Forecasting Future Student Mastery

Stefan Slater[a]* and Ryan Baker[a]

[a]*Graduate School of Education, University of Pennsylvania, Philadelphia, United States of America*

3700 Walnut St., Philadelphia PA, 19103; slater.research@gmail.com

Biography (Slater): Stefan Slater is a PhD student at the University of Pennsylvania Graduate School of Education, and a researcher at the Penn Center for Learning Analytics (PCLA). He studies knowledge inference and knowledge estimation methods, mixture models, and natural language processing.

Biography (Baker): Ryan Baker is an Associate Professor at the University of Pennsylvania Graduate School of Education, and the Director of the Penn Center for Learning Analytics (PCLA). His research interests include the automated detection of engagement, affect, and self-regulated learning strategies, as well as the use of automated detectors in longitudinal research on student success.

Forecasting Future Student Mastery

Considerable attention has been given to methods for knowledge estimation, a category of methods for automatic assessment of a student's degree of skill mastery or knowledge at a specific time. Knowledge estimation is frequently used to make decisions about when a student has reached mastery and is ready to advance to new material, but there has been little work to forecast how far a student is from mastery or predict how much more practice the student will need before he or she will reach mastery. This paper presents a method for predicting the point at which a student will reach skill mastery within an adaptive learning system, based on current approaches to estimating student knowledge. We apply this technique to two popular methods of modeling student learning – Bayesian Knowledge Tracing and Performance Factors Analysis – and compare prediction correctness. Potential applications and future steps for improving the method are discussed.

Keywords: Bayesian knowledge tracing, performance factors analysis, prediction, knowledge inference, forecasting

Introduction

The inference of student understanding of skills and content knowledge in adaptive learning systems is a major area of study within the fields of educational data mining and student modeling (Corbett & Anderson, 1995; Pavlik, Cen & Koedinger, 2009; Pardos & Heffernan, 2010; Pelánek, 2014). Within both distance learning and blended learning applications of adaptive learning, models of student knowledge are used both to assess and monitor the progress of student learning as well as to identify attributes of the knowledge domain itself. Automated assessment of student learning can aid in the intelligent assignment of problems and content to students (e.g. Murray & Arroyo, 2002) and ensure that students are not continuing to practice already mastered content (Corbett & Anderson, 1995; Koedinger et al., 2010), increasing the flexibility of the system to provide differentiated learning experiences. Recent years have seen these

technologies, originally developed in formal educational settings, moving into open distance learning as well. For example, Pardos and his colleagues (2013) developed a student modeling algorithm for an edX MOOC on introductory electrical engineering, and Rushkin and his colleagues (2017) used automated assessment of student knowledge to drive adaptive learning within a HarvardX MOOC on astrobiology. Automated assessment and adaptation is particularly powerful in open and flexible learning contexts, where there are too many students for an instructor to monitor each student's progress individually, and there is utility to having high-quality information on student knowledge quickly available.

Having learning support structures in place is vital to success in distance education contexts, as many distance education learners are part-time adult students who may not be as sure of their capabilities and levels of knowledge mastery as their younger peers (Kumari, 2018). Research by Thistoll & Yates (2017), for example, showed that support and monitoring of student learning, as well as the promotion of motivation and engagement in the learning process, are important success factors in whether students complete distance education courses. The ability to forecast a student's future performance and knowledge in a learning activity affords the ability for the system, or for an instructor, to act upon that information and ensure that distance education learners are receiving adequate support and scaffolding in their studies. Additionally, it gives learners agency over their learning process, and the ability to see their learning trajectories.

The function of a knowledge model is to infer how well a student knows each skill in the learning content, by looking at the pattern of correct and incorrect responses that a student gives (Koedinger, Corbett, & Perfetti, 2012). Models adopting this approach have been found to have good accuracy and reliability in modeling student

knowledge in a range of applications (Corbett & Anderson, 1995; Baker et al., 2010; Pardos & Heffernan, 2010; Sao Pedro, Baker & Gobert, 2013), including in open distance learning contexts (as discussed above; e.g. Pardos et al., 2013; Rushkin et al., 2017). These models are especially well-suited to adaptive learning systems and distance education contexts, as instructor intervention is not always feasible in these environments. Thus far, however, these models have only been able to identify student mastery as it happens or after it has already occurred.

While assessing a student's current level of mastery is useful, it may also be useful to assess their likely future knowledge. Forecasting has gained prominence in other domains such as weather and financial markets, but in education predictive analyses of future outcomes have typically focused either on immediate future performance (Pavlik, Cen, & Koedinger, 2009; Pardos & Heffernan, 2010) or on relatively coarse-grained and distal outcomes, such as who will drop out of a massive open online course (Tang & Xing, 2018) or what constitutes successful outcomes of course-taking (Henderikx, Kreijns, & Kalz, 2017). Students working in an adaptive learning environment, whether in a distance or blended context, may be interested in knowing how their learning is progressing on a shorter timescale, and how much further they have to go until they reach mastery. From a system design standpoint, forecasting student mastery can allow for more adaptive and flexible problem assignment. It may also be possible to use mastery forecasts to determine how much more content needs to be created within an adaptive system where students are not yet reaching mastery. Research by Moilenburg & Berge (2005) highlights time/interruptions, support services, and motivation as obstacles to student success in distance education contexts, and providing students access to real-time performance feedback that can tailor the

scaffolding and content that they receive could alleviate several of these obstacles to learning.

While some models have attempted to infer how long a student will retain factual content, towards optimizing the spacing of practice (e.g. Khajah, Lindsey, & Mozer, 2014; Pavlik et al., 2008), these models do not provide guidance regarding how close a student is to reaching mastery. In particular, both instructors and adaptive systems may find it useful to know when a student is ready to move to a different concept or lesson and whether the student potentially needs additional scaffolding and tutoring. This may be particularly relevant in distance contexts, where much of the information available to instructors in person is no longer available or where learners require additional learning supports and measures of their own competencies.

In this paper, we will examine whether student knowledge modeling can be adjusted to not only assess current knowledge, but forecast when a student will reach mastery, towards better personalizing instruction in distance contexts.

Background: Knowledge Modeling Algorithms

The question of when a student is likely to reach a future mastery state could be answered by using a modified knowledge tracing framework that not only tracks student mastery in real time, but also provides a prospective forecast for when a student is likely to reach some threshold of mastery. One of the few areas where this type of analysis has been seen in fine-grained educational data is in attempts to predict "wheel-spinning" (Beck & Gong, 2013). Wheel-spinning students are characterized by a failure to master a KC in a timely manner, operationalized by Beck and Gong as ten opportunities. Determining whether or not a student will reach mastery by a specific point (or ever

reach mastery) is a special case of the broader problem of forecasting future performance.

Within the fields of educational data mining and learning analytics, two models that are frequently employed to assess student mastery in running systems used by students are classical Bayesian Knowledge Tracing (BKT), developed by Corbett & Anderson (1995), and Performance Factors Analysis (PFA), developed by Pavlik, Cen & Koedinger (2009). These two models make distinct assumptions but have largely performed equivalently in attempts to predict immediate future student correctness (i.e. on the next problem). However, BKT has been more frequently extended and used in scientific analyses.

In classical BKT, student knowledge of a given KC is assumed to be either unknown or known, and the likelihood of a student being in either state is inferred via their pattern of correct and incorrect answers. This likelihood is inferred through four parameters, which are typically fit for each skill. $p(L_0)$, initial knowledge, is the probability that a knowledge component was learned before the first opportunity to answer a problem associated with that KC. $p(T)$, the learning parameter, is the probability that a KC is learned between a student's current opportunity to answer and the next opportunity. $p(G)$, guess, is the probability that a correct answer is given despite a student being in the unlearned state, and $p(S)$, slip, is the probability that a student in the learned state does not answer the question correctly. For each first attempt that a student makes at a new problem, a set of equations based on Bayes Theorem is used both to calculate the probability that they will answer that question correctly, and to update the probability that the student knows the skill.

In PFA, student mastery is expressed as the probability that a student will produce the correct answer to a problem, without the separate step of computing a

probability that the student knows the skill. PFA models student performance using a logistic regression equation with two variables – the cumulative number of correct and incorrect answers that a student has produced thus far for the current skill. It also uses three parameters, typically fit for each skill - γ, the degree to which a correct answer is associated with better future performance; ρ, the degree to which an *incorrect* answer is associated with better future performance; and β, the overall ease or difficulty of the KC (or KCs) to be learned. Unlike BKT, where parameters are probabilities and vary from 0 to 1, PFA's parameters are logits – they are unbounded, and can vary from negative to positive infinity. However, despite the lack of bounding on parameters, they still have relative magnitude – a skill with a γ of 0.6 and a ρ of 0.3 means that a correct answer is twice as likely to indicate mastery than an incorrect answer. In PFA, unlike BKT, the same parameter can represent both learning and performance. This is seen in the meaning of a negative parameter, which represents evidence *against* mastery. A skill with a negative ρ coefficient is one where wrong answers provide stronger evidence that the student does not know the skill than the typical improvement in performance seen from gaining more practice.

In this paper, we apply the extension of forecasting to both BKT and PFA. We conducted this research using classical BKT instead of one of its many variants (e.g. Baker et al., 2008; Pardos & Heffernan, 2010; Gonzalez-Brenes, Huang, & Brusilovsky, 2014). There is no a priori reason to prefer alternative versions of BKT for this application, so using the original formulation provides a 'base case' or point of reference that future models and forecasting techniques can be compared to in order to assess changes in accuracy, precision, or predictive ability.

Methods

Data used for this study were obtained from 22,225 students who used the

ASSISTments system during the 2012-13 school year (the publicly available 2012-2013 School Data with Affect Data Set -- https://sites.google.com/site/assistmentsdata/home/2012-13-school-data-with-affect). The use of ASSISTments has increased over time, with over 50,000 students a year now using ASSISTments for their mathematics learning; its growth in users demonstrates that schools are increasingly choosing this free and flexible learning platform to assist students. ASSISTments is an online platform that provides formative *assessment* and student support and *assistance*, used both in blended classroom situations and as homework. While ASSISTments is not generally thought of as an open flexible distance education platform, it does share some similarities with systems typically referred to in this fashion. Students frequently work on problems and assignments in the system from home, and teachers can upload and modify problem sets within the system remotely. ASSISTments is entirely free to use for teachers, formal students, and other learners. While the ASSISTments platform lacks the peer messaging and socializing structures that many OFDL systems contain, students and teachers can message each other within the system regarding problem content and pedagogy. The data set we analyzed consisted of problems and skills used in 'skill builder' problem sets in ASSISTments. In a skill builder, students complete a set of problems involving the same skill and can only advance when they get three consecutive correct answers, or another threshold chosen by teachers. This provides struggling students additional practice, while allowing students who demonstrate early mastery to move quickly to new content.

Our data consisted of 179,908 unique problems across 478 identified skills. In our forecasting of student knowledge mastery, we grouped these data into unique student-skill pairs, one string of records for each unique student's first attempt on each problem within each unique skill. Pairs where a student failed to reach mastery were not

included in the analysis, as they did not provide information on how much practice was needed for that student to reach mastery on that skill. The data originally contained 264,202 student-skill pairs consisting of 1,315,121 individual problem responses. After removing outlier skills (described in the following section) the data was reduced to 193,048 student-skill pairs consisting of 1,054,818 responses to problems in ASSISTments.

BKT parameters were calculated using an exhaustive grid search (BKT-BF, Baker et al., 2010) with constraints to guess and slip as in Corbett & Anderson (1995). PFA parameters were calculated using the Nelder-Mead optimization algorithm within the SciPy library for Python 2.7 (Nelder & Mead, 1965; Jones et al., 2001). PFA and BKT models were optimized to minimize the root mean squared error (RMSE).

Removal of Outlier Skills

After fitting BKT and PFA parameters to the data, each skill was examined for parameter values that could potentially indicate poorly-defined, outlier, or otherwise problematic skills. Descriptives for all parameter values can be found in Table 1. Table 2 shows how many skills were excluded due to each of the criteria listed below.

In the case of BKT, because we wanted to examine the ability to predict when students reach mastery, we excluded all skills which suggested that students had either reached mastery before beginning to practice the skill, or for which acquisition was very fast and mastery was virtually always obtained within the first three attempts. For BKT, skills with a $p(L_0)$ above 0.90 were excluded, as well as skills with a $p(T)$ above 0.60. PFA parameters, for the most part, had less extreme values, and outlier removal was based on outliers in the overall distribution. Skills with a γ parameter that fell outside of the range 0 – 2 were excluded, skills with a ρ parameter outside the range -2 – 0 were excluded, and skills with a β parameter below -2 were excluded. Functionally speaking,

this process excluded skills for which students were likely to be familiar with the material already, and where additional learning was unlikely.

Finally, for both BKT and PFA, we removed all skills for which fewer than 100 students practiced the skill, as there was not enough data within these skills to construct an accurate knowledge model. This left a total of 313 skills for analysis.

Parameter	*Mean*	*SD*	*Min*	*Max*	*Skewness*
$p(L_0)$	0.702	0.162	0.186	0.894	-1.358
$p(T)$	0.063	0.051	0.001	0.402	2.811
γ	0.938	0.332	0.239	1.952	0.738
ρ	-0.679	0.286	-1.760	-0.174	-1.235
β	-0.366	0.471	-1.929	1.872	-0.449

Table 1. Descriptive statistics for BKT and PFA parameters across all skills (after removal of outliers).

Criterion	*# of skills excluded*
BKT	
$p(L_0) > 0.90$	62
$p(T) > 0.60$	2
PFA	
$0 < \gamma < 2$	26
$-2 < \rho < 0$	5
$-2 < \beta$	22
Low *n*	162

Table 2. Skills excluded from analysis (not mutually exclusive – some skills were eliminated based on multiple criteria).

Data Processing

This work attempted to identify how quickly future student mastery could be predicted – a forecast of when the student was likely to reach mastery based on their current learning trajectory. For each algorithm, we determined from the data when that algorithm thought the student reached mastery. From there, we created multiple sets of data consisting of the first *n* attempts recorded by each student, and then applied BKT and PFA models to these sets of attempts, computing a predicted point of mastery based on the first *n* attempts, and comparing that forecast to the actual point of mastery in the full dataset. We used different sizes of *n* for this process, from $n = 1$ (only using the student's first opportunity to practice the skill) to $n = 15$ (the first 15 opportunities, for students who had at least 16 opportunities' worth of data). We chose $n = 15$ based on examining the data and determining that most problem sets within ASSISTments consisted of 15 or fewer problems. At each step, students who reached mastery before

the *n*th attempt were removed, as there was no future point of mastery to predict. Sample sizes for the full range of lists used in the analysis are presented in Table 3. This process was conducted for BKT and PFA separately, as the two methods produced different estimates of when mastery was reached.

An additional consideration had to be made in formatting the data for the PFA model. Depending on the learning rates estimated for a specific skill by the algorithm and the student's current performance estimate when forecasting occurs, a student may be forecast to never reach mastery. There were 35,742 total student-skill pairs in our data (18.5% of cases) where students were forecasted to never reach mastery by the model, despite having evidence of having done so in the data. These cases were dropped from the analysis.

For comparability, we performed similar filtering on BKT for extreme cases. Unlike PFA, BKT monotonically approaches 1 over time so long as the student achieves some level of correctness, as a consequence of modeling some errors as slips. Therefore, we could not set a strict "never reaches mastery" condition as we did with PFA. Instead, we identified the longest sequence of problems solved by a student for any skill (e.g. a student-skill pair) which resulted in BKT inferring mastery, which was 93 problems long, and used this value as an upper limit for the forecasting procedure. Any cases where mastery was predicted to require more than 93 problems solved were dropped from the analysis. This method removed 23,863 total student-skill pairs from the BKT set of data (12% of cases), a proportion similar to the proportion removed from the data for PFA.

Forecasting Procedure

The following forecasting procedure was used for BKT: at the n^{th} student attempt in a transaction, a probability of mastery, $p(L_n)$, was calculated, using BKT. For all

subsequent attempts, this probability of mastery at the previous problem was used to compute the correctness for the subsequent problem by using the standard equation for calculating *p*(correct) in BKT (Corbett & Anderson, 1995).

Classical BKT does not accommodate partial correctness, instead using separate equations for the calculation of $p(L_n)$ after either a correct or incorrect answer. Because the forecasting process does not produce a binary correct or incorrect value, we needed to adjust the information that was returned to the BKT model. One option for using this information would be to threshold this prediction at 0.5 and treat it as a fully correct answer. However, doing so would discard considerable information, treating 0.51 as identical to 0.99 and as fundamentally different from the much closer probability value of 0.49. In order to account for partial correctness, we followed the approach utilized in Sao Pedro et al. (2013) for integrating partially correct answers into BKT. We repeat this recursively, such that at each step forward in the forecasting process, we estimate the probability of correctness, use that estimate to re-estimate student knowledge, in turn re-calculate the probability of correctness, and continue until the student is assessed to have reached mastery. We can then compare this prediction to the assessment of mastery within the evaluation list.

For PFA the process was simpler. PFA does not model latent knowledge and then predict correctness adjusted for guess and slip; it simply predicts correctness directly. As such, *p*(m) – the knowledge estimate in PFA – is equivalent to *p*(correct). Therefore, when forecasting beyond problem *n,* for each subsequent $(n + 1)^{th}$ attempt, the most recent *p*(m) was added to the student's cumulative successful answers, and (1 – *p*(m)) was added to the student's cumulative unsuccessful answers. These modified successes and failures are then used recursively to calculate a new *p*(m) for each new

attempt in the forecasting model. Like BKT, this process continues until $p(\text{m})$ exceeded 0.95.

While the criterion for mastery in the ASSISTments system is a sequence of three correct answers, we chose instead to use the more standard (and model-relevant) criterion of $p(L_n)$ or $p(\text{m})$ exceeding 0.95. This choice also had the benefit of avoiding having to decide how to map $p(\text{correct})$ to a binary success or failure value. The point at which the student reached mastery according to the forecasting model was then recorded and compared to the actual point of mastery (according to PFA) in the evaluation list.

Inference size	n_{BKT}	n_{PFA}	*Inference size*	n_{BKT}	n_{PFA}
$n = 1$	56,958	83,546	$n = 9$	6,263	6,592
$n = 2$	37,946	42,323	$n = 10$	5,047	5,328
$n = 3$	31,406	37,676	$n = 11$	3,984	4,157
$n = 4$	24,465	26,590	$n = 12$	3,104	3,251
$n = 5$	17,576	19,270	$n = 13$	2,534	2,622
$n = 6$	13,968	15,049	$n = 14$	1,985	2,080
$n = 7$	10,307	10,883	$n = 15$	1,619	1,678
$n = 8$	8,191	8,641			

Table 3. Sample sizes for BKT and PFA, by number of attempts.

Results

We can compute the goodness of our forecasting by comparing our forecasted time of

reaching mastery (according to the forecasting algorithm) to the actual time the student reached mastery (according to the original algorithm – BKT or PFA), with the number of problems completed so far used as the measure of time. In doing this comparison, we compute the root mean squared error (RMSE; the square root of the mean squared error, which is the average difference between the prediction and the true value, squared), averaging across all student-skill pairs considered, for each set of *n* attempts. Model results are presented in Table 4.

On average, forecasting based on BKT performed quite poorly, only forecasting student mastery to within 24 attempts of the student's actual point of mastery in the data, with a range of 23 to 29 attempts across values of *n*. Forecasting based on PFA performed substantially better, with an average root mean squared error (RMSE) of 2.37 attempts, and a range of 2.00 to 2.57 attempts across values of *n*. Additionally, forecasting based on PFA performed consistently well across different values of *n*, and was able to predict when a student would master a skill to within two problems as early as after two opportunities to practice the skill.

	RMSE BKT	*RMSE PFA*		*RMSE BKT*	*RMSE PFA*
$n = 1$	26.12	2.54	$n = 9$	23.65	2.43
$n = 2$	28.63	**2.00**	$n = 10$	23.26	2.31
$n = 3$	25.01	2.21	$n = 11$	23.53	2.51
$n = 4$	24.26	2.28	$n = 12$	**22.95**	2.39
$n = 5$	24.48	2.19	$n = 13$	24.21	2.59
$n = 6$	23.22	2.42	$n = 14$	22.97	2.47
$n = 7$	23.65	2.32	$n = 15$	23.72	2.57
$n = 8$	23.32	2.33	*Average*	24.20	2.37

Table 4. RMSE of forecasting BKT and PFA for *n* attempts. The best performing set is bolded for each model.

We also evaluated model performance for skills that had different values for their model parameters. Results for BKT are presented in Table 5, and results for PFA are presented in Table 6. Within these tables we subdivide skills based on their learning and initial knowledge parameters into three groups: "low", "medium", and "high". For both BKT and PFA, we chose cutoffs between low/medium/high that appeared to follow existing divisions in the distribution of parameters across the model. The distribution of parameters for BKT was heavily skewed, while the distribution of parameters for PFA was more normally distributed. In both cases, we select sub-sets of one parameter while including all values for the other parameter.

For BKT, "low" initial knowledge ($p(L_0)$) skills were defined as skills which had a $p(L_0) < 0.45$, "medium" initial knowledge skills had a $p(L_0)$ between 0.45 and 0.60, and "high" initial knowledge skills had a $p(L_0)$ at or above 0.60. For the learning

parameter *p*(T), “low” learning skills were defined as having a *p*(T) below 0.07, “medium” learning skills had a *p*(T) between 0.07 and 0.10, and “high” learning skills had a *p*(T) at or above 0.10.

For PFA, learning can be conceptualized in terms of the γ and ρ parameters – there is not a single indicator of learning, rather, the γ parameter describes the evidence that a correct answer provides about mastery, and the ρ parameter describes the evidence that an incorrect answer provides about mastery. We examined the differential performance of the model across both these parameters. “Low” learning from correct answers was defined as skills with a γ below 0.60, “medium” skills had a γ between 0.60 and 1.26, and “high” skills had a γ at or above 1.26. “Low” learning from incorrect answers was defined as skills with a ρ below -1.0, “medium” skills had a ρ between -1.0 and -0.50, and “high” skills had a ρ at or above -0.50. Finally, for skill difficulty (represented by the β parameter), low difficulty skills were skills with a β greater than zero, medium difficulty skills were skills with a β between -0.75 and 0, and high difficulty skills were skills with a β less than -0.75.

	Low $p(L_0)$	Med $p(L_0)$	Hi $p(L_0)$	Low $p(T)$	Med $p(T)$	Hi $p(T)$
$n = 1$	47.38	25.79	28.88	37.44	15.88	9.12
$n = 2$	45.79	25.47	28.74	37.54	14.60	9.74
$n = 3$	37.00	22.42	27.38	34.90	12.67	8.35
$n = 4$	42.48	25.11	25.80	33.68	13.88	9.23
$n = 5$	37.59	25.11	24.79	31.67	13.50	9.16
$n = 6$	33.69	19.95	24.18	30.56	13.07	8.26
$n = 7$	31.90	19.21	22.74	30.37	12.00	8.46
$n = 8$	29.83	15.68	24.18	30.30	12.29	***7.95***
$n = 9$	27.10	***14.16***	24.81	30.98	13.93	9.81
$n = 10$	25.48	***14.16***	21.98	27.40	13.05	8.96
$n = 11$	***24.48***	21.16	20.48	26.64	13.31	9.30
$n = 12$	26.28	19.32	23.20	28.25	13.90	13.54
$n = 13$	27.10	17.42	21.97	26.64	***11.83***	14.53
$n = 14$	26.14	21.79	22.58	26.61	14.74	14.58
$n = 15$	28.55	17.11	***20.29***	***24.36***	15.88	12.88
Average	32.72	20.26	24.13	30.49	13.63	12.88
Skill n	29	19	203	212	42	57

Table 5. RMSE of BKT results by low, medium, and high values of $p(L_0)$ and $p(T)$, with the best-performing set for each group ***emphasized***.

Forecast accuracy for BKT was better when $p(T)$ was high and/or $p(L_0)$ had moderate values. but tended to be poorer when the number of attempts used was either small or large. We hypothesize that the poorer performance for large numbers of attempts was

due to the relatively large variation in problems left for the small proportion of remaining students (all of who had struggled and some of whom continued to struggle for dozens of problems). The overall predictive power remained low, still only arriving within 8 attempts of the test set even under ideal circumstances. This suggests that despite BKT's success in making immediate inferences and despite the accuracy of those inferences for predicting measures such as post-test scores (Corbett & Anderson, 1995; Baker et al., 2010) and standardized test scores (Pardos et al., 2014), BKT is not well-suited to the forecasting application tested in this paper.

	L. γ	M. γ	H. γ	L. ρ	M. ρ	H. ρ	L. β	M. β	H. β
n = 1	4.20	3.12	2.32	2.45	2.80	4.13	2.92	3.18	3.10
n = 2	3.09	2.02	1.80	1.83	1.87	2.73	1.69	2.21	2.19
n = 3	2.86	2.25	2.16	2.24	2.16	2.56	2.08	2.21	2.81
n = 4	2.57	2.25	1.98	1.86	2.09	2.67	2.22	2.29	2.10
n = 5	***1.77***	1.94	1.77	1.81	1.92	1.90	1.98	***1.81***	2.05
n = 6	2.17	2.10	2.16	***1.67***	2.16	2.23	2.33	2.08	2.03
n = 7	2.34	2.03	2.09	2.10	1.98	2.27	2.08	2.04	2.21
n = 8	2.14	2.02	1.80	1.79	2.03	2.00	1.94	2.07	1.81
n = 9	2.57	***1.90***	1.68	2.19	1.93	1.95	1.88	1.87	2.33
n = 10	2.40	1.97	2.14	1.98	2.20	***1.78***	1.92	2.13	1.88
n = 11	2.80	2.10	1.89	1.79	2.17	2.27	1.76	2.37	1.81
n = 12	1.89	1.98	2.14	1.88	2.09	***1.78***	2.37	1.96	***1.78***
n = 13	***1.77***	2.22	1.86	2.02	2.31	1.79	1.92	2.07	2.40
n = 14	1.83	1.95	***1.59***	1.98	***1.84***	2.06	***1.65***	1.88	2.09
n = 15	2.11	2.19	2.00	2.83	2.07	2.15	1.84	2.09	2.62
Average	2.43	2.14	1.96	2.03	2.11	2.28	2.04	2.15	2.21
Skill n	35	210	45	43	185	87	36	219	61

Table 6. RMSE of PFA results by low, medium, and high values of γ, ρ, and β, with the best-performing set for each group ***emphasized***.

Again in contrast to BKT's forecasting accuracy, PFA is able to predict point of student mastery to within about two attempts on average. Additionally, PFA appears to be consistent across a range of parameter values and is effective even with relatively limited data.

It should be mentioned that both techniques consistently over-predicted the amount of additional practice necessary to reach mastery – the error range of predictions for BKT and PFA are both in the positive direction (i.e. no cases of under-prediction). This is a consequence of using partial correctness as input to the model. A number of skills are not learned gradually, but rather through transformative "aha moments" (Baker, Goldstein & Heffernan, 2011). The gradual increases in prediction that our forecasting models use are not able to predict these sudden shifts in understanding, and have a bias towards over-practice. While over-practice has historically been a concern for educational data mining researchers (Cen, Koedinger & Junker, 2007), we believe that in the case of PFA a two-problem over-prediction of necessary practice opportunities is relatively minor.

Discussion

In this paper we developed a method for predicting the point at which a student in an online learning platform was likely to reach mastery of a given knowledge component, and tested the accuracy of this forecasting method when applied to two thoroughly-researched knowledge estimation techniques: classical Bayesian Knowledge Tracing and Performance Factors Analysis. We determined that this is a viable technique when using PFA, with the model being able to predict the number of opportunities that a student requires to reach mastery to within two opportunities, even as soon as upon completion of the student's second or third problem. BKT, however, performed poorly, and it does not seem like this method is feasible for use with this algorithm in its current articulation.

One hypothesis for the poor performance of forecasting based on BKT is the way in which it models learning. BKT uses its $p(T)$ parameter to model the acquisition of KCs, and by doing so attempts to model learning as occurring in each of the

individual opportunities a student has to learn the skill. However, work by Baker, Goldstein & Heffernan (2011) has shown that learning is not always a smooth and gradual process. Instead, learning can be 'lumpy', and students occasionally make great leaps of understanding on particular problems. In other words, if a student gets a skill consistently wrong on the first several attempts, our forecasting approach within BKT has no way to infer that the student may eventually "get it" and reach mastery.

PFA was largely resistant to this issue. Although it is equally unable to model leaps in understanding as BKT, it models learning as occurring in the aggregate, based on how much prior practice there has been. As such, it is more able to identify that a student's past failures do not permanently prevent success. However, PFA was a faulty model in some ways as well. Forecasting with PFA was occasionally unable to identify students who would eventually go on to master a skill. Instances where this happened looked similar to those that BKT struggled with – students who provided early consecutive incorrect answers. In cases where average student learning is slow, initial poor performance may prevent a student from ever reaching mastery, according to the PFA model. In this case, these students were necessarily eliminated from consideration, a potential bias in favor of PFA (but note that a comparable portion of students were omitted from analysis of BKT, for similar reasons). Future work on mastery forecasting may consider finding a method for identifying these cases where prediction is difficult, and handling them using some alternate method. On the other hand, it may not be desirable to assume t2hat all students will learn – the research on wheel-spinning indicates that some students do not learn specific skills from an online learning system quickly, or indeed at all (Beck & Gong, 2013). Forecasting could be a valuable technique for quickly identifying students likely to wheel-spin and finding them additional scaffolding or support for their learning.

In general, the capacity of this forecasting technique to predict future student learning could be of interest to the adult, part-time learners that make up the bulk of distance education platforms. Forecasting affords learners the ability to consider their current knowledge state, and when the system believes they will reach mastery. Given the adaptive nature of learning supports in open flexible distance learning systems, these forecasted predictions create the potential to extend the knowledge – and ultimately the agency -- that learners have over their own learning processes.

For instance, while PFA was less vulnerable to the difficulties of modeling lumpy learning than BKT, both models were not fully able to account for more rapid shifts in student performance, especially in cases where a student struggles early but goes on to drastically improve their performance, thanks potentially to a eureka or "aha" moment (Moore, Baker, & Gowda, 2015). Both models weight data at all time points equally, and have trouble adjusting to sudden shifts of understanding that students may have. Combining this forecasting technique with models of moment-by-moment student learning (Baker, Goldstein, & Heffernan, 2011) could improve overall performance and allow learners to see the impact of their work in the system.

It is worth noting that classic BKT's lack of success does not mean that all knowledge modeling algorithms and approaches based on Bayesian Knowledge Tracing will be ineffective. Within this paper, we used Corbett & Anderson (1995)'s classical BKT model, due to its wide adoption within the field of educational data mining and learning analytics, as well as its relative computational simplicity. However, there are several variants of BKT, and some of these expanded models may be able to better forecast skill mastery. One potentially useful approach is seen in in Falakmasir, Yudelson, Ritter & Koedinger (2015), which breaks down student transactions into overlapping triads of responses. Each potential triad is then coded as indicative or not

indicative of mastery (i.e., {0,1,1} would be indicative of mastery, while {0,1,0} would not). This algorithm may be especially relevant for use in forecasting due to the structure of consecutive student attempts as *n*-grams. It is possible that this approach could potentially minimize the difficulty that our models had with multiple early incorrect attempts. More generally, it is worth analyzing whether other model frameworks can enhance the predictive accuracy of our forecasting method, and whether alternative methods for adding forecasting to a student model may prove more accurate.

Applications for Knowledge Forecasting

The forecasting technique outlined here may prove useful for a range of implementations. For students using an online learning system who have not yet reached mastery on a certain skill, forecasting can be used to determine an approximate amount of additional practice that the student may need. For learning environments which do not actively intervene based on student knowledge inferences, forecasting could be used in external analyses by curriculum developers or instructors creating assignments, to determine the ideal length for assignments or activities and ensure that students are not spending too much or too little time practicing a given skill.

Another use of the forecasts provided here is for teachers and instructors. Teachers' roles as technology experts and instructional designers are increasing in distance education contexts, yet teachers do not see themselves as technology experts (Roberts, 2018) and often their perceptions of student knowledge are inaccurate (Mavrikis, Holmes, Zhang, & Ma, 2018). Learning analytics provide key information and an opportunity for teachers to decide whether to intervene to support a student (e.g. Miller et al., 2015) – a mastery forecast may be a more useful tool than a simple estimate of current knowledge. By determining which students are furthest from mastering the skill,

a teacher can allocate their time where it has the best chance of improving outcomes. Even in open distance contexts, making high-quality forecasting readily available may help instructors and community teaching assistants in identifying groups of students who are wheel-spinning and need alternate forms of assistance. Equally importantly, by predicting future success relatively early in the learning process, forecasting may be able to identify students at risk of wheel-spinning faster than methods which rely on evidence the student is already wheel-spinning (e.g. Beck & Gong, 2013). In open distance learning contexts with hundreds or thousands of learners, this approach can help focus facilitator attention and resources to students who may be most in need. Without support, a student who is not making progress in an open distance learning context is likely to drop out of the course (Tang & Xing, 2018); if we can identify this struggle before it goes on for too long, we may be able to address the problem and help students get back on track.

One potential issue involving the validity of these results is the relative age of the ASSISTments dataset used. The intelligent tutoring system landscape continues to shift and evolve, and is different than it was six years ago. It may be that these forecasting techniques rely on features of the ASSISTments platform, and its users, that are less common today. This data set was chosen for its public availability (allowing replication and comparison of our findings) and its extensive use in prior published work by a substantial number of research groups (indicating data quality and usability). As this work is replicated and examined on more recent datasets, we may find that the predictive power of forecasting technique improves or degrades for these two algorithms, depending on what manner of scaffolds, behaviors, and knowledge models are being employed in the context of study. An area of future work would be to validate these results across multiple different learning platforms and timeframes, and ensure

that this forecasting process is robust and generalizable to multiple distance learning contexts.

Overall, by forecasting how much practice a student is likely to still need, we can provide information and guidance on where more support is needed or where educational resources are best spent – helping the mission of adaptive learning to provide each student with the support that he or she needs. We believe that the addition of these forecasting techniques to the learning scaffolds currently employed by open flexible distance education platforms has the potential to give learners greater understanding of the progression of their learning, and greater agency to understand and improve their current knowledge state.

Acknowledgements

This research was supported by the National Science Foundation (NSF) (DRL 1252297 and DRL-1535340) and the Penn Center for Learning Analytics. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF. We also would like to thank Neil and Cristina Heffernan and the rest of the ASSISTments team for providing us with the data analyzed in this paper.

References

Baker, R.S.J.d., Corbett, A.T., Aleven, V. (2008) More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, 406-415.

Baker, R.S.J.d., Corbett, A.T., Gowda, S.M., Wagner, A.Z., MacLaren, B.M., Kauffman, L.R., Mitchell, A.P., Giguere, S. (2010) Contextual Slip and Prediction of Student Performance After Use of an Intelligent Tutor.

Proceedings of the 18th Annual Conference on User Modeling, Adaptation, and Personalization, 52-63.

Baker, R. S., Goldstein, A. B., & Heffernan, N. T. (2011). Detecting learning moment-by-moment. *International Journal of Artificial Intelligence in Education*, *21*(1-2), 5-25.

Beck, J. E., & Gong, Y. (2013, July). Wheel-spinning: Students who fail to master a skill. In *Artificial Intelligence in Education*, 431-440.

Cen, H., Koedinger, K. R., & Junker, B. (2007). Is Over Practice Necessary? Improving Learning Efficiency with the Cognitive Tutor through Educational Data Mining. *Frontiers in Artificial Intelligence and Applications*, *158*, 511.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4), 253-278.

Falakmasir, M., Yudelson, M., Ritter, S., & Koedinger, K. (2015). Spectral bayesian knowledge tracing. In OC Santos, JG Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, JM Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, and M. Desmarais (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid, Spain, 360-364.

González-Brenes, J., Huang, Y., & Brusilovsky, P. (2014). General features in knowledge tracing to model multiple subskills, temporal item response theory, and expert knowledge. In *The 7th International Conference on Educational Data Mining*, 84-91.

Henderikx, M., Kreijns, K., & Kalz, M. (2017). Refining success and dropout in massive open online courses based on the intention-behavior gap. *Distance Education, 38* (3), 353-368.

Jones, E., Oliphant, E., Peterson, P., *et al.* (2001). **SciPy: Open Source Scientific Tools for Python**, http://www.scipy.org/

Khajah, M. M., Lindsey, R. V., & Mozer, M. C. (2014). Maximizing students' retention via spaced review: Practical guidance from computational models of memory. *Topics in cognitive science*, *6*(1), 157-169.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, *36*(5), 757-798.

Koedinger, K., Pavlik Jr, P. I., Stamper, J., Nixon, T., & Ritter, S. (2010, June). Avoiding problem selection thrashing with conjunctive knowledge tracing. In *Educational Data Mining 2011*, 91-100.

Kumari, S. (2018). Open and Distance Education System and Learner Support Services: An Introduction. In Anjana (Ed.), *Technology for Efficient Learner Support Services in Distance Education*. Springer, Singapore.

Miller, S. M. (2015). Teacher learning for new times: Repurposing new multimodal literacies and digital-video composing for schools. *Handbook of research on teaching literacy through the communicative and visual arts*, *2*, 441-453.

Moore, G., Baker, R.S., Gowda, S.M. (2015) The Antecedents of Moments of Learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 1631-1636.

Muilenburg, L., & Berge, Z. (2005). Student barriers to online learning: A factor analytic study. *Distance Education, 26,* 29-48. doi: 10.1080/01587910500081269

Murray, T., & Arroyo, I. (2002). Toward measuring and maintaining the zone of proximal development in adaptive instructional systems. In *International Conference on Intelligent Tutoring Systems* (pp. 749-758). Springer, Berlin, Heidelberg.

Nelder, J.A. & Mead, R. (1965). A simplex method for function minimization. *Computer Journal 7*, 308–313.

Mavrikis, M., Holmes, W., Zhang, J., & Ma, N. (2018). Fractions Lab Goes East: Learning and Interaction with an Exploratory Learning Environment in China. In *International Conference on Artificial Intelligence in Education*, 209-214. Springer.

Pardos, Z.A., Baker, R.S., San Pedro, M.O.C.Z., Gowda, S.M., Gowda, S.M. (2014) Affective states and state tests: Investigating how affect and engagement during the school year predict end of year learning outcomes. *Journal of Learning Analytics, 1* (1), 107-128.

Pardos, Z. A., Bergner, Y., Seaton, D. T., & Pritchard, D. E. (2013). Adapting Bayesian Knowledge Tracing to a Massive Open Online Course in edX. *Proceedings of the International Conference on Educational Data Mining*, *13*, 137-144.

Pardos, Z. A., & Heffernan, N. T. (2010). Modeling individualization in a bayesian networks implementation of knowledge tracing. In *User Modeling, Adaptation, and Personalization*, 255-266.

Pavlik Jr, P., Bolster, T., Wu, S. M., Koedinger, K., & Macwhinney, B. (2008, June). Using optimally selected drill practice to train basic facts. In *International Conference on Intelligent Tutoring Systems*, 593-602.

Pavlik Jr, P. I., Cen, H., & Koedinger, K. R. (2009). Performance Factors Analysis--A New Alternative to Knowledge Tracing. In V. Dimitrova & R. Mizoguchi (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 531-538.

Pelánek, R. (2014). Application of Time Decay Functions and the Elo System in Student Modeling. *Proc. of Educational Data Mining*, 21-27.

Roberts, J. (2018). Future and changing roles of staff in distance education: a study to identify training and professional development needs. *Distance Education*, *39*, 1-17. doi:10.1080/01587919.2017.1419818.

Rushkin, I., Rosen, Y., Ang, A., Fredericks, C., Tingley, D., Blink, M. J., & Lopez, G. (2017). Adaptive assessment experiment in a HarvardX MOOC. In *Proceedings of the 10th International Conference on Educational Data Mining*.

Sao Pedro, M.A., Baker, R.S.J.d., Gobert, J., Montalvo, O. Nakama, A. (2013) Leveraging Machine-Learned Detectors of Systematic Inquiry Behavior to Estimate and Predict Transfer of Inquiry Skill. *User Modeling and User-Adapted Interaction, 23* (1), 1-39.

Tang, H., & Xing, W. (2018). Exploring the temporal dimension of forum participation in MOOCs. *Distance Education, 39* (3), 353-372.

Thistoll, T., & Yates, A. (2016). Improving course completions in distance education: an institutional case study. *Distance Education, 37* (2), 180-195.