# Studying Memory Decay and Spacing within Knowledge Tracing

**Cristina MAIER[a*], Isha SLAVIN[b], Ryan S. BAKER[c] & Steve STALZER[d]**
[a]*McGraw Hill Education, USA*
[b]*McGraw Hill Education, USA*
[c]*University of Pennsylvania, USA*
[d]*McGraw Hill Education, USA*
*cristina.maier@mheducation.com

## ABSTRACT
Knowledge Tracing estimates a student's knowledge on a set of skills and predicts whether the student will answer correctly if given a question linked to subsets of such skills. We conduct an in-depth analysis on capturing cognitive science principles such as memory decay and spacing and measure their effects within knowledge tracing. To do this, we propose a new algorithm called MemDec which incorporates memory decay theory into knowledge estimation. This model is further expanded to consider the spacing effect, another pivotal cognitive science concept. We explore different methods of modeling the rate and weight of decay, with and without the spacing effect, and analyze the role they play in predicting student performance within real-world data. Variations of the model are compared between each other as well as against other existing algorithms.

## Keywords
Knowledge tracing, memory decay, spacing effect, learning systems

## 1. INTRODUCTION
*Knowledge tracing* is an area in the educational data mining field that is concerned with the estimation of mastery on a set of skills. Knowledge tracing attempts to represent student learning trajectories as a sequence of knowledge state changes to predict when a student has mastered a skill. The outputs of knowledge tracing have several uses, including use in behavioral and self-regulated learning models, reports to instructors with actionable insights, and driving mastery learning within adaptive learning systems [30].

Although the majority of the most recent work in knowledge tracing has investigated refinements to deep learning algorithms [31, 40, 8], there has also been interest in developing knowledge tracing algorithms that leverage findings from cognitive science [5, 27]. In this paper, we investigate the implementation of cognitive science principles within knowledge tracing in detail, specifically focusing on memory decay and the spacing effect when applied to variants of some components of Logistic Knowledge Tracing (LKT) [27] with an emphasis on the components from Performance Factor Analysis (PFA) [28] and Recent-Performance Factor Analysis (R-PFA) [10]. We chose to study memory decay and spacing within this proposed algorithm because of PFA's interpretability and its support of multi-skill items. To study these concepts, we introduce a new algorithm that incorporates memory decay into the knowledge estimation and investigate how the spacing effect and different ways of governing the increase in memory decay influence modeling student performance. We consider two methods for modeling memory decay: one that uses the practice order (as in [12, 10]), and a second that uses a time window approach. Incorporating these concepts is important for tracing knowledge as it allows for the differentiation between a student's current knowledge versus previous comprehension that may have been lost over time.

In Section 2 we discuss the memory decay and spacing effect theories, and describe the well-known knowledge tracing algorithm, Performance Factor Analysis (PFA) [28]. Section 3 contains Related Work investigating issues surrounding memory within knowledge tracing. In Section 4 we introduce a new algorithm, *MemDec*, and a variant, *MemDec Spacing,* that incorporates the spacing effect. Section 5 presents the real-world dataset used for the experiments. Analysis and experimental results are presented in Section 6. Finally, conclusions and final remarks are discussed in Section 7.

## 2. KNOWLEDGE TRACING AND MEMORY DECAY

### 2.1 Memory Decay and Spacing Effect
Psychological effects influence student learning in classrooms and virtual learning environments. Hence, the consideration of cognitive science principles has become crucial for analyzing student performance in educational systems [24]. One such principle is the *decay theory*, a principle of forgetting which states that memory fades with the passage of time. Unless reinforced by repetition, the information we learn is forgotten over time. Without incorporating decay into knowledge estimation, it is difficult to differentiate between current knowledge (i.e. "student knows") and past knowledge (i.e. "student knew"). Therefore, when estimating a student's mastery of a skill, it is important to consider the possibility of memory decay.

Memory decay has been widely studied by cognitive scientists. Some researchers have attempted to capture the rate of time-dependent memory decay in an effort to measure the rate of forgetting [14]. Many studies have examined decay in short-term memory

over time [32] while others look specifically at the effects of a lack of repetition over time on long-term memory [38].

Memory decay has additionally been studied with applications to student learning. In MCM, memory decay is incorporated through the power function in which each item-specific memory trace decays exponentially [17]. This model was then used to determine optimal practices to yield the highest retention of material in a classroom. Similarly, a study was conducted to measure and compare the relearning of forgotten material by three computational models, all of which incorporate a component of decay over time in its prediction equation [36].

A variety of projects in the learning sciences community have shown that a range of pedagogical techniques could improve the retention of learned concepts [20]. Applying such teaching practices to a variety of online contexts might help improve memory retention. One such practice is a cognitive phenomenon frequently studied by learning scientists called the *spacing effect*, or distributed learning. This is based on the observation that concepts which are practiced in a distributed schedule over time tend to be better retained than those taught within so-called massed schedules, in which practice attempts related to the same set of skills are reviewed in quick succession. Research in this area has demonstrated that both time elapsed between practices of the same material and time elapsed between final study episode and an exam affect final-test retention [39].

Additional work has proposed ways to determine optimal spacing methods, such as [23, 26] which introduced a model that can predict the influence of specific study schedules on retention for specific items. Other research has focused on studying this effect in specific learning domains such as the effect of time gaps on retention of foreign vocabulary, science, and music [3, 15, 35, 16]. In total, there is considerable evidence for the importance of spacing for long-term retention of knowledge. However, there is a very limited amount of work that incorporates spaced learning into predicting student outcomes via knowledge tracing. To combat this shortcoming, we extend our memory decay algorithm to capture the spacing effect in knowledge estimation. We also analyze how these models are influenced when using different methods of enabling memory decay, through altering the rate and strength of decay over time.

## 2.2 Performance Factor Analysis

Performance Factor Analysis (PFA) is a knowledge tracing algorithm [28] that can be used with multi-skill items, in contrast to some of its extensions which only work for single-skill items [10]. Our proposed algorithm and experiments use PFA as a baseline model.

The original PFA model predicts the performance of a student on a given item/problem, at a given time. It does this by using the student's past number of successes, multiplied by a weight $\gamma$ fit to each of the item's skills; a student's past number of failures, multiplied by a weight $\rho$ fit to each of the item's skills; a weight $\beta$ which represents the difficulty of an item. Depending on the variant of PFA, $\beta$ is either applied across all contexts, across all items linked to the current skill (the most common approach and what we will use here), across all items of the same "item-type", or for individual items**.** These features are inputted into a logistic function to obtain a prediction, p(m), which gives the probability of success for a given student on a given (future) item. The PFA formula is given below:

$$m(i;\ j \in KC; s;\ f) = \sum_{j \in KC}(\beta_j + \gamma_j s_{i,j} + \rho_j f_{i,j})$$

$$p(m) = \frac{1}{1 + e^{-m}}$$

where $i$ represents a student, $KC$ are the knowledge components (i.e skills) linked to the item, $j$ represents a skill. Parameters $\beta_j$, $\gamma_j$ and $\rho_j$ are the learned parameters for skill $j$. Throughout this paper, the terms "skill" and "knowledge component" are used interchangeably.

It is evident from the formula that the PFA model in its original form does not incorporate the notion of memory decay. All previous practices are given the same weight, regardless of the time or order in which they took place.

Some previous studies have incorporated the notion of memory decay into their knowledge tracing models. Doing so addresses the phenomenon that memory fades with the passage of time. Ignoring decay may be temporarily safe when practice sessions are massed (as in some intelligent tutoring systems) but will lead to less accurate inference when the student's work on a skill is spread out over time.

# 3. RELATED WORK

A variety of knowledge tracing models have been proposed and studied, based on a range of frameworks. This includes methods based on Hidden Markov Models such as Bayesian Knowledge Tracing (BKT) [6], which models a student's latent knowledge of concepts (knowledge components) as a set of binary variables which represent mastery or lack of mastery of each concept, and neural network models such as Deep Knowledge Tracing [31], which uses RNNs (recurrent neural networks) to learn concept patterns using long-short term memory (LSTM) without human annotations. There are also recent efforts to combine knowledge tracing and Item Response Theory [25, 18] with a decay effect, modeled as elapsed time and a forgetting parameter, to improve accuracy on real world datasets.

In the last few years, extensions to these categories of models have attempted to include item recency/decay. Researchers demonstrated that by incorporating a forgetting parameter to represent recency of learning into classic BKT (which assumes that once a skill is mastered it cannot be forgotten), the algorithm is more sensitive to the effect of interspersed trials [19]. In another variant of BKT called Multistate BKT [1], the model incrementally increases the weight of newer attempts. In Attentive Knowledge Tracing [11], researchers used attention networks to draw connections between a target question and every question the learner had responded to in the past, implementing a monotonic attention mechanism that uses an exponential decay curve to down-weight past questions. In DAS3H [5], the authors modeled both learning and forgetting curves, extending factorization machines to handle multiple skills tagging. Learning Process-consistent Knowledge Tracing (LPKT) [34] models the student learning process as a set of tuples which includes the time series information of the assignments, thereby embedding both the answer time as well as the interval between activities.

Previous research investigated variations of PFA that incorporate decay into their mastery predictions. One notable model, PFA-Decay [13], took practice order into account using a decay factor $\delta$ ($0 < \delta <= 1$) raised to a power representing the distance in practice

number, and multiplied to the counts of successes and counts of failures.

A modified version of PFA called *Recent-PFA* (R-PFA) was introduced in [10]. R-PFA incorporated memory decay into the model's performance prediction in the form of a weighted proportion of success, with weight being dependent on the recency of the practice. However, R-PFA does not take time specifically into account; it just considers practice order. R-PFA modifies the PFA formula by replacing the total number of failed practices, $f_{ij}$, with the total number of all practices thus far (essentially equaling $f_{ij} + s_{ij}$), and replacing the total number of successful practices $s_{ij}$ with a component, $R_{ijt}$, that incorporates the notion of memory decay:

$$R_{ijt} = \frac{\sum_{p=-2}^{t-1} b^{(t-p)} X_{ijp}}{\sum_{p=-2}^{t-1} b^{(t-p)}}$$

where $b \in (0,1]$ represents the decay factor, and $X_{ijp}$ represents the outcome of practice $p$ (1 if successful and 0 if failed) for student $i$ on skill $j$. In the original R-PFA paper, authors used three ghost/synthetic practices [10].

Note that in the original R-PFA work each item was linked to a single skill, losing one of the main original benefits of PFA. In particular, the formula introduced in the original R-PFA paper can only handle problems coded with a single knowledge component. To the best of our knowledge, there is no previous work on extending R-PFA to create a variation which can handle data containing multi-skill items [37, 10]. Other work has been done to consider recency, including adding time-based weights to components of existing models and incorporating a weighted proportion for previous failures [12, 27].

Logistic Knowledge Tracing (LKT) [27] is a logistic regression-based framework that can enable multiple components from different existing models, such as Additive Factors Model (AFM) [2], Instructional Factors Analysis (IFA) [4], PFA [28], PFA-Decay [13], and R-PFA [10]. LKT showcased a suite of components that could be combined to form new models. Some of these components incorporate the notion of recency and decay. Two of the comparison models we implement, Alg1 and Alg2 (described further in Section 6), are based on components described in LKT. Important components from [27] that we used in our comparison models are as follows:

- Exponential decay of proportion, which uses the prior probability correct for each knowledge component, introduced as in the R-PFA model. This uses a parameter to describe the exponential rate of decay, or recency, for observations of a knowledge component. This is used in Alg2.
- Intercept for each KC/skill, which is a simple linear model intercept. Used in Alg1 and Alg2.
- Log performance (logsuc, logfail), which is the log-transformed performance factor (the total successes or failures), representing declining marginal returns, e.g. $\ln(s_{ij}) + \ln(f_{ij})$. These are used in Alg1.
- Recency, which is defined as the power log decay applied to the time interval since the previous encounter with the KC. This feature considers only the just prior observation and simulates performance improvement when the prior practice was recent. This version of recency was introduced in LKT. This is used in Alg2.

In this paper we propose a model called MemDec that captures memory decay and spacing, which is shown to outperform PFA, R-PFA, and components of LKT, while accounting for multiple skills per item. We implement and analyze two methods of incorporating decay through practice order and through time windows. This proposed algorithm goes beyond the aforementioned studies by considering the effects of spacing between practices, thus modeling the spacing effect when predicting student performance. It also studies multiple ways of inducing decay, whereas previous studies focus only on one method, mostly practice-order.

To the best of our knowledge, capturing the spacing effect in knowledge tracing has not been thoroughly studied. While a few algorithms have utilized the notion of time windows [5, 21], their time windows overlap and are defined in a more restrictive and limited manner. Additionally, [5, 21] do not consider time elapsed between practices when modeling spacing, whereas our proposed model MemDec Spacing considers the time elapsed between each current and previous practice, which we believe is crucial for modeling spacing accurately. Also, in MemDec, this is done for both practice-order and time-window variations. For MemDec and MemDec Spacing with time windows, we decay the weight of a practice based on the time window the practice falls into. Our usage of equivalent, disjoint time windows allows for consistency in the exponential rate of decay through time, offering a more simple and interpretable implementation of capturing spacing effects. Additionally, we conduct an in-depth analysis on differing values of a decay factor applied to both successful and failed practices as well as a varying number of ghost practices.

# 4. MEMDEC ALGORITHM

## 4.1 MemDec (PFA Memory Decay)

We propose a new model, MemDec, a variation of the PFA algorithm inspired by R-PFA components, that can also be seen as fitting within the LKT framework. In this approach, memory decay is applied to both successful and failed practices, and the model can be used with multi-skill items (whereas the original R-PFA formula only supports single-skill items). In addition to this, we build on the approach in [22], which splits skills into "common" or "rare" categories. This makes MemDec suitable to be applied to existing learning systems, even situations in which some skills are very rare.

When applying knowledge tracing algorithms to real educational systems it is very common to have to deal with rare skills. Such rare skills can occasionally occur in datasets for various reasons. For example, some items might be tagged with skills that represent prerequisites that are not taught in the courseware. When training PFA with datasets containing rare skills, several challenges including degenerate parameters can occur [22]. Depending on how rare some skills are, there might not be enough data points to precisely estimate parameters when training a model. In [22], authors investigate this challenge and as a remediation they propose a PFA variant that splits the skills into common and rare skills. When training a model, each common skill will learn its own set of parameters, and all rare skills will train a single common set of default parameters. Such a model was shown to improve the results and reduce the degenerate parameters [22].

Given that it is common for datasets from real educational products to have rare skills, for all the proposed models or models used for comparison, we incorporate the concept of common vs. rare skills.

The formula for MemDec is given below (note that as for PFA and R-PFA, m will be inputted into a logistic function to obtain a prediction, p(m)):

$$m(i,j,KC,RS,RF) = \sum_{j \in common\ KC} \left(\beta_j + \gamma_j RS_{ijt_{ij}} + \delta_j RF_{ijt_{ij}}\right) + \sum_{j \in rare\ KC} \left(\beta_d + \gamma_d RS_{ijt_{ij}} + \delta_d RF_{ijt_{ij}}\right)$$

$$RS_{ijt_{ij}} = \frac{\sum_{p=0}^{t_{ij}-1} b_s^{(t_{ij}-p)} X_{ijp}}{\sum_{p=0}^{t_{ij}-1} b_s^{(t_{ij}-p)}}$$

$$RF_{ijt_{ij}} = \frac{\sum_{p=0}^{t_{ij}-1} b_f^{(t_{ij}-p)} (1 - X_{ijp})}{\sum_{p=0}^{t_{ij}-1} b_f^{(t_{ij}-p)}}$$

where $i$ represents a student, $j$ represents a skill, $t_{ij}$ is the current trial (i.e. practice) number student *i* is on with skill *j*. $X_{ijp}$ represents the correctness of the practice (i.e. it is 1 if student *i*'s practice *p* with skill *j* was successful, and 0 otherwise). Constant $b_s \in (0,1]$ is the decay rate for successful practices, and constant $b_f \in (0,1]$ represents the decay factor for failed practices. RS and RF represent the recency-weighted proportions of past successes and past failures, respectively. The values for $\beta_j$, $\gamma_j$ and $\delta_j$ are parameters that are learned for each skill $j$ during the training. Parameters $\beta_d$, $\gamma_d$ and $\delta_d$ are the default parameters learned for the rare skills. Only one value is learned for each parameter for the set of rare skills.

Like R-PFA, MemDec can incorporate ghost/synthetic practices in the RS and RF formulas. To allow for ghost practices, we start $p$ from a negative number (instead of starting from 0). In the original R-PFA formula, the authors proposed three ghost practices (all failed practices). In the experimental results, similar to some components from LKT, we investigate using no ghost practices, two ghost practices (one successful and one failed), and three ghost practices (all failed).

The main differences between R-PFA and MemDec base variant is that MemDec does not use a total term, and instead it contains a component that takes into consideration the weighted proportion of failed practices, giving more weight to the recent ones. Also, MemDec can handle multi-skill items whereas R-PFA is designed for only single-skill items, losing one of the main original benefits of PFA. Additional, R-PFA only considered a practice order approach, whereas as shown below, MemDec also has a variant that considers a time window approach.

In most models that incorporate the notion of decay, the order of practices plays an important role. Every time the student completes a new practice, the model introduces more decay to previous practices. This means that the more practices the student has, the more decay is applied to older practices. Note that while decay is incremented by order of practice, the time elapsed between each practice does not affect the calculation of decay. This is also the case for MemDec.

Sometimes practices with a given skill might be done within seconds, minutes, or hours. Other times, practices might be separated by days, weeks, or months. It is unlikely that substantial forgetting will occur with small amounts of elapsed times, such as seconds or minutes. This could be a limitation to the practice-order approach, particularly if the amount of elapsed time can vary considerably. Thus, we propose a variation of MemDec that uses a time window instead of a practice-order approach. A time window is a constant duration of time (for example: 1 day) in which items answered within the same time window are given equal decay. This is an important factor to consider because memory decay does not occur instantaneously. Practices answered in time windows farther from that of the current practice can be expected to have decayed more than practices from more recent time windows. If for example the time window is two weeks, then the model assumes that all practices done within last two weeks have no decay, the practices done within four to two weeks ago have some decay, within six to four weeks have more decay, etc.

For this variant, the MemDec's RS and RF formulas are modified as shown below:

$$RS_{ijt_{ij}} = \frac{\sum_{p=0}^{t_{ij}-1} b_s^{timewindow(time(t_{ij})-time(p))} X_{ijp}}{\sum_{p=0}^{t_{ij}-1} b_s^{weeks(time(t_{ij})-time(p))}}$$

$$RF_{ijt_{ij}} = \frac{\sum_{p=0}^{t_{ij}-1} b_f^{timewindow(time(t_{ij})-time(p))} (1 - X_{ijp})}{\sum_{p=0}^{t_{ij}-1} b_f^{weeks(time(t_{ij})-time(p))}}$$

## 4.2 MemDec Spacing

The spacing effect has to do with the temporal distribution of practices linked to the same skill. If minimal time has elapsed between practices, the learning is said to be massed. Existing research suggests that if practices are spaced out, information is retained longer in memory [7, 29]. In this section, we investigate an extension of MemDec that incorporates the notion of the spacing effect into knowledge estimation. The MemDec model was adjusted to use $b_s$ values that are calculated based on a formula which takes into account how spaced apart practices are. The values of $b_f$ are calculated in a similar manner. In this approach, $b_s$ and $b_f$ (if not constant) become functions.

$$b_s(t_{ij}) = \begin{cases} b_{s_{min}}, & if\ no\ prev\ practices, or\ elapsed\ time = 0 \\ \min(b_{s_{min}} + \log_m(time(t_{ij}) - time(t_{ij}-1)), b_{s_{max}}), & otherwise \end{cases}$$

$$b_f(t_{ij}) = \begin{cases} b_{f_{min}}, & if\ no\ prev\ practices, or\ elapsed\ time = 0 \\ \min(b_{f_{min}} + \log_m(time(t_{ij}) - time(t_{ij}-1)), b_{f_{max}}), & otherwise \end{cases}$$

where $b_{s_{min}} \in (0,1]$ and $b_{s_{max}} \in (0,1]$ represent the min and the max values that we allow for $b_s$. Constants $b_{f_{min}} \in (0,1]$ and $b_{f_{max}} \in (0,1]$ represent the min and the max values that can be used for $b_f$. Constant $m$ is the base of the logarithm, and expression $time((t_{ij}) - time(t_{ij}-1))$ calculates the elapsed time between the current practice with skill j and the previous practice with skill j performed by student *i.* Note that if the student has no previous practices with a skill, or the elapsed time is 0 (elapsed time being 0 is impossible in theory, but can occur in real systems if timestamps are not captured at enough granularity), then we assign the minimum values.

In the experimental results, we present how MemDec Spacing performs when compared with MemDec, as well as some of the previous algorithms discussed in the Related Work section.

# 5. DATASET

For the experiments presented in this article, we used data from Reveal Math Course 1, a McGraw Hill digital math product for grade 6. The items from the assessments from this data are tagged with one or more skills.

The data we used for the experiments came from two Midwestern school districts and one Southwestern school district. One of these

school districts is within a large U.S. city where over half of students are classified as Black, around 10% of students are classified as Hispanic, and a fifth of families live under the poverty line. A second district is within a small town where around 5% of students are classified as Black, around 90% of students are classified as White, and around 10% of families live under the poverty line. A third is within a larger town where just over half of students are classified as Hispanic, just under half of students are classified as White, and about ¾ of families live under the poverty line. They all adopt the Common Core Standards which come from the NGA Center/CCSSO Authority.

Extracted data spans between August 2019 and May 2021. There are 4,363 unique items, out of which 2,009 items (representing about 46% of the total number of items) are tagged with at least two skills. The items are of different types such as multiple choice, fill in the blank, and entering equations. Overall, the dataset had 71 unique skills which were linked to the items from the dataset. For a skill to be considered common, we require (based on [22]) at least 200 students that have at least 3 practices with the skill. Out of these 71 skills, 42 were classified as common and the remaining 29 were classified as rare.

The dataset has 489,359 datapoints. Datapoints represent students' responses and their normalized scores (1 if the response is correct, 0 if incorrect). 1.25% of the datapoints contained a partially correct score, which were treated as 0 for the purposes of this analysis. For the experiments, we split our dataset into training and testing sets. We randomly selected about 20% of the students (647 students, 98,604 data points, 64 skills) for the testing set, leaving 80% of the students (2,588 students, 390,755 data points, 71 skills) for training.

# 6. EXPERIMENTAL RESULTS

For validation, we ran several experiments using the proposed approaches from this article, as well as other existing algorithms. All algorithms were implemented in Python.

For comparison reasons, we implemented the original PFA [28] with adjustments to handle rare skills as described in [22]. We call this the Baseline model. We also implemented other algorithms to benchmark against MemDec and MemDec Spacing: R-PFA [12], and two algorithms that were inspired by models from LKT [27]. We provide information on those in the Comparison Models subsection. For all models, we allowed for multi-skill items by using a summation factor across multiple skills linked to an item.

In an effort to study the differences and the effectiveness between each model, we calculated the AUC and RMSE validation metrics. Also, we present validation results for different groups of datapoints within the testing dataset: "all data" means we validated against all datapoints; "at least 1 non-default skill" means that we only used datapoints for which the item was linked to at least one common skill; "only non-default skills" means we only used datapoints whose items were linked with exclusively common skills; "at least 1 default skill" means we only used datapoints whose items were linked to at least one rare skill; and "only default skills" means datapoints with items solely tagged to rare skills.

## 6.1 Baseline Results

We trained a model that learned three parameters for each of the 42 common skills and three parameters for the rare skills. The validation results are presented in Table 1 below:

**Table 1. Baseline PFA**

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **All data** | 98604 | 0.6975 | 0.4443 |
| **At least 1 non-default skill** | 97460 | 0.6959 | 0.4446 |
| **Only non-default skills** | 95859 | 0.6952 | 0.4444 |
| **At least 1 default skill** | 2745 | 0.7554 | 0.4417 |
| **Only default skills** | 1144 | 0.8083 | 0.4163 |

## 6.2 MemDec Models Results

To study the difference between methods used to govern the increase in decay, we implemented and tested a variation of MemDec that used the practice-order approach, as well as a variation that used the time-window method. For the decay factors $\boldsymbol{b_s}$ and $\boldsymbol{b_f}$ we tried several combinations of values from (0,1]. While other combinations gave similar results, the best were obtained with $\boldsymbol{b_s}$= 0.6 and $\boldsymbol{b_f = 0.7}$, which we present below:

**Table 2. MemDec, with practice-order ($\boldsymbol{b_s}$= 0.6, $\boldsymbol{b_f = 0.7}$)**

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **All data** | 98604 | 0.7679 | 0.4076 |
| **At least 1 non-default skill** | 97460 | 0.7675 | 0.4074 |
| **Only non-default skills** | 95859 | 0.7677 | 0.407 |
| **At least 1 default skill** | 2745 | 0.7628 | 0.431 |
| **Only default skill** | 1144 | 0.8077 | 0.4244 |

We can observe a significant improvement when compared with the Baseline model. By incorporating the notion of memory decay, MemDec achieved an AUC of about 0.77 on all testing datapoints, whereas the baseline reached only an AUC of about 0.7. Significant improvements were observed in all other categories of datapoints, except for categories involving default skills, for which the two models achieved similar performance. This finding is expected, because rare skills do not contain enough datapoints in the dataset for our model to substantially learn from.

We also ran experiments with a time-window of 14-day, 7-day, 2-day, and 1-day. The best results were observed for a 1-day time window with an AUC equaling 0.756 across all datapoints from the testing set, which is slightly lower than the AUC of the model that uses practice-order.

The 2-day window model obtained an AUC of 0.754 for all data, the 7-day window an AUC equaling 0.749, and the 14-day window model an AUC of 0.747. This demonstrated that for this dataset, the model that uses the practice-order approach performs slightly better. For the time-window variation we observed that the smaller the

window, the better the results.

**Table 3. MemDec, with time-window ($b_s$= 0.6, $b_f = 0.7$), 1-day time-window**

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **All data** | 98604 | 0.7561 | 0.4126 |
| **At least 1 non-default skill** | 97460 | 0.7557 | 0.4124 |
| **Only non-default skills** | 95859 | 0.7557 | 0.4119 |
| **At least 1 default skill** | 2745 | 0.7564 | 0.4346 |
| **Only default skills** | 1144 | 0.8053 | 0.4287 |

**Table 4. MemDec, with time-window ($b_s$= 0.6, $b_f = 0.7$), 2-day time-window**

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **All data** | 98604 | 0.7541 | 0.4134 |
| **At least 1 non-default skill** | 97460 | 0.7536 | 0.4132 |
| **Only non-default skills** | 95859 | 0.7536 | 0.4127 |
| **At least 1 default skill** | 2745 | 0.7552 | 0.4343 |
| **Only default skills** | 1144 | 0.8018 | 0.4294 |

**Table 5. MemDec, with time-window ($b_s$= 0.6, $b_f = 0.7$), 7-day time-window**

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **All data** | 98604 | 0.7494 | 0.4155 |
| **At least 1 non-default skill** | 97460 | 0.7489 | 0.4153 |
| **Only non-default skills** | 95859 | 0.7488 | 0.4149 |
| **At least 1 default skill** | 2745 | 0.7552 | 0.435 |
| **Only default skills** | 1144 | 0.7991 | 0.4308 |

**Table 6. MemDec, with time-window ($b_s$= 0.6, $b_f = 0.7$), 14-day time-window**

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **All data** | 98604 | 0.7468 | 0.4166 |
| **At least 1 non-default skill** | 97460 | 0.7462 | 0.4164 |
| **Only non-default skills** | 95859 | 0.746 | 0.416 |
| **At least 1 default skill** | 2745 | 0.7556 | 0.4357 |
| **Only default skills** | 1144 | 0.7992 | 0.4311 |

To study whether modeling a combination of both decay and spacing could further improve the predictions, we ran experiments with the MemDec Spacing model. We experimented with different values for the hyperparameters which represents the lower and upper bounds of the decay factor, and for practice order we obtained very similar results compared to the non-spacing practice order MemDec models. One of our best performing MemDec Spacing models, presented in Table 7, gave a slightly better result than MemDec without spacing, with an overall AUC of 0.768.

For the time window of 1-day, the AUC was 0.756, which is slightly lower than the AUC when using practice-order. Overall, with time windows, MemDec Spacing gave slightly poorer results than the MemDec model. This finding may be due to certain properties of the dataset we use. Many datapoints are not spaced apart more than a few seconds in time, which would cause incorporating the effects of spacing (through time windows) to have negligible effects on the model's calculation of student knowledge. When using practice order to represent decay, the effect of spacing seems to be negligible. It is possible that if a different dataset with more widely spaced practices is used, the effect of spacing on MemDec with practice order might be more beneficial. In Section 7 we will discuss the interpretation of these results further.

**Table 7. MemDec Spacing, with practice-order ($b_{s_{min}}$ = 0.55, $b_{s_{max}}$ = 0.65, $b_{f_{min}} = 0.7$, $b_{f_{max}} = 0.7$)**

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **All data** | 98604 | 0.768 | 0.4076 |
| **At least 1 non-default skill** | 97460 | 0.7675 | 0.4074 |
| **Only non-default skills** | 95859 | 0.7678 | 0.407 |
| **At least 1 default skill** | 2745 | 0.7635 | 0.4307 |
| **Only default skills** | 1144 | 0.8087 | 0.4245 |

**Table 8. MemDec Spacing, with time-window ($b_{s_{min}}$ = 0.55, $b_{s_{max}}$ = 0.65, $b_{f_{min}} = 0.7$, $b_{f_{max}} = 0.7$), 1-day time-window**

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **All data** | 98604 | 0.7558 | 0.4127 |
| **At least 1 non-default skill** | 97460 | 0.7554 | 0.4125 |
| **Only non-default skills** | 95859 | 0.7554 | 0.4121 |
| **At least 1 default skill** | 2745 | 0.7569 | 0.4347 |
| **Only default skills** | 1144 | 0.8038 | 0.4291 |

**Table 9. MemDec Spacing, with time-window ($b_{s_{min}}$ = 0.55, $b_{s_{max}}$ = 0.65, $b_{f_{min}} = 0.7, b_{f_{max}} = 0.7$), 7-day time-window**

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **All data** | 98604 | 0.7492 | 0.4156 |
| **At least 1 non-default skill** | 97460 | 0.7487 | 0.4155 |
| **Only non-default skills** | 95859 | 0.7485 | 0.4151 |
| **At least 1 default skill** | 2745 | 0.7559 | 0.4351 |
| **Only default skills** | 1144 | 0.7992 | 0.431 |

So far, it appears that both MemDec and MemDec Spacing achieved significantly better performance compared to baseline PFA. We also provided an analysis on several variations that used different ways of representing decay, with or without the spacing effect being applied to the model. Next, we want to look at how MemDec variations perform when compared to other existing algorithms.

## 6.3 Comparison Models

We compare our models with R-PFA [12], as well as two algorithms that incorporate components from LKT [27], which we call *Alg1* and *Alg2*. All models were implemented in Python. The two algorithms are described below.

Alg1:

$$\boldsymbol{m}(\boldsymbol{i};\ \boldsymbol{j} \in \boldsymbol{KC};\boldsymbol{s};\ \boldsymbol{f}) = \sum_{\boldsymbol{j}\in\boldsymbol{KC}} \left(\boldsymbol{\beta_j} + \boldsymbol{\gamma_j}\boldsymbol{\ln}\left(\boldsymbol{s_{i,j}}\right) + \boldsymbol{\rho_j}\boldsymbol{\ln}\left(\boldsymbol{f_{i,j}}\right) + \boldsymbol{\alpha_j}\boldsymbol{t_{i,j}^{d}}\right)$$

Alg2:

$$\boldsymbol{m}(\boldsymbol{i};\ \boldsymbol{j} \in \boldsymbol{KC};\boldsymbol{s};\ \boldsymbol{f}) = \sum_{\boldsymbol{j}\in\boldsymbol{KC}} \left(\boldsymbol{\beta_j} + \boldsymbol{\gamma_j}\ \frac{\sum_{\boldsymbol{p}=-\boldsymbol{2}}^{\boldsymbol{t-1}} \boldsymbol{b}^{(\boldsymbol{t-p})}\ \boldsymbol{X_{ijp}}}{\sum_{\boldsymbol{p}=-\boldsymbol{2}}^{\boldsymbol{t-1}} \boldsymbol{b}^{(\boldsymbol{t-p})}}\right)$$

The *m* function of each model is inputted into the sigmoid function, to get a probability value between 0 and 1.

Alg1 contains a recency component that captures the elapsed time (*t*) between current and previous practice of the student *i* with skill *j* raised to a decay factor *d*. It also takes the natural logarithm of the number of successes $\boldsymbol{s_{i,j}}$ and number of failures $\boldsymbol{f_{i,j}}$.

Alg2 contains a component that uses a weighted proportion of previous practices along with a parameter *b* that represents the exponential rate of decay. Alg2 contains either two or three ghost practices.

Because many of these models used two (1 failed, 1 successful) ghost practices, or three (3 failed) ghost practices, we also implemented and ran experiments with MemDec and MemDec Spacing using this combination of two or three ghost practices. We present the validation results for all of these models below.

For MemDec, the presence of ghost practices had a negligible influence on the results. Tables that display the results for both two and three ghost practices are provided below.

**Table 10. MemDec, with practice-order ($b_s$= 0.6, $b_f = 0.7$), 2 ghost practices (1 success, 1 fail)**

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **All data** | 98604 | 0.765 | 0.4094 |
| **At least 1 non-default skill** | 97460 | 0.7644 | 0.4093 |
| **Only non-default skills** | 95859 | 0.7646 | 0.4088 |
| **At least 1 default skill** | 2745 | 0.7661 | 0.4301 |
| **Only default skills** | 1144 | 0.8315 | 0.4207 |

**Table 11. MemDec, with practice-order ($b_s$= 0.6, $b_f = 0.7$), 3 ghost practices (3 fail)**

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **All data** | 98604 | 0.7517 | 0.4159 |
| **At least 1 non-default skill** | 97460 | 0.7508 | 0.4159 |
| **Only non-default skills** | 95859 | 0.7511 | 0.4153 |
| **At least 1 default skill** | 2745 | 0.7545 | 0.4359 |
| **Only default skills** | 1144 | 0.8196 | 0.4142 |

The results of R-PFA with 3 ghosts (all failure), as given in its original paper, are given below:

**Table 12. R-PFA ($b_s$= 0.6), 3 ghost practices (3 fail)**

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **All data** | 98604 | 0.6065 | 0.4639 |
| **At least 1 non-default skill** | 97460 | 0.6051 | 0.4637 |
| **Only non-default skills** | 95859 | 0.6069 | 0.463 |
| **At least 1 default skill** | 2745 | 0.574 | 0.492 |
| **Only default skills** | 1144 | 0.7614 | 0.4802 |

The results of R-PFA with 2 ghost practices (1 successful and 1 failure) are given below:

**Table 13. R-PFA ($b_s$= 0.6), 2 ghost practices (1 success, 1 fail)**

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **All data** | 98604 | 0.6067 | 0.4638 |
| **At least 1 non-default skill** | 97460 | 0.6053 | 0.4636 |

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **Only non-default skills** | 95859 | 0.6071 | 0.463 |
| **At least 1 default skill** | 2745 | 0.574 | 0.492 |
| **Only default skills** | 1144 | 0.7614 | 0.4802 |

We also experimented with R-PFA with no ghost practices, and the results are very similar to the versions with two and three ghosts:

**Table 14. R-PFA ($b_s$= 0.6), 0 ghost practices**

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **All data** | 98604 | 0.6068 | 0.4638 |
| **At least 1 non-default skill** | 97460 | 0.6053 | 0.4636 |
| **Only non-default skills** | 95859 | 0.6071 | 0.463 |
| **At least 1 default skill** | 2745 | 0.5743 | 0.4919 |
| **Only default skills** | 1144 | 0.7615 | 0.4801 |

These findings show that MemDec significantly outperformed R-PFA, regardless of the number of ghost practices. Similarly to MemDec, the number of ghost practices did not seem to have a significant influence on the R-PFA results.

Alg1 performed better than R-PFA, with an AUC of 0.7234 for all test data points, but worse than MemDec and MemDec Spacing. The results for Alg1 for all categories of datapoints are given below:

**Table 15. Alg1, 0 ghost practices**

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **All data** | 98604 | 0.7235 | 0.4258 |
| **At least 1 non-default skill** | 97460 | 0.7223 | 0.4258 |
| **Only non-default skills** | 95859 | 0.7224 | 0.4253 |
| **At least 1 default skill** | 2745 | 0.7492 | 0.4438 |
| **Only default skills** | 1144 | 0.8234 | 0.428 |

Alg2, with both two and three ghost practices, performed better than Alg1, with an overall AUC of 0.747 for two ghost practices and 0.739 for the model with three ghost practices. These results are still worse than MemDec and MemDec Spacing.

**Table 16. Alg2 ($b_s$= 0.6), 2 ghost practices (1 success, 1 fail)**

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **All data** | 98604 | 0.7471 | 0.4208 |
| **At least 1 non-default skill** | 97460 | 0.7469 | 0.4205 |
| **Only non-default skills** | 95859 | 0.7476 | 0.4199 |
| **At least 1 default skill** | 2745 | 0.735 | 0.4509 |
| **Only default skills** | 1144 | 0.8294 | 0.4489 |

**Table 17. Alg2 ($b_s$= 0.6), 3 ghost practices (3 fail)**

| Category | # of Data Points | AUC | RMSE |
|---|---|---|---|
| **All data** | 98604 | 0.7395 | 0.4227 |
| **At least 1 non-default skill** | 97460 | 0.7387 | 0.4226 |
| **Only non-default skills** | 95859 | 0.7395 | 0.422 |
| **At least 1 default skill** | 2745 | 0.7279 | 0.4485 |
| **Only default skills** | 1144 | 0.8191 | 0.4355 |

Overall, MemDec and MemDec Spacing outperformed all other models implemented in this study, including PFA, R-PFA, Alg1, and Alg2. We find that the practice-order variation of MemDec Spacing and MemDec provided the best predictions, with a minimal higher performance seen in MemDec Spacing. Both were followed by the time-window MemDec variation with a slightly more significant difference.

While MemDec Spacing with time-window was outperformed by MemDec with time-window, it was still more effective than any other tested models in this experiment. The practice-order model was able to estimate student knowledge much more accurately than Baseline PFA or R-PFA. Additionally, within the MemDec variants, practice-order models were more effective than time-window models, and ghost practices had a negligible effect on performance predictions.

# 7. DISCUSSION AND CONCLUSIONS

In this work we studied the cognitive science concepts of memory decay and the spacing effect in the context of variants on Logistic Knowledge Tracing, a knowledge tracing framework. We created a new algorithm called MemDec which expands on the R-PFA components, a variation of PFA that incorporates decay. Despite the early emphasis on multi-skill items being a strength of PFA [28], to the best of our knowledge, there is no previous work on R-PFA or other time-involved extensions that looked at data containing multi-skill items [37, 10], although components have been introduced that can handle multi-skill items [27].

We further expanded MemDec to capture the spacing effect in a model called MemDec Spacing. Our new algorithms were able to handle multi-skill items, unlike the existing R-PFA model which can only handle items coded with a single knowledge component. Thus, MemDec and MemDec Spacing are more applicable to real-world educational systems in which items are often associated with multiple skills.

We also studied different ways of enabling and increasing decay through either the order of practices (practice-order) or by intervals of time elapsed between practices (time-window). To the best of our knowledge previous extensions of PFA and LKT components were mostly focused on a practice-order approach. We tested whether different values of the decay factor led to improved model predictions, in all variations. To measure effectiveness, our new algorithms were compared against two comparison algorithms based on LKT components, Alg1 and Alg2, as well as against PFA and R-PFA. The results of this study showed that MemDec and MemDec Spacing outperformed all other comparison models.

Practice-order MemDec variations showed better results than time-window variations. We investigated different time window sizes, from 1-day to 14-day windows, and the experiments showed that the smaller the time-window, the better the results. More specifically, the 1-day time window produced the best results. However, these findings may be due to the relatively massed nature of the dataset, in which many practices are spaced apart by small time intervals, potentially causing the time-window approach to have a smaller effect than the order of practices.

Also, the study shows that modeling decay with the spacing effect did not seem to provide an advantage over solely modeling decay. For practice-order, MemDec Spacing exhibited a slight advantage over MemDec. For time-window, MemDec outperformed MemDec Spacing by a small increase in performance across all time windows tested. It is possible that this finding may also be due to the relatively massed nature of the non-spaced dataset, as mentioned above.

Therefore, it may be valuable for future work to compare these models against datasets containing more spaced items, to determine whether the time-window approach could be beneficial over practice-order in datasets where practices are spaced out. This would also show whether incorporating the spacing effect along with decay can have a high positive impact on predicting student performance on such datasets.

Another area of future work following from this paper involves looking into how well the approaches presented perform at predicting retention long-term, including on standardized examinations [9, 33]. Finally, future work in this area may benefit from going beyond simply assessing predictive goodness to assessing the practical implications of when instructors are told a student has mastered a skill, when in fact they have forgotten it.

Overall, the fact that MemDec and MemDec Spacing outperformed the other models highlights the importance of capturing cognitive science principles such as memory decay and spacing when modeling student knowledge and predicting future performance. The analysis conducted also showcases the difference in model performance between increasing decay by either order or through time. The results show that the proposed models are suitable knowledge tracing approaches for real-world adaptive learning systems with multi-skill items, where the real possibility of students forgetting skills can significantly impact the results.

# 8. REFERENCES


[1] Argawal, D., Baker, R.S., and Muraleedharan, A. 2020. Dynamic knowledge tracing through data driven recency weights. *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*, pp. 725 – 729.

[2] Cen H., Koedinger K., and Junker B. 2006. Learning factors analysis – A general method for cognitive model evaluation and improvement. *Intelligent Tutoring Systems: 8th Intl. Conference (ITS 2006),* pages 164–175, Berlin, Heidelberg. Springer Berlin Heidelberg.

[3] Cepeda, N.J., Coburn, N., Rohrer, D., Wixted, J.T., Mozer, M.C., and Pashler, H. 2009. Optimizing Distributed Practice – Theoretical Analysis and Practical Implications. *Experimental Psychology* 56(4):236–246. DOI: 10.1027/1618-3169.56.4.236.

[4] Chi, M., Koedinger, K. R., Gordon, G., Jordan, P., and VanLehn, K. 2011. Instructional Factors Analysis: A cognitive model for multiple instructional interventions. In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, & J. Stamper (Eds.), *4th International Conference on Educational Data Mining, Eindhoven, The Netherlands,* (pp. 61-70).

[5] Choffin, B., Popineau, F., Bourda, Y., and Vie, J.J. 2019. DAS3H: Modeling Student Learning and Forgetting for Optimally Scheduling Distributed Practice of Skills. *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019).* arXiv:1905.06873v1.

[6] Corbett, A.T., Anderson, J.R. 1995. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction* 4:253-278. Kluwer Academic Publishes, The Netherlands.

[7] Dempster, F. N. and Farris, R. 1990. The spacing effect: Research and practice. *Journal of Research & Development in Education, 23*(2), 97–101.

[8] Ding, X. and Larson, E. C., 2021. On the Interpretability of Deep Learning Based Models for Knowledge Tracing. *Association for the Advancement of Artificial Intelligence*.

[9] Feng, M., Heffernan, N.T., and Koedinger, K.R. 2006. Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required. In Ikeda, Ashley & Chan (Eds.). *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*. Springer-Verlag: Berlin. pp. 31-40.

[10] Galyardt, A. and Goldin, I. 2014. Recent-performance factors analysis. In *7th Int. Conf. Educational Data Mining*, J. Stamper, Z. A. Pardos, M. Mavrikis, and B. McLaren, Eds., London, United Kingdom, Jul. 4–7, 2014, pp. 411–412.

[11] Ghosh, A., Heffernan, N., and Lan, A.S. 2020. Context-Aware Attentive Knowledge Tracing. *KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* arXiv:2007.12324v1.

[12] Goldin, I.M. and Galyardt, A. 2015. Convergent validity of a student model: Recent-Performance Factors Analysis. In *Proceedings of 8th International Conference on Educational Data Mining*. Madrid, Spain.

[13] Gong Y., Beck J. E., and Heffernan N. T. 2011. How to construct more accurate student models: Comparing and optimizing knowledge tracing and performance factor analysis. *Int. J. Artif. Intell. Educ.,* vol. 21, pp. 27–46, Jan. 2011, http://doi:10.3233/JAI-2011-016.

[14] Jenkins, P., Earle-Richardson, G., Slingerland, D. T., and May, J. 2002. Time dependent memory decay. *American Journal of Industrial Medicine*, *41*(2), 98–101. https://doi.org/10.1002/ajim.10035.

[15] Kaipa, R., Kaipa, R., and Keithly, A. 2022. The role of lag effect in distributed practice on learning novel vocabulary.

*Logopedics, Phoniatrics, Vocology.* DOI: 10.1080/14015439.2021.2022197.

[16] Katz, J.K, Ando, M., and Wiseheart, M. 2021. Optimizing song retention through the spacing effect. *Cognitive Research: Principles and Implications* 6(1)79. DOI: 10.1080/09658211.2011.631550.

[17] Khajah, M. M., Lindsey, R. V., and 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.

[18] Khajah, M., Huang, Y., Gonzalez-Brenes, J.P., Mozer, M.C., and Brusilovsk, M.C. 2014. Integrating Knowledge Tracing and Item Response Theory: A Tale of Two Frameworks. *Proceedings of Workshop on Personalization Approaches in Learning Environments (PALE2014) at the 22nd International Conference on User Modeling, Adaptation, and Personalization*, pp. 7–12, 2014.

[19] Khajah, M., Lindsey, R.V., and Mozer, M. 2016. How Deep is Knowledge Tracing?. *Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016).* arXiv:1604.02416v2.

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

[21] Lindsey, R.V., Shroyer, J.D., Pashler, H., and Mozer, M.C. 2014. Improving students' long-term knowledge retention through personalized review. *Psychological science* 25(3), 639–647.

[22] Maier, C., Baker, R.S., and Stalzer, S. 2021. Challenges to Applying Performance Factor Analysis in Existing Learning Systems. *Proceedings of the 29th International Conference on Computers in Education.* Asia-Pacific Society for Computers in Education.

[23] Mozer, M.C., Pashler, H., Wiseheart, M., Lindesey, R.A., and Vul, E. 2009. Predicting the Optimal Spacing of Study: A Multiscale Context Model of Memory. *23rd Annual Conference on Neural Information Processing Systems*. Advances in Neural Information Processing Systems 22. Vancouver, British Columbia, Canada.

[24] National Academies of Sciences, Engineering, and Medicine. 2018. How people learn II: Learners, contexts, and cultures. *National Academies Press*.

[25] Oeda, S. and Asai, K. 2016. Student Modeling Method Integrating Knowledge Tracing and IRT with Decay Effect. In *EKM@ EKAW,* (pp. 19–26).

[26] Pavlik, P. and Anderson, J. 2008. Using a Model to Compute the Optimal Schedule of Practice. *Journal of Experimental Psychology Applied* 14(2):101-17. DOI: 10.1037/1076-898X.14.2.101.

[27] Pavlik, P., Eglington, L., and Harrell-Williams, L. 2021. Logistic knowledge tracing: A constrained framework for learner modeling. *IEEE Transactions on Learning Technologies*, *14*(5), 624–639. https://doi.org/10.1109/tlt.2021.3128569.

[28] Pavlik, P.I., Cen, H., and Koedinger, K.R. 2009. Performance Factors Analysis – A New Alternative to Knowledge Tracing. *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 531-538). Amsterdam, The Netherlands: IOS Press.

[29] Pavlik, P.I., Jr. and Anderson, J.R. 2005. Practice and Forgetting Effects on Vocabulary Memory: An Activation-Based Model of the Spacing Effect. *Cognitive Science*, 29: 559-586. https://doi.org/10.1207/s15516709cog0000_14.

[30] Pelánek, R. 2017. Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Modeling and User-Adapted Interaction, 27*, 313-350.

[31] Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J., and Sohl-Dickstein, J. 2015. Deep Knowledge Tracing. *Advances in Neural Information Processing Systems* 28:505-513. Curran Associates, Inc.

[32] Reitman, J. S. 1974. Without surreptitious rehearsal, information in short-term memory decay. *Journal of Verbal Learning and Verbal Behavior*, *13*(4), 365–377. https://doi.org/10.1016/s0022-5371(74)80015-0.

[33] Ritter, S., Joshi, A., Fancsali, S., and Nixon, T. 2013. Predicting Standardized Test Scores from Cognitive Tutor Interactions. *Educational Data Mining.*

[34] Shen, S., Liu, Q., Chen, E., Huang, Z., Huang, W., Yin, Y., Su, Y., and Wang, S. 2021. Learning Process-consistent Knowledge Tracing. *KDD '21, August 14–18, 2021, Virtual Event, Singapore.* DOI:10.1145/3447548.3467237.

[35] Vlach, H.A. and Sandhofer, C.M. 2012. Distributing Learning Over Time: The Spacing Effect in Children's Acquisition and Generalization of Science Concepts. *Child Development* 84(4):1137-1133. DOI: 10.1111/j.1467-8624.2012.01781.x.

[36] Walsh, M. M. et al. 2018. Mechanisms underlying the spacing effect in learning: A comparison of three computational models. *Journal of Experimental Psychology: General*, *147*(9), 1325.

[37] Wei, H., Li, H., Xia, M., Wang, Y., and Qu, H. 2020. Predicting Student Performance in Interactive Online Question Pools Using Mouse Interaction Features. *LAK*. Frankfurt, German, 10 pages. https://doi.org/10.1145/3306307.3328180.

[38] Wickelgren, W. A. 1972. Trace resistance and the decay of long-term memory. *Journal of Mathematical Psychology*, *9*(4), 418–455. https://doi.org/10.1016/0022-2496(72)90015-6.

[39] Wiseheart, M., Pashler, H., Vul, E., Wixted, J., and Rohrer, D. 2006. Distributed Practice in Verbal Recall Tasks: A Review and Quantitative Synthesis. *Psychological Bulletin* 132(3):354-80. DOI: 10.1037/0033-2909.132.3.354.

[40] Zhang, J., Shi, X., King, I., and Yeung, D.Y. 2017. Dynamic Key Value Memory Networks for Knowledge Tracing. *International World Wide Web Conference Committee.* arXiv:1611.08108.