closer $\hat{p}$ is to the boundary values of 1.0 or 0.0, the more confident the model is in the prediction. The abbreviations are: TP - true positive, FP - false positive, TN - true negative, FN - false negative. Win and loss are for R-PFA relative to PFA A $\hat{p} > 0.5$ is a prediction of a correct outcome, and $\hat{p} < 0.5$ is a prediction of an incorrect outcome.

the estimated difficulty of the skill, and both models generate a large number of false positives. Instead, R-PFA gains an advantage as the length of practice history increases. Consider three histories $X_{Aj} = \{0, 1, 0, 1, 0, 1\}$, $X_{Bj} = \{0, 0, 0, 1, 1, 1\}$, and $X_{Cj} = \{1, 1, 0, 1, 0, 0\}$. These three histories look equivalent to PFA on opportunity $t = 7$ since it only uses counts of successes and failures, but they look quite different to R-PFA; Student B has a fairly high proportion of recent success, while student A has a moderate value of R_{Aj7}, and R_{Cj7} is quite low. Thus, for students like student C, PFA will generate a false positive prediction while R-PFA generates a true negative prediction. This improvement in predicting when a student will fail to answer correctly is an important contribution of R-PFA for intelligent tutors and adaptive systems.

Second, when the student has had successes on the most recent items, R-PFA is more likely to predict a correct outcome than PFA. This is true both when the observed outcome is correct, and when it is not, i.e., when the observed incorrect outcome is likely a slip. Ultimately, the number of false positive losses for R-PFA (top row, R in $(0.5, 0.7]$ or $(0.7, 1.0]$) is much lower than the number of true positive wins (bottom row, same R). To an intelligent tutor, accurately predicting slips is arguably unimportant. An intelligent tutor using R-PFA rather than PFA would be more aggressive and more accurate at predicting student mastery of a skill, allowing students to graduate from practicing a skill more quickly than PFA.

When the student has had 2 correct answers in the last 3-4 attempts (R in $(0.3, 0.5]$), it is hard to know whether to expect a correct or an incorrect outcome. In the aggregate, PFA and R-PFA perform comparably in this case.

In sum, the exponentially decayed proportion of recent success outperforms separate success and failure counts because it is more accurate at predicting true negatives in early practice, and true positives for students with a history of successful practice.

6. Simulation Study

Simulations allow the exploration of model performance under a variety of controlled conditions. They address the question, "if student behavior has these features, how does model performance change?" Even though simulated data are not real, they are realistic. Thanks to simulations, we can understand more deeply how aspects of models and data affect model performance, and what performance we might expect on future datasets.

Simulation also removes some of the limitations of real data. Two aspects of the Assistments dataset examined above complicate model comparison. First, the sparsity in the Assistments data (section 5.) complicates the use of cross-validation for model comparison on the full dataset. In a simulation, sparsity can be controlled, so that we can compare model performance using both AIC and cross-validation. Second, the stopping rule used in the Skill Builder feature of Assistments leads to data missing non-randomly. Once Assistments determines that a student has mastered a skill, Assistments does not assign further practice on this skill. Because the stopping rule for the mastery criterion is often set at either 3 or 5 items in a row correct, this aspect of the data may impact the optimal bandwidth for recent performance. Simulation allows us to determine the optimal bandwidth when no stopping rule affects data generation.

We compare the predictive value of the recent proportion of success (R-PFA), a count of total successes and failures (PFA), and the total length of practice (AFM), across three different assumed features of student behavior. In simulation 1, we assume an idealized student behavior where each student learns at a consistent rate. In simulation 2, we assume that while students are practicing a skill, they may exhibit inconsistent performance, with relatively long sequences

of interwoven correct and incorrect performance. In simulation 3, we assume that some students may generate long sequences of incorrect responses on some skills (e.g., due to skill difficulty, or to a student gaming the system); we call these skills 'stumper skills'.

6.1. Simulation 1: Idealized Student Behavior

The Bayesian Knowledge Tracing (BKT) model (Corbett and Anderson, 1995) provides a generative model of idealized student behavior. BKT is a hidden Markov model with two states: a learned state where the student has a high probability of correctly responding to a question, and an unlearned state where the student has a low probability of correctly responding. At each practice attempt students transition from the unlearned state to the learned state with a fixed probability.

BKT does not represent complex detail of human cognition, but it is useful to treat it as an idealized, abstract model of observable student behavior. BKT is naive in multiple ways, among them that it treats all students as having the same ability and learning at the same rate; it does not consider that skills may be unexpectedly difficult for some students, or that students may attempt to game the system; it also assumes that students never forget a skill.

Simulating from BKT should favor PFA and AFM, since the probability that a student is in the learned state increases at a rate of $1-q_j^t$ with each attempt t. In this simple model, the length of practice should be a sufficient summary of student performance.

Methods We generated 100 datasets from this model, each with 50 skills and 3500 students, so that the total size of the data is near the size of the Assistments data. Each student practiced a random number of skills, generated by a Poisson distribution with a mean of 5. The number of opportunities for a student to practice each skill also varied randomly, generated by a Poisson distribution with a mean of 8. This means the number of opportunities for practice is statistically independent of the skill. This eliminates the uneven sparsity observed in the real data set. Full technical details for data generation are in appendix A1. Code for each of the simulations is posted online[6].

On each dataset, we fit seven of the logistic test models; AFM, using total practice length T_{ijt}; PFA, using undecayed counts of failures F_{ijt} and successes S_{ijt}; and R-PFA (exponential), at 5 different values of the decay weight for the weighted proportion of successes R_{ij}: $d = 0.2, 0.4, 0.6, 0.8, 1.0$. For the count of failures in R-PFA, we fixed the decay weight $d = 0.1$, because the smallest decay parameter for failures was always optimal for the Assistments data. As in the Assistments data, we used $g = 3$ ghost attempts. Student parameters were omitted from the models to allow for student-stratified cross-validation. Additionally, it was the uneven sparsity in the Assistments data that necessitated random effects for the skill parameters; randomness in the simulation means that there is sufficient data at each level to fit skill parameters as fixed effects. Because there is sufficient data at each level, the estimates for random effects from the R function `glmer` and fixed effects from the R function `glm`, used here, will be effectively the same. We compared the fits using AIC and 5-fold student-stratified cross-validation MAD (L_1 loss).

Results Across all 100 simulation replications, both AIC and CV-MAD ranked the 7 models in the same order: R-PFA at any bandwidth performed better than PFA, and AFM was ranked

[6] https://sites.google.com/site/aprilgalyardt/research

last. Among the R-PFA bandwidths, AIC ranked $d = 0.6$ as best in about 80% of the simulations and $d = 0.8$ as best in the other 20%. CV-MAD ranked $d = 0.6$ as best in 40% of the simulations with $d = 0.8$ best in about 60%. Note that these decay rates bracket the best-performing rate $d = 0.7$ observed on the Assistments data. This is the kind of behavior we expect when two models are similar. AIC and CV-MAD clearly agree that these are the best two decay parameters compared to the rest of the available models.

6.2. Simulation 2: Practicing

Simulation 1 represents idealized student behavior; in Simulation 2, we add an aspect of realistic student behavior. When students are learning a new skill, they often exhibit some length of inconsistent performance where correct and incorrect attempts are interwoven, and practice histories are similar to sequences such as $X_{ij} = (0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1)$. The classic BKT model does not adequately capture this phenomenon. With only the two states of unlearned and learned, BKT tends to generate practice sequences where students suddenly transition from mostly incorrect to mostly correct practice, e.g., $X_{ij} = (0, 0, 0, 0, 1, 1, 1, 1)$.

Methods To simulate gradual transition from incorrect to correct practice, we created a 3-state BKT model, adding an intermediate state to BKT. Students in the unlearned state have a low probability of answering correctly. In the new 'practicing' state, students are learning the skill, but their understanding is not complete, so they have only moderate probabilities of a correct response. Students in the 'fluent' state have a very high probability of answering correctly. We assume that no student begins practice in the fluent state, so that practice will benefit all students. Details are in appendix A2.. This model is more realistic than the idealized student behavior generated by the classic 2-state BKT model, but it should still favor AFM and PFA, since the probability that a student will reach the fluent state increases at a consistent rate across each practice opportunity t.

As in Simulation 1 (section 6.1.), we generate 100 datasets, fit the same 7 models, and compare them using AIC and 5-fold student-stratified CV-MAD scores.

Results Across all 100 simulation replications, both AIC and CV-MAD ranked the 7 models in the same order: R-PFA at any bandwidth performed better than PFA, and AFM was ranked last. In 100% of the simulations, both AIC and CV-MAD ranked $d = 0.6$ as the best bandwidth. The bandwidth $d = 0.4$ was second best in about 60% of the simulations, with $d = 0.8$ falling in third place.

6.3. Simulation 3: Stumper Skills

In the first two simulations, all students progress towards mastery at a steady rate (that depends on individual skills). But real students sometimes encounter skills that for a variety of reasons, they are not able to master and leave them "stumped". Students may lack a pre-requisite skill, and legitimately struggle and fail to master the target skill; or they may become de-motivated and engage in hint abuse (Aleven and Koedinger 2000) or in gaming the system (Baker et al. 2004). Such a student will generate a practice history that is a sequence of predominantly incorrect practice. The presence of such sequences can dramatically affect model parameters since the behavior of the stumped students does not match the behavior of the other students, and may affect the relative utility of the different summaries of prior student practice under consideration.

METHODS To simulate this behavior, we adapted a BKT model. A randomly generated segment of students (about 8%) will occasionally get stumped. The probability that one of these students gets stumped on a particular skill depends on the skill. When a student gets stumped, they have a low probability of correctly responding to any item on that skill. When a student is not stumped, their practice history is generated according to the classic 2-state BKT model. Full technical details are in appendix A3. Once again, we generate 100 datasets, fit the same 7 models, and compare them using AIC and 5-fold student-stratified CV-MAD scores.

RESULTS As in the previous simulations, in 100% of simulated datasets both AIC and CV-MAD ranked R-PFA at any bandwidth above PFA, and AFM was the worst performing model. AIC and CV-MAD agree that the best two bandwidths are $d = 0.6$ and $d = 0.8$; AIC ranks $d = 0.6$ as better in 60% of the simulations, while CV-MAD ranks $d = 0.8$ as better in 60% of the simulations.

6.4. SIMULATION DISCUSSION

In all three simulation scenarios, recent history is a better predictor of future success than a complete history of successes and failures, or the total length of practice. This was true even in the first simulation, which should have favored AFM and PFA. In a BKT model with no forgetting, the more opportunities that a student has to practice, the more likely it is that a student will transition from the unlearned state to the learned state. Therefore, on average, the total number of opportunities to practice should be proportional to the log-odds of a correct response. Yet even in this case, recent history makes better predictions.

The scenarios that students may go through periods of "practicing" (Sec 6.2.) or encounter "stumper skills" (Sec 6.3.) reflect real-world phenomena by construction. R-PFA is robust to this for the same reason that it is robust to interleaved skill practice, namely that it disaggregates learning curves, in effect allowing different students different amounts of time to begin to demonstrate successful practice (Goldin and Galyardt, 2015b).

In all 3 simulation models, the optimal decay parameter is in the range [0.6, 0.8]. If the number of ghost attempts changes from $g = 3$, this may change slightly; but since by the student's third attempt the ghost attempts are largely forgotten, adding additional ghost attempts will make little difference. With these decay rates 75-93% of the weight is on the last 5 attempts, and 55-78% of the weight is on the last 3 attempts. Thus, the last 3-5 practice opportunities contain sufficient information to judge whether or not a student has learned the skill.

7. CONCLUSIONS

The primary contributions of this work are:

- a definition of several representations of recent practice
- a thorough comparison of these representations in terms of predictive accuracy on real-world and simulated data, which demonstrates how a student's recent performance history evidences whether or not they have acquired a particular knowledge component
- the novel Recent-Performance Factors Analysis model that embeds the most effective representation (the kernel-transformed proportion of recent successes and the decayed count of prior failures)

- the publicly available implementation of the recency representations and R-PFA model

ON RECENCY We proposed that recent history ought to be a better predictor of student performance than the entire history of practice. We validated this insight by embedding various representations of recent history in the well-established Linear Logistic Test Model framework, and by leveraging prior work: the separation of student and item characteristics (IRT), the grouping of items by skill and the significance of past performance (AFM), the separation of prior successful and unsuccessful practice (PFA), and discounting of older evidence by Gong et al.

We found that, first, a decay-weighted proportion of successes is a better predictor than a decay-weighted count of successes. Second, decay weights should be tuned, rather than determined heuristically (as in the work by Gong et al.). Third, decay weights for successes and failures should be tuned separately. Fourth, it is sensible and effective to inform the model with a prior "belief" that students who have never attempted the skill will likely fail to answer correctly, e.g., using ghost attempts. In aggregate, these insights lead to improvements in predictive accuracy in the true negative rate when recent history contains few correct attempts, and in the true positive rate when recent history mostly consists of correct attempts.

The optimal amount of recent history for modeling is consistent across all of the simulations, and the Assistments data; the best decay parameter for recent successes is consistently $d = \{0.6, 0.7, 0.8\}$. Weights in this range place almost all of the weight on the last 3-5 attemtps. Thus, empirically, these last 3-5 attempts contain sufficient information about the student's knowledge state to make accurate predictions. Interestingly, this decay rate supports the heuristic, implemented in some adaptive learning systems (Heffernan and Heffernan, 2014), that a student has mastered a skill if a student has a "streak" of 3-5 correct responses in a row on the skill. However, because the exponentially decayed kernel performs better than a box kernel, we can conclude that the streak model can be easily improved.

The exponential kernel is relatively simple mathematically, adding only three tuning weights beyond the parameter structure of PFA (the number of ghost attempts, and decay rates for failures and successes). The stability of the decay weights identified in this work implies these weights might be reasonably treated as starting values in new uses of R-PFA, further reducing complexity of R-PFA (Goldin and Galyardt, 2015b). Importantly, the proportion of successes not only has higher predictive accuracy than a count predictor, but it also has clearer interpretation, because it avoids the scaling issues associated with the unlimited count predictor.

The findings here cast doubt on the validity of the AFM model, because its treatment of total practice had the lowest predictive accuracy of all the other logistic model variants, including even the non-decayed count of successes only, i.e., S-only with $d = 1$. At present, AFM has uses aside from prediction, including in skill model selection in Learning Factors Analysis (Cen et al., 2006), which may need to adopt different models.

We evaluated models on both real-world and simulated data. Simulated data allowed us to test out representations of recency under conditions where we could control different kinds of noise, and to ensure that our ranking of recency representations generalized beyond the Assistments dataset. Although simulated data evaluations are rare in the educational data mining literature, they are very popular in statistics. In fact, we argue that real-world datasets have sparse data properties that necessitate both kinds of comparisons.

CURRENT LIMITATIONS AND FUTURE WORK This investigation of recency leveraged the logistic regression framework. In the future, we will consider the relationship of R-PFA to

Dynamic Bayesian Networks such as Knowledge Tracing. Graphical models also reflect change over time, but in a less transparent way than a simple proportion of recent success. Preliminary work suggests that the two kinds of models reveal interesting aspects in each other.

In this study, the decay rate for recent history was held constant over all skills, and the optimal decay rate was consistently in the same window [0.6, 0.8]. It is possible that the optimal bandwidth is different for different skills, perhaps we need a smaller bandwidth for harder skills where students have very little chance of guessing an answer correctly. Future work may investigate this potential.

We will consider how R-PFA may incorporate richer Q-matrices (multiple skills per item), and we have begun to look at how R-PFA may be used to improve cognitive models (Goldin and Galyardt, 2015b).

A CONJECTURE One interpretation of the recency investigation is that a handful of the most recent observations are a better summary of the learner's mastery of a skill than the student's entire history of practice. Why might that be? One explanation is that because data are noisy, by saving all older data, we retain too much noise. Another explanation is that in a system where change (e.g., learning) can happen, older data may be not merely noisy, but erroneous.

Just as a human tutor can make accurate inferences about tutee mastery on the basis of watching a student solve a problem or two, ideally, a machine ought to be able to do the same.

We conjecture that the more information can be extracted from the most recent practice attempts, the fewer attempts are necessary to make valid and accurate inference about student knowledge. Beyond correct and incorrect outcomes, some information that we might extract from an attempt includes the time taken to solve the problem, the interaction of the student with the problem, student affect and engagement, use of hints and feedback, likely and apparent misconceptions, and scratch work.

A SIMULATION DETAILS

A1. SIMULATION 1: IDEALIZED STUDENT BEHAVIOR

Simulated data is generated according to the usual BKT model, it is a hidden Markov model with an unlearned and a learned state.

- Knowledge components are indexed $j = 1, \ldots, K$
- Students indexed $i = 1, \ldots, N$
- Student i's response on the t^{th} opportunity to practice skill j:

$$X_{ijt} = \begin{cases} 0 & \text{if incorrect} \\ 1 & \text{if correct} \end{cases}$$

- Denote student i's unobserved knowledge of skill j on the t^{th} opportunity as

$$Z_{ijt} = \begin{cases} 1 & \text{if unlearned state} \\ 2 & \text{if learned state} \end{cases}$$

- Probability of initially knowing skill j: $Pr(Z_{ij1} = 2) = \pi_j \sim Beta(1, 2)$.

  This distribution has positive probability for all values on the interval [0,1], but is centered at a mean of $\mathbb{E}[\pi_j] = \frac{1}{3}$. This encapsulates our expectation that in a well-targeted educational intervention, most of the students would not already know the majority of topics which will be taught. The density is shown in figure 5.

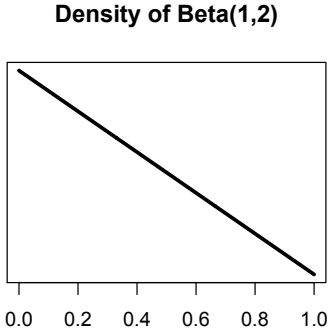

Figure 5: Density of $Beta(1, 2)$ distribution.

- Transition matrices for the Markov process are

$$P_j = \begin{pmatrix} 1 - L_j & L_j \\ 0 & 1 \end{pmatrix}$$

  L_j is the probability of learning skill k following a practice attempt, generated according to $L_j \sim Beta(2, 2)$.

  This distribution positive probability for all values on the interval [0,1], but is centered at $\mathbb{E}[L_k] = \frac{1}{2}$. If L_k is near 1, then a student has a high probability of learning the skill after a single practice attempt. In the same way if L_k is near 0, then a student has a low probability of learning the skill, regardless of how much they practice. This Beta distribution places more probability near 0.5, and lower probability near 0 or 1, reflecting the idea that most students need to practice skills a couple of times before they learn them. The density is shown in figure 6.

- Probability of a correct answer in the unlearned state (guessing): $C_{uj} \sim Unif(0.02, 0.3)$

$$Pr(X_{ijt} = 1|Z_{ijt} = 1) \; = \; C_{uj}$$

- Probability of a correct answer in the learned state (1-slip): $C_{lj} \sim Unif(0.7, 0.98)$

$$Pr(X_{ijt} = 1|Z_{ijt} = 2) \; = \; C_{lj}$$

- Average number of skills seen by each student is fixed at `K.n = 5`
- Number of skills seen by student i is generated $J_i \sim \min\{K, Poisson(K.n)\}$.

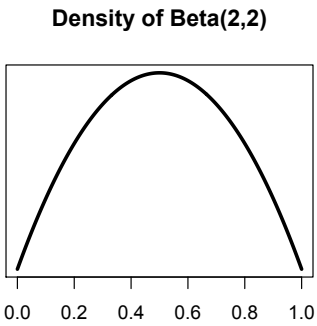

Figure 6: Density of $Beta(2, 2)$ distribution.

- The skills that student i answers are drawn without replacement from $\{1, \ldots, K\}$.
- `T.avg = 8` is the average number of practice opportunities for any student on any skill.
- The number of practice opportunities for student i on skill k is $O_{ij} \sim \max\{Poisson(T.avg), 2\}$. So that if a student practiced a skill, they practiced it at least twice.

A2. Simulation 2: Practicing Student Behavior

The 3-state BKT model uses the states unlearned, practicing, and fluent. Students in the unlearned state have a low probability of answering correctly. Students in the practicing state have moderate probabilities of answering correctly. We may think of students in this state as largely understanding the ideas and knowing what to do, but slipping frequently perhaps due to high working memory loads or other causes. Students in the fluent state have a very high probability of answering correctly.

- Student i's response on the t^{th} opportunity to practice skill j:

$$X_{ijt} = \begin{cases} 0 & \text{if incorrect} \\ 1 & \text{if correct} \end{cases}$$

- Denote student i's unobserved knowledge of skill j on the t^{th} opportunity as

$$Z_{ijt} = \begin{cases} 1 & \text{if unlearned state} \\ 2 & \text{if practicing state} \\ 3 & \text{if fluent state} \end{cases}$$

- Probability for initial states: $\pi_j = (\pi_{j1}, \pi_{j2}, \pi_{j3})$.

$$\begin{aligned} P(Z_{ij0} = 1) &= \pi_{j1} &&\sim Beta(2, 2) \\ P(Z_{ij0} = 2) &= \pi_{j2} &&= 1 - \pi_{j1} \\ P(Z_{ij0} = 3) &= \pi_{j3} &&= 0 \end{aligned}$$

This distribution for π_j assumes that no student begins practice in the fluent state, so that practice will benefit all students. The $Beta(2,2)$ distribution is shown in figure 6. π_{j1} can take any value between 0 an 1, but it is more likely to take values nearer to 0.5. This simulates the idea that for an average skill approximately half the students will start out not knowing the skill at all, and the other half of the students need more practice.

- Transition matrices for the Markov process are

$$L_j = \begin{bmatrix} L_{j11} & 1 - L_{j11} & 0 \\ 0 & L_{j22} & 1 - L_{j22} \\ 0 & 0 & 1 \end{bmatrix}$$

where

$$L_{j11}, L_{j22} \sim Beta(2,2).$$

With these transition matrices, a student may not transition directly from the unlearned to the fluent state over a single opportunity. However, since this is a 1st order Markov process, it is possible and fairly likely that some students will transition from unlearned to fluent within 2 practice opportunities.

- Probability of a correct answer in the unlearned state (guessing): $C_{uj} \sim Unif(0.02, 0.2)$

$$Pr(X_{ijt} = 1 | Z_{ijt} = 1) \; = \; C_{uj}$$

- Probability of a correct answer in the practicing state: $C_{pj} \sim Unif(0.4, 0.7)$

$$Pr(X_{ijt} = 1 | Z_{ijt} = 2) \; = \; C_{pj}$$

- Probability of a correct answer in the fluent state (1-slip): $C_{fj} \sim Unif(0.85, 1)$

$$Pr(X_{ijt} = 1 | Z_{ijt} = 3) \; = \; C_{fj}$$

- Average number of skills seen by each student is fixed at `K.n = 5`
- Number of skills seen by student i is generated $J_i \sim \min(K, Poisson(K.n))$.
- The skills that student i answers are drawn without replacement from $\{1, \ldots, K\}$.
- `T.avg = 8` is the average number of practice opportunities for any student on any skill.
- The number of practice opportunities for student i on skill k is $O_{ij} \sim \max(Poisson(T.avg), 2)$. So that if a student practiced a skill, they practiced it at least twice.

A3. SIMULATION 3: STUMPER SKILLS

The second adaptation to the familiar 2-state BKT model includes a small proportion of students who occasionally engage in unproductive learning behavior, which produces long strings of incorrect responses, or *failure sequences*. This behavior might appear for many different reasons, such as the student engaging in hint-abuse or other gaming behaviors, or the student may simply lack a key prerequisite skill. As a shorthand, we shall refer to students who engage in this behavior as stumped students.

On each skill, the stumped students will have a probability of engaging in the stumped behavior for that skill. The probability that these students will engage in the behaviors depends on the skill, not the student. *Whether* a student ever engages in the stumped behavior depends on the student. *When* a student does so depends on the skill.

When a student does engage in stumped behavior, their responses will be a string of primarily incorrect responses with high probability. When a student is not engaging in stumped behavior, data is generated according to an unmodified two-state BKT model.

- To simulate the stumped behavior:
  - For each student draw the indicator G_i for whether student i engages in the stumped behavior, $G_i \sim Bernoulli(0.08)$.
  - For each skill j, draw a probability that one of the stumped students will engage in this behavior on this skill. $B_j \sim Uniform(0, 1)$.
  - Draw an indicator for whether student i will engage in this behavior on skill j

$$\begin{aligned} W_{ij}|G_i = 1 &\sim Bernoulli(B_j) \\ W_{ij}|G_i = 0 &= 0 \end{aligned}$$

  - If $W_{ij} = 0$, then generate X_{ij} from the 2-state BKT model (exactly as in appendix A1.).
  - If $W_{ij} = 1$, then for $t = 1, \ldots, T_{ij}$, $\quad X_{ijt}|W_{ij} = 1 \sim Bernoulli(0.2)$.
- Average number of skills seen by each student is fixed at `K.n = 5`
- Number of skills seen by student i is generated $J_i \sim \min(K, Poisson(K.n))$.
- The skills that student i answers are drawn without replacement from $\{1, \ldots, K\}$.
- `T.avg = 8` is the average number of practice opportunities for any student on any skill.
- The number of practice opportunities for student i on skill k is $O_{ij} \sim \max(Poisson(T.avg), 2)$. So that if a student practiced a skill, they practiced it at least twice.

ACKNOWLEDGEMENTS

We thank Neil Heffernan, Ryan Baker, and Yutao Wang for providing the Assistments data set.

This study was supported in part by resources from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgia's Office of the Vice President for Research and Office of the Vice President for Information Technology.

REFERENCES

ADAMS, R. J., WILSON, M., AND WANG, W.-C. 1997. The multidimensional random coefficients multinomial logit model. *Applied psychological measurement 21,* 1, 1–23.

AKAIKE, H. 1973. Information theory and an extension of the maximum likelihood principle. In *Proceedings of Second International Symposium on Information Theory*, B. Petrov and F. Caski, Eds. Akademiai Kiado, Budapest, 267–281.

AKAIKE, H. 1985. Prediction and entropy. In *A Celebration of Statistics*, A. Atkinson and S. Fienberg, Eds. Springer: New York, 1–24.

ALEVEN, V. AND KOEDINGER, K. R. 2000. Limitations of student control: Do students know when they need help? In *Intelligent Tutoring Systems*, G. Gauthier, C. Frasson, and K. VanLehn, Eds. Vol. 1839. Springer Berlin Heidelberg, Berlin, Heidelberg, 292–303.

BAKER, R. S., CORBETT, A. T., KOEDINGER, K. R., AND WAGNER, A. Z. 2004. Off-task behavior in the cognitive tutor classroom: when students game the system. In *Proceedings of SIGCHI conference on Human factors in computing systems*. ACM, 383–390.

BAKER, R. S., GOLDSTEIN, A. B., AND HEFFERNAN, N. T. 2011. Detecting learning moment-by-moment. In *IJAIED*. Vol. 21. 5–25.

BATES, D., MAECHLER, M., BOLKER, B., AND WALKER, S. 2013. lme4: Linear mixed-effects models using eigen and s4. Computer Program.

BECK, J. E. AND CHANG, K.-M. 2007. Identifiability: A fundamental problem of student modeling. In *User Modeling 2007*, C. Conati, K. McCoy, and G. Paliouras, Eds. Number 4511 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, 137–146.

BECK, J. E., CHANG, K.-M., MOSTOW, J., AND CORBETT, A. 2008. Does help help? Introducing the Bayesian evaluation and assessment methodology. In *Intelligent Tutoring Systems*, B. P. Woolf, E. Aïmeur, R. Nkambou, and S. Lajoie, Eds. Vol. 5091. Springer, Berlin, 383–394.

BOOTS, B., SIDDIQI, S. M., AND GORDON, G. J. 2011. Closing the learning-planning loop with predictive state representations. *The International Journal of Robotics Research 30,* 7 (June), 954–966.

CEN, H., KOEDINGER, K., AND JUNKER, B. 2006. Learning Factors Analysis – a general method for cognitive model evaluation and improvement. In *Proc of 8th ITS Conf*, M. Ikeda, K. D. Ashley, and T.-W. Chan, Eds. Vol. 4053. Springer Berlin Heidelberg, Berlin, Heidelberg, 164–175.

CHI, M., KOEDINGER, K., GORDON, G., JORDAN, P., AND VANLEHN, K. 2011. Instructional factors analysis: A cognitive model for multiple instructional interventions. In *Proceedings of 4th International Conference on Educational Data Mining*.

CORBETT, A. T. AND ANDERSON, J. R. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction 4,* 4, 253–278.

DE BOECK, P. AND WILSON, M., Eds. 2004. *Explanatory item response models: a generalized linear and nonlinear approach.* Springer, New York.

FALAKMASIR, M. H., PARDOS, Z. A., GORDON, G. J., AND BRUSILOVSKY, P. 2013. A spectral learning approach to knowledge tracing. In *Proceedings of 6th International Conference on Educational Data Mining.*, S. K. D'Mello, R. A. Calvo, and A. Olney, Eds. Memphis, TN, 28–34.

FALAKMASIR, M. H., YUDELSON, M., RITTER, S., AND KOEDINGER, K. 2015. Spectral bayesian knowledge tracing. In *Proceedings of the 8th International Conference on Educational Data Mining.*, O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, and M. Desmarais, Eds. Madrid, Spain, 360–364.

FISCHER, G. H. 1973. The linear logistic test model as an instrument in educational research. *Acta Psychologica 37*, 359–374.

GALYARDT, A. AND GOLDIN, I. 2014. Recent-Performance Factors Analysis. In *Proceedings of 7th International Conference on Educational Data Mining*, J. Stamper, Z. Pardos, M. Mavrikis, and B. McLaren, Eds. 411–412. (Poster paper).

GOLDIN, I. AND GALYARDT, A. 2015a. Viz-R: Using Recency to Improve Student and Domain Models. In *Proceedings of Second (2015) ACM Conference on Learning @ Scale*. L@S '15. ACM, New York, NY, USA, 417–420.

GOLDIN, I. M. AND GALYARDT, A. 2015b. Convergent validity of a student model: Recent-Performance Factors Analysis. In *Proceedings of 8th International Conference on Educational Data Mining*. Madrid, Spain.

GOLDIN, I. M., KOEDINGER, K. R., AND ALEVEN, V. A. W. M. M. 2012. Learner Differences in Hint Processing. In *Proceedings of 5th International Conference on Educational Data Mining*, K. Yacef, O. Zaïane, A. Hershkovitz, M. Yudelson, and J. Stamper, Eds. Chania, Greece, 73–80.

GONG, Y., BECK, J. E., AND HEFFERNAN, N. T. 2011. How to construct more accurate student models: Comparing and optimizing knowledge tracing and performance factor analysis. *International Journal of Artificial Intelligence in Education 21,* 1, 27–46.

GONZÁLEZ-BRENES, J., HUANG, Y., AND BRUSILOVSKY, P. 2014. General features in knowledge tracing: Applications to multiple subskills, temporal item response theory, and expert knowledge. In *Proceedings of 7th International Conference on Educational Data Mining*, J. Stamper, Z. Pardos, M. Mavrikis, and B. McLaren, Eds. London, England, 84–91.

HEFFERNAN, N. T. AND HEFFERNAN, C. L. 2014. The ASSISTments ecosystem. *IJAIE 24,* 4 (Dec.), 470–497.

JUNKER, B. W. 2011. Modeling hierarchy and dependence among task responses in educational data mining. In *Handbook of Educational Data Mining*. Chapman & Hall/CRC.

KASER, T., KOEDINGER, K., AND GROSS, M. 2014. Different parameters-same prediction: An analysis of learning curves. In *Proceedings of 7th International Conference on Educational Data Mining*. London, UK.

KHAJAH, M. M., HUANG, Y., GONZÁLEZ-BRENES, J., MOZER, M. C., AND BRUSILOVSKY, P. 2014. Integrating knowledge traching and item response theory: A tale of two frameworks. In *Fourth International Workshop on Personalization Approaches in Learning Environments (PALE 2014)*, M. Kravcik, O. C. Santos, and J. G. Boticario, Eds. 7–15.

KHAJAH, M. M., WING, R. M., LINDSEY, R. V., AND MOZER, M. C. 2014. Integrating latent-factor and knowledge-tracing models to predict individual differences in learning. In *Proceedings of 7th International Conference on Educational Data Mining*, J. Stamper, Z. Pardos, M. Mavrikis, and B. McLaren, Eds. London, England.

PAVLIK, P. I., CEN, H., AND KOEDINGER, K. 2009. Performance Factors Analysis–a new alternative to Knowledge Tracing. In *Proceedings of 14th International Conference on Artificial Intelligence in Education*. IOS Press, 531–538.

PAVLIK, P. I., YUDELSON, M., AND KOEDINGER, K. R. 2015. A measurement model of microgenetic transfer for improving instructional outcomes. *International Journal of Artificial Intelligence in Education*, 1–34.

RUPP, A., TEMPLIN, J., AND HENSON, R. 2010. *Diagnostic Measurement: Theory, Methods and Applications*. Guilford Press, New York, NY.

STONE, M. 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological) 39,* 1, 44–47.

WASSERMAN, L. 2006. *All of Nonparametric Statistics*. Springer, New York, NY.

YUDELSON, M., HOSSEINI, R., VIHAVAINEN, A., AND BRUSILOVSKY, P. 2014. Investigating automated student modeling in a Java MOOC. In *Proceedings of 7th International Conference on Educational Data Mining*. London, UK.