Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Transferable Student Performance Modeling for Intelligent Tutoring Systems

Robin SCHMUCKER[a*] & Tom M. MITCHELL[a]
[a]*Carnegie Mellon University, USA*
*rschmuck@cs.cmu.edu

Abstract: Millions of students worldwide are now using intelligent tutoring systems (ITSs). At their core, ITSs rely on student performance models (SPMs) to trace each student's changing ability level over time, in order to provide personalized feedback and instruction. Crucially, SPMs are trained using interaction sequence data of *previous* students to analyze data generated by *future* students. This induces a *cold-start* problem when a new course is introduced, because no students have yet taken the course and hence there is no data to train the SPM. Here, we consider transfer learning techniques to train accurate SPMs for new courses by leveraging log data from existing courses. We study two settings: (i) In the *naive transfer* setting, we first train SPMs on existing course data and then apply these SPMs to new courses without modification. (ii) In the *inductive transfer* setting, we fine tune these SPMs using a small amount of training data from the new course (e.g., collected during a pilot study). We evaluate the proposed techniques using student interaction sequence data from five different mathematics courses taken by over 47,000 students. The naive transfer models that use features provided by human domain experts (e.g., difficulty ratings for questions in the new course) but no student interaction training data for the new course, achieve prediction accuracy on par with standard BKT and PFA models that use training data from thousands of students in the new course. In the inductive setting our transfer approach yields more accurate predictions than conventional SPMs when only limited student interaction training data (<100 students) is available to both.

Keywords: performance modeling, knowledge tracing, transfer learning

1. Introduction

Intelligent tutoring systems (ITSs) are an educational technology that provides millions of students worldwide with access to learning materials and personalized instruction. Even though ITS offerings come at a much lower cost, they can in certain cases be as effective as a personal human tutor (VanLehn, 2011). ITSs can mitigate the academic achievement gap and help disadvantaged students (Huang et al., 2016). At their core, ITSs rely on student performance models (SPMs), to trace each student's changing ability level over time (Corbett & Anderson, 1994), to enable personalized curricula and feedback.

The increasing popularity of ITSs induces a need for SPM techniques that are flexible enough to support frequent releases of new courses, as well as changes to existing courses. The *cold-start* problem, which arises when a new course is released for which no student log data is available for SPM training, prevents us from applying conventional modeling approaches. In practice this means that the first batch of students does not enjoy the full benefits offered by the ITS. Future students then have the advantage that the log data generated by the early students can be used to train an accurate SPM.

In this paper we consider transfer learning (TL) techniques to improve the learning experience of early adopter students by mitigating the SPM cold-start problem for new courses. We show that TL can be used to train accurate SPMs for a new course by leveraging student log data collected from existing courses. We study two settings: (i) In the *naive transfer* setting where no data is available for the new course, we learn *course-agnostic* SPMs – i.e., models whose parameters can be trained using student interaction sequence data from existing courses and that can be applied to any new course. (ii) In the *inductive transfer* setting where small-scale new course data is available, we tune pre-trained course-agnostic SPMs to the new course by learning new course-specific question and knowledge component (KC) (i.e., skill) difficulty parameters. This inductive transfer setting mimics the case where the course designer can run a pilot with a small number of students before large-scale deployment.

We evaluate the proposed TL techniques using learning trajectory data from over 47,000 students collected from five different mathematics courses offered by a single ITS organization. In both settings, we find that the proposed techniques mitigate the cold-start problem for all courses. We hope that TL methods will become a standard tool for ITS designers and improve the learning experience of early students. The key contributions of this paper include:

- **Course-agnostic student performance models**. We present the first course-agnostic modeling techniques for predicting student performance on future questions in newly introduced courses where no previous students have yet taken this course. Even though our course-agnostic models have no access to training data logs of students taking the new course, they exhibit predictive performance comparable to conventional BKT and PFA models – found in many real-world ITSs – which were trained on data from thousands of students taking the new course.
- **Inductive transfer learning for efficient tuning**. We use transfer learning techniques to efficiently tune our pre-trained course-agnostic performance models to individual new courses by learning question- and KC-specific parameters. Our experiments show how our approach leads to more accurate performance predictions than conventional modeling techniques in settings in which only limited student log data from the new course is available (<100 students).
- **Guidance for practice**. By analyzing data from five different courses offered by a large-scale ITS this work provides insights which can inform the design of future ITSs. Among others, our experiments show how manually assigned difficulty ratings and information about distinct learning contexts provided by human domain experts during content creation can be used to boost the prediction accuracy of course-agnostic SPMs. Further, going against common guidance, our study of various existing SPM approaches reveals that large logistic regression models can outperform classical lower dimensional SPMs even in data starved settings (when training on <10 students).

2. Related Work

2.1 Transfer Learning

Transfer learning (TL) techniques are a class of machine learning (ML) algorithms which aim to improve model performance in a *target* domain (e.g., a new course) by leveraging data from a different but related *source* domain (e.g., existing courses) (Zhuang et al., 2020). TL is particularly attractive when only limited target domain data is available, but source domain data is abundant. Via pre-training on source domain data, TL can acquire a model for the target domain even when *no target domain data* is available. TL techniques enjoy great popularity in domains such as image classification and machine translation but have also been applied to various educational data mining (EDM) problems.

In the context of learning management systems (LMS), TL methods that combine data from multiple different courses or from multiple offerings of the same course have been explored for predicting academic performance (e.g., Tsiakmaki et al., 2020). Data collected from multiple courses has been used to predict the student's likelihood of completing future courses (Huynh et al., 2020) and their degree program (Hunt et al., 2017). In the setting of massive open online courses (MOOCs) TL can improve dropout predictions (e.g., Boyer & Veeramachaneni, 2015). Unlike all above-mentioned transfer approaches, in this work we do not predict a single attribute related to a current course (e.g., pass/fail, student dropout), but rather trace the changing likelihood with which students answer individual questions inside an ITS correctly over time based on their interaction history.

More related to the ITS setting considered in this paper, Paquette et al. (2015) studied the transfer of student gaming detection models between different courses and ITSs. Using simulated students, Spaulding et al. (2021) investigated an approach for transferring cognitive models of language learning between educational games. Multi-task learning has been proposed to learn useful representations via pre-training on response correctness and interaction time prediction tasks (Kim et al., 2021). Baker (2019) framed the problem of transferring student models (e.g., gaming detection models, SPMs, …)

between different learning systems as an open challenge at the EDM2019 conference. Recently, Baker et al. (2021) surveyed related work and discussed directions for future research on sharing models across different learning systems. While our work does not consider the transfer of SPMs across different ITSs, it focuses on the question of transferring SPMs between different courses inside the *same* ITS.

2.2 Student Performance Modeling

Tutoring systems rely on SPMs to estimate a student's ability level based on sequential log data that describes their prior interactions with the system. There are three major categories of SPMs: (i) Markov process-based inference, (ii) logistic regression and (iii) deep learning-based approaches. Markov process-based techniques, such as Bayesian Knowledge Tracing (BKT) (Corbett & Anderson, 1994) and BKT+ (Khajah et al., 2016), are well established and can for example be found in the Cognitive Tutor (Koedinger & Corbett, 2006) and the ASSISTments system (Feng et al., 2009). Most probabilistic approaches estimate a student's ability level by performing inference in a two state Hidden Markov Model – one state to represent mastery and one for non-mastery. Logistic regression models rely on a set of manually specified features which summarizes the student's interaction sequence. Given an input vector with feature values, the regression-based SPM estimates the probability that the student is proficient in a certain question or KC. Some approaches in this class are PFA (Pavlik et al., 2009), DAS3H (Choffin et al., 2019), Best-LR (Gervet et al., 2020) and Best-LR+ (Schmucker et al., 2022). Deep learning-based approaches take as input the same sequence data, but unlike logistic regression techniques can learn suitable features on their own without requiring human feature engineering. Deep learning models benefit from large-scale training data, but as of today, BKT- and logistic regression-based SPMs are still competitive with deep learning in multiple domains (e.g., Khajah et al., 2016; Schmucker et al., 2022). A survey on recent deep learning-based SPMs is provided by Liu et al. (2021).

Importantly, all above-mentioned SPM approaches rely on course-specific parameters (e.g., parameters that represent the difficulty of individual questions and KCs in the target course) that need to be learned from target course data. This makes these models inapplicable in our cold start setting where a new course is first introduced and there is no data for training these parameters available yet.

Lastly, we want to mention recent works (Gervet et al., 2020; Zhang et al., 2021) which investigated another SPM related cold-start problem. There, the question is how accurate are SPM predictions for new students for which we have only observed a few interactions. This is different from the cold-start problem studied in this paper – it addresses the question of how to handle a new cold-start student in an existing course, whereas we address the question of how to handle a new cold-start course. Related to the inductive transfer setting studied in this work, is a short-paper by Zhao et al. (2020) which proposed an Attentive Neural Turing Machine architecture that requires less training data than an LSTM based approach. Unlike our study, they only experiment with small-scale student log data (<30 students, <1000 responses) and do not leverage data collected from existing courses for knowledge transfer.

3. Problem Setting

3.1 The Student Performance Modeling Problem

Formally, we denote the sequence of student s's past interaction with the system as $\boldsymbol{x}_{\boldsymbol{s},\boldsymbol{1}:\boldsymbol{t}} = (x_{s,1}, \dots, x_{s,t})$. The tuple $x_{s,t} = (y_{s,t}, q_{s,t}, c_{s,t})$ represents the data collected for student s at time-step t. Variable $q_{s,t}$ indicates the answered question, $y_{s,t} \in \{0,1\}$ is binary response correctness and $c_{s,t}$ is an aggregation of additional information about question difficulty, learning context, read materials, watched videos and time. Provided student s's history $\boldsymbol{x}_{\boldsymbol{s},\boldsymbol{1}:\boldsymbol{t}}$ and a question $q_{s,t+1}$, a SPM f_w estimates $p(y_{s,t+1} = 1 \mid q_{s,t+1}, \boldsymbol{x}_{\boldsymbol{s},\boldsymbol{1}:\boldsymbol{t}})$ as the probability of s responding correctly to $q_{s,t+1}$ if it were asked next.

All SPMs considered in this paper are parametric and defined by a vector $w \in R^d$. Using training data $D = \{\boldsymbol{x}_{\boldsymbol{s_1},\boldsymbol{1}:\boldsymbol{t_1}}, \dots, \boldsymbol{x}_{\boldsymbol{s_n},\boldsymbol{1}:\boldsymbol{t_n}}\}$ capturing interaction logs from *previous* students one can determine a vector w_D for predicting the performance of *future* students by solving the minimization problem

$$w_D = arg \min_{w \in R^d} \sum_{s \in D} \sum_{t=1}^{t_s} L\big(f_w(q_{s,t}, \boldsymbol{x}_{\boldsymbol{s},\boldsymbol{1}:\boldsymbol{t-1}}), y_{s,t}\big). \quad (1)$$

Here, $L(\hat{y}_{s,t}, y_{s,t}) = -(y_{s,t} log(\hat{y}_{s,t}) + (1 - y_{s,t})\log(1 - \hat{y}_{s,t}))$ is the negative conditional log-likelihood of observed student response correctness $y_{s,t}$ given model prediction $\hat{y}_{s,t} = f_w(q_{s,t}, \boldsymbol{x}_{\boldsymbol{s},\boldsymbol{1}:\boldsymbol{t-1}})$ and student history $\boldsymbol{x}_{\boldsymbol{s},\boldsymbol{1}:\boldsymbol{t-1}}$. This function penalizes predictions $\hat{y}_{s,t}$ that deviate from observation $y_{s,t}$.

3.2 Dataset

For our analysis we rely on the *Squirrel Ai ElemMath2021* dataset (Schmucker et al., 2022) which provides log data from multiple mathematics courses for elementary school students collected over a 3-month period. Overall, the dataset describes about 62,500,000 interactions from over 125,000 students. Going beyond pure question-solving activities, *ElemMath2021* provides insights into how students interact with learning materials. During content creation human domain experts assign each question a difficulty rating between 10 and 90 and specify a prerequisite graph to describe dependencies between individual KCs. *ElemMath2021* further records information about the learning context by assigning each learning activity to one of six categories of study modules (e.g., pre-test, post-test, review, …).

Our study of the transferability of SPMs partitions *ElemMath2021* into multiple course-specific datasets. We selected the five courses with the most students, which we refer to as *C6*, *C7*, *C8*, *C9* and *C40*. Together, these five courses capture approximately 26,300,000 interactions from over 47,000 students (Table 1). Each student only participates in a single course which implies *disjoint* student populations across courses. In terms of covered KCs and used questions the courses are also *disjoint* except for *C9* and *C40* which have an overlap of less than 5%. These properties allow us to measure the transferability of SPMs to different courses involving disjoint students and disjoint questions and KCs.

Table 1. *Five largest ElemMath2021 courses by student number. Avg. responses is the average number of submitted responses per student. Avg. correctness is the proportion of correct student responses.*

course	*C6*	*C7*	*C8*	*C9*	*C40*
# of students	11,864	9,423	10,296	8,531	7,487
# of questions	2,483	2,226	2,438	2,407	1,307
# of KCs	164	145	159	157	87
# of responses	3,262k	1,934k	2,142k	1,407k	1,228k
avg. responses	275	227	187	165	164
avg. correctness	71.30%	69.62%	69.47%	68.68%	62.39%

4. Approach

4.1 Naive Transfer Approach

The naive transfer setting is concerned with using student log data D_S from existing *source* courses $S = \{S_1, \dots, S_k\}$ to learn an SPM that can be applied to any future *target* course T. Crucially, such a *course-agnostic* SPM approach *cannot* rely on any parameters that describe *course-specific* features. Because existing SPMs rely on parameters that capture properties of individual questions and KCs, they require access to target course data D_T for training and are thus not applicable when such data is not available.

As a first step in the design of course-agnostic SPMs we identify a set of features which induces model parameters that do not require target course data for training. For this we study existing logistic regression-based SPMs. Each regression model relies on a distinct feature function $\Phi = (\phi_1, \dots, \phi_d)$ which outputs a real-valued feature vector that describes student s's prior interaction history $\boldsymbol{x}_{\boldsymbol{s},\boldsymbol{1}:\boldsymbol{t}}$ and information about the next question $q_{s,t+1}$. The trained model then uses this feature vector as input to estimate the probability that s will respond correctly to question $q_{s,t+1}$ if it were asked next as

$$p\left(y_{s,t+1} = 1 \,\middle|\, q_{s,t+1}, \boldsymbol{x}_{\boldsymbol{s},\boldsymbol{1}:\boldsymbol{t}}\right) = \sigma\left(w^{\mathsf{T}}\Phi\left(q_{s,t+1}, \boldsymbol{x}_{\boldsymbol{s},\boldsymbol{1}:\boldsymbol{t}}\right)\right). \tag{2}$$

Here $w \in R^d$ is the learned weight vector that defines the model and $\sigma(x) = 1/(1 + e^{-x}) \in [0,1]$ is the sigmoid function whose output can be interpreted as the probability of correct response.

Because conventional SPMs use feature functions that target course-specific features they do not generalize to new courses. As an example, consider the Best-LR model by Gervet et al. (2020). It features an ability parameter α_s for each individual student and difficulty parameters δ_q and β_k for each individual question q and KC k. Further, Best-LR uses count features for the number of prior correct (c_s) and incorrect (f_s) responses of student s overall and for each individual KC k (i.e., $c_{s,k}$ and $f_{s,k}$). Defining scaling function $\phi(x) = log(1 + x)$, the Best-LR prediction is

$$\begin{aligned} p_{\textit{Best-LR}}\left(y_{s,t+1} = 1 \,\middle|\, q_{s,t+1}, \boldsymbol{x}_{\boldsymbol{s},\boldsymbol{1}:\boldsymbol{t}}\right) = \sigma(&\alpha_s - \delta_{q_{s,t+1}} + \tau_c\phi(c_s) + \tau_f\phi(f_s) \\ &+ \sum_{k \in KC(q_{s,t+1})} \beta_k + \gamma_k\phi(c_{s,k}) + \rho_k\phi(f_{s,k})). \end{aligned} \tag{3}$$

One can interpret the Best-LR feature function as a tuple $\Phi = (\Phi_A, \Phi_T)$ where Φ_A is course-agnostic (i.e., total counts) and Φ_T is target course-specific (i.e., student ability, question and KC difficulty and counts). Because – to the best of our knowledge – this is the first work that investigates course-agnostic SPMs we introduce simple but reasonable baselines by taking conventional SPM approaches and reducing them to their course-agnostic feature sets.

From Best-LR we derive a course-agnostic SPM called A-Best-LR. A-Best-LR uses overall count features c_s and f_s to indicate the number of student s's prior correct and incorrect responses. The two parameters γ and ρ consider the number of prior correct ($c_{s,k}$) and incorrect responses ($f_{s,k}$) for the *current* KC k – the *same* γ and ρ parameters are used for *all* KCs. Best-LR's ability parameters are reduced to a *single* bias term α that is constant over time for *all* students. The A-Best-LR prediction is

$$p_{\textit{A-Best-LR}}\left(y_{s,t+1} = 1 \,\middle|\, q_{s,t+1}, \boldsymbol{x}_{\boldsymbol{s},\boldsymbol{1}:\boldsymbol{t}}\right) = \sigma\left(\alpha + \tau_c\phi(c_s) + \tau_f\phi(f_s) + \gamma\phi\left(c_{s,k}\right) + \rho\phi\left(f_{s,k}\right)\right). \tag{4}$$

By avoiding course-specific features A-Best-LR can be trained on source data D_S from existing courses and then be used for any new course T. Giving a similar treatment to other common SPMs we define:

- A-BKT: We train a single BKT parameter set shared for all KCs. We then estimate student performance by using this parameter set to initialize a separate BKT model for each individual KC.
- A-IRT: We train an IRT (Rasch) model that uses the same difficulty parameter (δ) for all questions. We then use this single difficulty parameter to trace each student's ability over time for each KC and derive performance predictions. The student ability parameters are updated after each response.
- A-PFA: We train a reduced 3-parameter PFA model that uses the same difficulty (δ), correctness (γ), and incorrectness count parameters (ρ) for all KCs.
- A-DAS3H: We train a reduced DAS3H model that uses a shared difficulty parameter (δ) for all questions and KCs, a shared constant ability bias term (α) for all students and a single a set of time-window based correctness and incorrectness count parameters for all KCs.
- A-Best-LR+: We train a reduced Best-LR+ model that augments the A-Best-LR feature set (Equation 3) with response pattern and smoothed average correctness features (Schmucker et al., 2022). In addition, the model learns a single set of DAS3H time-window (Choffin et al., 2019), R-PFA (Galyardt & Golding, 2015) and PPE (Walsh et al., 2018) count parameters used for all KCs.

Related to A-BKT, Corbett & Anderson (1994) evaluated a version of BKT which uses a single set of BKT parameters for all KCs that is trained and tested on data from the same course. Related to A-PFA, Maier et al. (2020) proposed to learn PFA parameters for KCs with enough training data and to use the average of the parameters to model KCs with insufficient data in the same course. A-BKT and A-PFA are different in that they train on data from existing courses and then make predictions for a new course.

Conventional SPMs – including all the above – base their estimates solely on log data that describes the student's question-answering behavior. Recently, it has been shown how alternative types of log data collected by modern ITSs can improve logistic regression-based SPMs (Schmucker et al., 2022). The use of such alternative types of features is particularly interesting in the naive transfer setting because most conventional SPM features are course-specific and are thus not transferable. The *ElemMath2021* dataset captures various types of student interaction data. In our experiments we consider information related to student video and reading material usage, learning context, question difficulty ratings assigned by human domain experts during content creation, KC prerequisite structure and the response- and lag-time features introduced by SAINT+ (Shin et al., 2021).

4.2 Inductive Transfer Approach

Most conventional SPM approaches rely on parameters that capture question- and KC-specific attributes. By training and testing on *target* course data D_T using a 5-fold cross validation, Table 2 compares the performance of course-agnostic SPMs with models that use the same *course-agnostic* feature set, but which are allowed to learn additional *course-specific* parameters to capture question- and KC-difficulty. We observe that the inclusion of question- and KC-specific parameters leads to large improvements in prediction accuracy and closes the gap to conventional SPM techniques (Table 3).

Table 2. *When training and testing on data from the same course, adding course-specific question- and KC-difficulty parameters to the course-agnostic A-AugLR model increases accuracy (ACC) and AUC.*

	ACC (%)	AUC (%)
A-AugLR	72.02	69.48
A-AugLR+KC	74.00	74.99
A-AugLR+quest.	76.34	79.39
A-AugLR+KC+quest.	**76.37**	**79.39**

Motivated by this observation, we propose an inductive TL approach that uses small-scale target course data D_T to tune a pre-trained course-agnostic SPM to a new course T by learning additional question- and KC-specific parameters. Formally, the pre-trained agnostic and target model are defined by weight vectors $w_S \in R^{|\Phi_S|}$ and $w_T \in R^{|\Phi_S|+|\Phi_T|}$ respectively. We use L_2 regularization to subject the target weights w_T to a Gaussian prior $\mathcal{N}((w_S, 0)^\intercal, 1)$ and control the degree of regularization using a penalty parameter $\lambda \in R_{\geq 0}$. The regularized maximum likelihood objective is

$$w_T = arg \min_{w \in R^d} \frac{\lambda}{2} ||w - \binom{w_S}{0}||_2^2 + \sum_{s \in D_T} \sum_{t=1}^{t_s} L\big(f_w\big(q_{s,t}, \boldsymbol{x}_{\boldsymbol{s},\boldsymbol{1}:\boldsymbol{t-1}}\big), y_{s,t}\big). \quad (5)$$

By using a prior for w_T that is based on the previously learned w_S, we can mitigate overfitting and can learn a suitable target model using only very limited training data D_T. With increasing amounts of recorded learning histories in D_T the objective focuses increasingly on model fit. For our experiments we determine the penalty parameter value by evaluating $\lambda \in \{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 100\}$ using the first split of a 5-fold cross validation on the *C6* training data. We found $\lambda = 5$ to be most effective for different amounts of tuning data and use it for all our experiments.

5. Experiments

5.1 Evaluation Methodology

As is common in prior work (e.g., Choffin et al., 2019; Gervet et al., 2020) we filter out students with less than ten answered questions. In the naive transfer setting, we use each course once to simulate a new target course $T \in \{C6, C7, C8, C9, C40\}$. For each target course T we train one course-agnostic SPM using source data D_S from the other four courses and then evaluate predictions on the unseen target dataset D_T. For the inductive transfer experiments, we perform a 5-fold cross-validation on the student level where in each fold 80% of students are used as training set $D_{T,train}$ and the remaining 20% are used as test set $D_{T,test}$. To simulate small-scale training data, we sample a limited number of students $(5, 10, \ldots)$ from training set $D_{T,train}$. Because the *ElemMath2021* courses tend to introduce topics in the same sequential order, we only sample students that reached the last topic – sampled students might have skipped or revisited individual topics. This approach mimics the case where the course designer can collect interaction log data from a small number of students during a pilot study before large-scale deployment. We report model performance using accuracy (ACC) and area under curve (AUC) metrics. AUC is a common evaluation metric for SPMs which can be interpreted as the probability that the model ranks a random correct student response higher than a random incorrect response.

Table 3. *Reference model performance. Performance metrics achieved by conventional course-specific student performance models that were trained and tested on data from the same course.*

	C6		*C7*		*C8*		*C9*		*C40*		Averaged	
model \ in %	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Always correct	71.30	50.00	69.62	50.00	69.47	50.00	68.68	50.00	62.38	50.00	68.29	50.00
BKT	74.89	73.39	71.66	69.35	72.24	70.43	72.01	70.09	68.09	71.00	71.78	70.85
PFA	74.66	73.02	71.52	69.19	72.13	70.21	71.87	69.94	67.85	70.87	71.61	70.65
IRT	75.52	75.66	73.05	73.22	73.28	73.21	72.40	72.36	68.66	72.05	72.58	73.30
DAS3H	77.31	78.15	74.59	76.06	75.05	76.18	74.09	75.38	70.87	75.20	74.38	76.19
Best-LR	78.42	80.30	75.95	78.44	76.58	78.97	76.33	79.08	73.10	78.07	76.08	78.97
Best-LR+	**78.75**	**80.85**	**76.18**	**78.83**	**76.90**	**79.39**	**76.69**	**79.58**	**73.62**	**78.81**	**76.43**	**79.49**

Our code builds on the public GitHub repository by Schmucker et al. (2022) which implements various SPMs. We build on their regression models and leave their hyperparameter choices unchanged. We use pyBKT (Badrinath et al., 2021) to implement the BKT experiments. For our naive and inductive transfer experiments we use PyTorch and train each model for 200 epochs using the Adam optimizer with learning-rate $\alpha = 0.001$. As a reference, Table 3 shows average performance metrics of common SPM approaches that were trained and tested on the *same* course using a 5-fold cross-validation. To increase reproducibility, we provide detailed descriptions of the evaluated features and SPMs in an external appendix hosted on GitHub (https://github.com/rschmucker/TransferableSPM-Appendix).

5.2 Naive Transfer Experiments

Feature Evaluation. We evaluate the benefits of different features for course-agnostic SPMs. For each feature, we train an augmented A-Best-LR+ model using the A-Best-LR+ feature set plus one of several possible additional features, described below. We use A-Best-LR+ because it combines features that were found most useful in earlier SPMs and it yields the most accurate predictions among all considered course-agnostic baseline models in our experiments (Table 5).

Table 4 shows the ACC and AUC scores when adding each of several additional features to A-Best-LR+. The most useful additions are the one-hot features that encode question difficulty ratings assigned by human domain experts during content creation – these improve performance on average over all five courses by 0.24% ACC and 1.07% AUC. The one-hot learning context features improve the average AUC score by 0.14%. The count features that track the number of prior correct and incorrect responses to questions of a certain difficulty or learning context, lead to smaller improvements compared to their one-hot counterparts. The lag time and response time features improve AUC scores on average by 0.15% and 0.11%. The post- and pre-requisite features derived from the KC dependency graph did not benefit the course-agnostic SPMs. Similarly, the count features that summarize the students' video and reading material usage did not improve the performance predictions. One limitation of these two count features is that they do not capture the relationship between the content covered by individual learning materials and questions.

Table 4. *Naive transfer feature evaluation. Using the A-BestLR+ feature set augmented with one additional feature we trained course-agnostic models on four source courses and evaluated on a new target course. The marker* **X** *indicates which additional features yielded the largest improvements.*

	C6		*C7*		*C8*		*C9*		*C40*		Averaged		
model \ in %	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	
A-BestLR+ (base)	73.86	67.33	71.62	65.41	71.92	65.79	72.36	68.92	67.59	68.31	71.47	67.15	
current lag time	73.91	67.48	71.55	65.42	72.00	66.00	72.40	69.02	67.65	68.38	71.50	67.26	**X**
prior resp. time	73.94	67.61	71.57	65.39	72.02	65.98	72.39	69.12	67.46	68.41	71.48	67.30	**X**
context one-hot	73.83	67.24	71.65	65.56	71.94	65.95	72.38	69.02	67.65	68.70	71.49	67.29	**X**

context count	73.87	67.38	71.53	65.33	72.04	65.79	72.41	69.09	67.71	68.54	71.51	67.23	**X**
difficulty one-hot	74.09	68.63	71.88	66.80	72.22	67.21	72.54	69.71	67.82	68.84	71.71	68.24	**X**
difficulty count	73.84	67.34	71.60	65.56	71.93	66.00	72.33	69.00	67.59	68.52	71.46	67.28	**X**
prereq count	73.88	67.27	71.58	65.44	71.91	65.83	72.31	68.92	67.55	68.28	71.45	67.15	
postreq count	73.61	66.48	71.68	65.03	71.98	65.96	72.38	69.22	67.39	68.15	71.41	66.97	
videos count	73.84	67.32	71.59	65.41	71.95	65.75	72.30	68.91	67.52	68.20	71.44	67.12	
readings count	73.83	67.41	71.60	65.48	71.96	65.82	72.37	68.93	67.53	68.19	71.46	67.17	

Agnostic AugmentedLR. Given the experimental results in Table 4, we created a final model (which we call A-AugLR) that incorporates all the features from A-Best-LR+ (base), plus all features marked by X in Table 4, namely lag and response time, learning context and question difficulty. To complete our experiments on the naive transfer setting, we then compared the performance of A-AugLR to the naive transfer baselines defined in Subsection 4.1. The results of this comparison are shown in Table 5. We observe that the course-agnostic SPMs derived from BKT, PFA, IRT and DAS3H struggle in the naive transfer setting and yield low AUC scores. The A-Best-LR+ model uses additional features to capture aspects of long- and short-term student performance over time. On average, its predictions yield 0.37% higher ACC and 1.09% higher AUC scores compared to the A-Best-LR model it builds on.

The A-AugLR model yields the best predictions in the naive transfer setting. Compared to A-Best-LR+, the A-AugLR models are on average 0.38% more accurate and their AUC scores are 1.76% higher. This shows how additional information provided by domain experts during content creation can enable accurate performance predictions on cold start courses in the naive transfer setting. Importantly, even though our course-agnostic A-AugLR models were fitted using data from different source courses their prediction accuracy is on par with course-specific BKT and PFA models which were trained on target course data (compare the A-AugLR row in Table 5 to the BKT and PFA rows from Table 3).

Table 5. *Naive transfer performance. We used each of the five courses to simulate a new target course and trained course-agnostic performance models using student data from the other four source courses.*

	C6		*C7*		*C8*		*C9*		*C40*		Averaged	
model \ in %	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
A-BKT	73.25	63.29	70.58	60.87	71.04	60.85	70.93	62.30	65.56	63.87	70.27	62.57
A-PFA	73.27	63.54	70.75	60.59	71.13	61.23	70.92	62.29	65.55	63.13	70.32	62.16
A-IRT	59.50	61.82	58.36	59.55	57.50	59.45	58.04	60.96	58.33	62.72	58.35	60.90
A-DAS3H	73.29	63.70	70.81	60.84	71.15	61.31	70.98	62.41	65.59	63.54	70.36	62.36
A-Best-LR	73.55	66.38	71.35	64.17	71.74	65.01	71.99	67.87	67.13	66.86	71.15	66.06
A-Best-LR+	73.86	67.33	71.62	65.41	71.92	65.79	72.36	68.92	67.59	68.31	71.47	67.15
A-AugLR	**74.28**	**69.11**	**71.80**	**67.21**	**72.35**	**68.19**	**72.76**	**70.52**	**68.05**	**69.52**	**71.85**	**68.91**

5.3 *Inductive Transfer Experiments*

Here, we evaluate our inductive TL approach (I-AugLR) that uses small-scale target course data $\mathrm{D_T}$ to tune a course-agnostic A-AugLR model – pre-trained on log data from the other courses – to the target course by learning course-specific difficulty parameters. We also evaluate a course-specific model (S-AugLR) which use the same feature set as I-AugLR but does not use a pre-trained model. We measure the amount of target course data for tuning in number of students and experiment with values in $\{0, 5, 10, 25, 50, 100, 250, 500, 1000\}$. The 0-student case is equivalent to the naive transfer setting.

Figure 1 compares the performance of our inductive TL method (I-AugLR) with conventional SPM approaches and S-AugLR trained using only target course data D_T. Due to the page limit, we only plot model performance for *C40*. By tuning a pre-trained A-AugLR model, I-AugLR can mitigate the cold-start problem for all five courses and benefits from small-scale log data. Given as little as data from 10 students, the I-AugLR models consistently outperform standard BKT and PFA models that were trained on logs from thousands of target course students (Table 3). Among all considered SPMs, I-AugLR yields the most accurate performance prediction up to 25 students for C7, up to 100 students for *C6* and *C8* and up to 250 students for *C9* and *C40*. Among the non-TL approaches, Best-LR is most data efficient and yields the best performance predictions when training on up to 500 students.

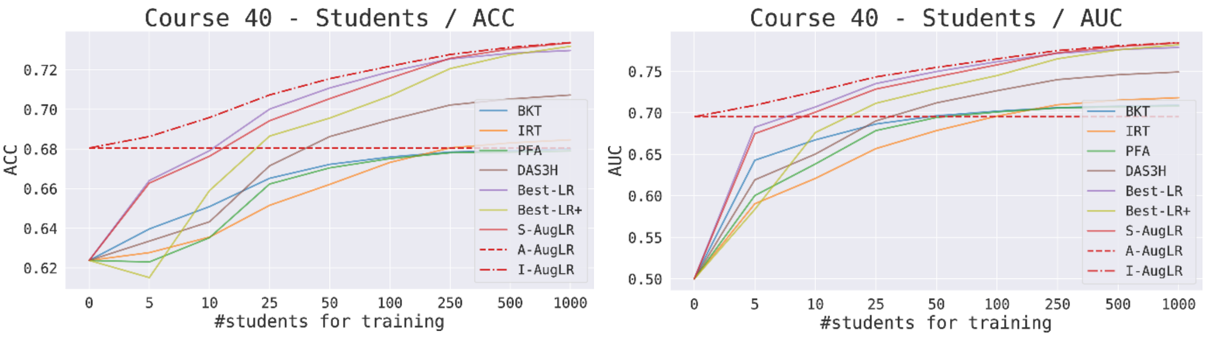

Figure 1. Training student performance models with different amounts of student log data.

6. Discussion

Our experiments show that our proposed TL techniques can mitigate the SPM cold-start problem for new courses by leveraging student interaction data from other existing courses. In the naive transfer setting where no target course data is available, the course-agnostic A-AugLR models trained on data from existing source courses yielded prediction accuracy on par with standard BKT and PFA models that use training data from thousands of students in the target course. One key ingredient of our course-agnostic SPMs is additional information about question difficulty and learning context provided by domain experts during content creation. While these features improve SPM predictions, the need for manual annotations puts an additional load on the content creators. Further, the success of our transfer approach depends to a degree on the domain expert's ability to assign accurate question difficulty labels.

In the inductive transfer setting we use small-scale target course data (e.g., collected during a pilot study) to tune pre-trained course-agnostic SPMs. This allows us to overcome the limitations of the naive transfer setting by learning target course specific question- and KC- difficulty parameters. Our parameter regularized approach yields better predictions than conventional SPM approaches when only limited target course data (<100 students) is available. Surprisingly, we found that among the non-TL approaches, Best-LR (Gervet et al., 2020) yielded the most accurate predictions when training on less than 500 students for all five courses. This is interesting because low-parameter models such as BKT (Corbett & Anderson, 1994), PFA (Pavlik et al., 2009) and IRT (Rasch, 1960) are commonly believed to be more data efficient than more complex logistic regression models that contain many more parameters. What sets Best-LR apart from these three models, is that its parameters describe student performance using multiple levels of abstraction (question-, KC- and overall-level). Future work might investigate this phenomenon further using log data from multiple different ITSs.

One limitation of our study is that it focuses on a set of five courses offered by the same ITS. This has advantages because the course log data is of consistent format and content creators follow similar protocols. Still, it prevents us from answering the question of whether SPMs are transferable between different tutoring systems (Baker, 2019). Another related limitation is that all considered courses cover mathematics topics for elementary school students. Our study did not investigate the transferability of SPMs across different subjects or grade levels (e.g., middle school, high school, …).

7. Conclusion

The increasing popularity of intelligent tutoring systems (ITSs) induces a need for student performance modeling (SPM) techniques that are flexible enough to support frequent new course releases as well as changes to existing courses. This paper proposes two transfer learning approaches for mitigating the cold-start problem that arises when a new course is introduced for which no training data is available. In the naive transfer setting where no new course data is available, we rely on student interaction sequences from existing courses to learn course-agnostic SPMs that can be applied to any future course. In the inductive transfer setting where small-scale new course data is available (e.g., collected during a pilot study), we show how one can tune pre-trained course-agnostic models to a specific course by learning question- and KC- (i.e., skill) difficulty parameters. Our experimental evaluation on student log data from five different mathematics courses showed how both transfer approaches mitigate the cold-start problem successfully. This work represents a first step in the design of SPMs that are transferable between different ITS courses. The success of our approach depends significantly on (i) automatically learned course-independent parameters that characterize how quickly students learn a skill as a function of the number of prior practice attempts, and (ii) information provided by human domain experts in the form of difficulty values for questions in the new, target course. We hope that transfer learning techniques such as the ones discussed in this paper will enable ITS designers to provide effective adaptive instruction for early adopter students.

Acknowledgements

We thank Richard Tong and Dan Bindman from Squirrel Ai Learning for providing the *ElemMath2021* dataset. This work was supported in part through the CMU - Squirrel Ai Research Lab on Personalized Education at Scale, and in part by the AFOSR under award FA95501710218.

References

Badrinath, A., Wang, F. & Pardos, Z. (2021). pyBKT: An Accessible Python Library of Bayesian Knowledge Tracing Models. In *Proc. of the 14th Int. Conference on EDM* (pp. 468-474). Paris, France: EDM.

Baker, R. S. (2019). Challenges for the Future of Educational Data Mining: The Baker Learning Analytics Prizes. *Journal of Educational Data Mining, 11*(1), 1-17.

Baker, R. S., McLaren, B. M., Hutt, S., Richey, J. E., Rowe, E., Almeda, M., Mogessie, M. & Andres, J. M. (2021). Towards sharing student models across learning systems. In *Proc. of the 22nd Int. Conference on AIED* (pp. 60-65). Utrecht, Netherlands: Springer

Boyer, S., & Veeramachaneni, K. (2015). Transfer Learning for Predictive Models in Massive Open Online Courses. In *Proc. of the 17th Int. Conference on AIED* (pp. 54-63). Madrid, Spain: Springer.

Choffin, B., Popineau, F., Bourda, Y., & Vie, J. J. (2019). DAS3H: Modeling Student Learning and Forgetting for Optimally Scheduling Distributed Practice of Skills. In *Proc. of the 12th Int Conference on EDM* (29-38). Montreal, Canada: EDM

Corbett, A. T., & Anderson, J. R. (1994). Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User modeling and user-adapted interaction*, 4(4), 253-278.

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User modeling and user-adapted interaction, 19*(3), 243-266.

Galyardt, A., & Goldin, I. (2015). Move Your Lamp Post: Recent Data Reflects Learner Knowledge Better than Older Data. *Journal of Educational Data Mining*, *7*(2), 83-108.

Gervet, T., Koedinger, K., Schneider, J., & Mitchell, T. (2020). When is Deep Learning the Best Approach to Knowledge Tracing?. *Journal of Educational Data Mining*, *12*(3), 31-54.

Huang, X., Craig, S. D., Xie, J., Graesser, A., & Hu, X. (2016). Intelligent tutoring systems work as a math gap reducer in 6th grade after-school program. *Learning and Individual Differences*, 47, 258-265.

Hunt, X. J., Kabul, I. K., & Silva, J. (2017). Transfer Learning for Education Data. In *KDD Workshop on Advancing Education with Data*. Nova Scotia, Canada: ACM

Huynh-Ly, T. N., Le, H. T., & Nguyen, T. N. (2020). Integrating courses' relationship into predicting student performance. *Int. Journal of Advanced Trends in Computer Science and Engineering*, 9(4). 6375-6383.

Khajah, M., Lindsey, R. V., & Mozer, M. C. (2016). How Deep is Knowledge Tracing?. In *Proc. of the 9th Int. Conference on EDM* (pp. 94-101). Raleigh, NC, USA: EDM

Kim, B., Yu, H., Shin, D., & Choi, Y. (2021). Knowledge Transfer by Discriminative Pre-training for Academic Performance Prediction. In *Proc. of the 14th Int. Conference on EDM* (pp. 287-294). Virtual: EDM

Koedinger, K. R., & Corbett, A. (2006). Cognitive Tutors: Technology Bringing Learning Sciences to the Classroom. In R. K. Sawyer (Ed.), *The Cambridge handbook of: The learning sciences* (pp. 61–77). Cambridge University Press.

Liu, Q., Shen, S., Huang, Z., Chen, E., & Zheng, Y. (2021). A Survey of Knowledge Tracing. *arXiv preprint arXiv:2105.15106*.

Maier, C., Baker, R. S. & Stalzer, S. (2021). Challenges to Applying Performance Factor Analysis to Existing Learning Systems. In *Proc. of the 29th ICCE* (pp. 57-62). Virtual: APSCE

Paquette, L., Baker, R. S., Carvalho, A. D., & Ocumpaugh, J. (2015). Cross-system Transfer of Machine Learned and Knowledge Engineered Models of Gaming the System. In *Proc. of the 23rd Int. Conference on UMAP* (pp. 183-194). Dublin, Ireland: Springer.

Pavlik Jr, P. I., Cen, H., & Koedinger, K. R. (2009). Performance Factors Analysis – A New Alternative to Knowledge Tracing. In *Proc. of the 14th Int. Conference on AIED* (pp. 531-538). Brighton, UK: IOS Press

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

Schmucker, R., Wang, J., Hu, S., & Mitchell, T. M. (2022). Predicting the Performance of Online Students - New Data, New Approaches, Improved Accuracy. *Journal of Educational Data Mining*, *14*(1), 1–45.

Shin, D., Shim, Y., Yu, H., Lee, S., Kim, B., & Choi, Y. (2021). SAINT+: Integrating temporal features for EdNet correctness prediction. In *Proc. of the 11th Conference on LAK* (pp. 490-496). New York, NY, USA: ACM

Spaulding, S., Shen, J., Park, H., & Breazeal, C. (2021). Towards Transferrable Personalized Student Models in Educational Games. In *Proc. of the 20th Int. Conference on AAMAS* (pp. 1245-1253). Virtual: ACM

Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., & Ragos, O. (2020). Transfer Learning from Deep Neural Networks for Predicting Student Performance. *Applied Sciences*, *10*(6), 2145.

VanLehn, K. (2011). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational psychologist*, *46*(4), 197-221.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., Krusmark, M., Myung, J. I., Pitt, M. A., & Zhou, R. (2018). Mechanisms Underlying the Spacing Effect in Learning: A Comparison of Three Computational Models. *Journal of Experimental Psychology: General*, *147*(9), 1325–1348.

Zhang, J., Das, R., Baker, R. S., & Scruggs, R. (2021). Knowledge Tracing Models' Predictive Performance when a Student Starts a Skill. In *Proc. of the 14th Int. Conference on EDM* (pp. 625-629). Virtual: EDM

Zhao, J., Bhatt, S., Thille, C., Gattani, N., & Zimmaro, D. (2020). Cold Start Knowledge Tracing with Attentive Neural Turing Machine. In *Proc. of the 7th Conference on L@S* (pp. 333-336). New York, NY, USA: ACM

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. & He, Q. (2020). A Comprehensive Survey on Transfer Learning. In *Proc. of the IEEE*, *109*(1), 43-76.