The Mobile Fact and Concept Textbook System (MoFaCTS) Computational Model and Scheduling System

Philip I. Pavlik Jr. and Luke G. Eglington

Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152, USA
ppavlik@memphis.edu, lgglngtn@memphis.edu

Abstract. An intelligent textbook may be defined as an interaction layer between the text and the student, helping the student master the content in the text. The Mobile Fact and Concept Training System (MoFaCTS) is an adaptive instructional system for simple content that has been developed into an interaction layer to mediate textbook instruction and so is being transformed into the Mobile Fact and Concept Textbook System (MoFaCTS). In this paper, we document the several terms of the logistic regression model we use to track performance adaptively. We then examine the contribution of each component of our model when it is fit to 4 semesters of Anatomy and Physiology course practice data. Following this documentation of the model, we explain how it is applied in the MoFaCTS system to schedule performance by targeting practice for each item at an optimal efficiency threshold.

Keywords: intelligent tutoring systems, e-learning, instructional design, cloze, models of learning, adaptive scheduling

1 Introduction

Many adaptive learning systems (ALS) are inspired by the idea of personalizing the selection of practice items for a student. However, existing ALS rarely quantify student-level individual differences (e.g., learning rate) or consider efficiency when sequencing practice, so the potential of adaptive practice selection is rarely achieved in practice. Since the prominence of behaviorism, there has been an interest in using automated methods to sequence practice items to help students learn (e.g., [1]). Skinner advocated the construction of sequences to promote error-free learning through a content domain. Advocates expanded upon these ideas to produce systems with adaptive branching depending on student responses (e.g., [2]). The transition to the information processing approach offered rich opportunities for adaptive practice since by proposing hypothetical cognitive constructs like memories or skills; it became easier to make computational models that tracked such constructs (e.g., [1, 3, 4]. This early work was so influential that one of the largest systems with adaptive sequencing, Carnegie Learning's Cognitive Tutor series of products, currently uses a system with many similarities to earlier models (e.g., [3, 5, 6].

We are currently developing and testing an adaptive textbook system to teach community college students Anatomy and Physiology content (see Fig. 1 below) using cloze practice. Students have used our system during four semesters as a supplement to course content. The practice content is currently cloze sentences created with NLP algorithms and the course textbook [7, 8]. Over the past two years, we have iteratively refined all aspects of the system based on practice data, teacher feedback, student survey data, and system log data (e.g., student performance).

Below we describe the motivating literature of our approach, followed by our iterative process to improve our ALS to integrate these additional student variables. While we describe this work embedded in our development context, we are explicitly designing the system as a generic textbook supplement that might be applied to any textbook to help students practice that content [8].

"The __________ __________ is composed predominantly of neural tissue, but also includes blood vessels and connective tissue." (fill-in is nervous system)

"Once __________ __________ are secreted, phosphates filtered into the fluid of the renal tubule buffer them, aided by ammonia." (fill-in is hydrogen ions)

Fig. 1. Easy and hard example cloze items.

2 Review of Adaptive Learning Algorithms

Our adaptive practice system is specifically designed for learning materials containing a set of knowledge components (KCs) that are interdependent and for which multiple KCs may be required to solve an individual problem. Despite this focus on complex content, the adaptive sequencing applies to both independent items (e.g., unrelated vocabulary words) and dependent items within a KC model (e.g., practice questions nested within a concept). There are two primary novel components of our system that distinguish our approach. First, our system uses an adaptive practice algorithm, with practice chosen based on an empirically derived Optimal Efficiency Threshold (OET) [8, 9]. Second, the learner model in our system includes features that account for elapsed time between practice events and the difficulty of those events. Below we begin by describing relevant literature concerning adaptive practice algorithms and the existing adaptive learning systems within which they are used. Most of the relevant literature and existing adaptive systems involve what is essentially flashcard learning. The key distinctions among them are whether practice or not practice is scheduled adaptively (the practice algorithm) and whether decisions are made with the help of a learner model. Subsequently, we describe relevant literature on individual student and item differences and why learner models need to account for these differences.

The Pimsleur method [10] was an early attempt to leverage spacing effects into language learning practice. With this method, new vocabulary was introduced and tested with increasingly wide spacing intervals. Expanding practice intervals can be effective [11], but they may not be effective unless the interval is adaptive to the user's performance [9, 12]. With the Pimsleur method, spacing intervals are increased regardless of item difficulty or student performance. This heuristic leads to overly difficult (and inefficient) practice for some items. The Leitner system [13] offered an improved algorithm that was adaptive to student performance. Put simply, practice schedules for items practiced according to the Leitner method were increased or decreased according to whether the student was answering correctly (increase spacing) or incorrectly (decrease spacing). Decreasing spacing for harder items can improve learning [14] and has been successfully implemented in adaptive practice algorithms [9, 15, 16].

In their simplest forms, ALS do not have a model estimating knowledge; they have a decision rule dictating when a concept is sufficiently well understood and adaptive feedback triggered by incorrect answers. For example, the Assistments system considers content mastered when the student answers questions about that content correctly three times in a row [17, 18]. Relevant feedback and hints are provided if the student answers incorrectly. While this approach is superior to undifferentiated instruction and problem sequencing [19], including a model can further improve practice efficiency. A particularly prominent model-informed adaptive system is the Cognitive Tutor System. It uses a Bayesian Knowledge Tracing (BKT) model to trace the learning of mathematical skills or KCs (knowledge components), adapting practice to drop items from the practice set when they have been "mastered" according to the BKT model. BKT avoids inefficiencies that can arise from solely decision-rule-based scheduling (e.g., a student getting two in a row repeatedly but not progressing due to a 3 in a row requirement). While BKT accounts for learning and adapts to performance, it and other popular models (e.g., PFA) typically do not account for other important factors that can predict knowledge states, such as elapsed time between practice attempts, the increased predictive utility of recent vs. older attempts, and forgetting.

Recently, more advanced adaptive practice algorithms informed by psychological theories of memory have been tested. The Half-life Regression algorithm [20] and the difficulty-threshold system introduced by Eglington & Pavlik [9] produce practice scheduling similar to these methods but further improved upon compared to these earlier heuristics by scheduling practice according to predictions by a learner model inspired by psychological theories of spacing and forgetting. Lindsey et al. [12] also demonstrated how scheduling practice according to a model could provide robust learning benefits over simpler scheduling algorithms. An important detail of these approaches is that the decision rule for which item to practice next is informed by prior theory that items more likely to have been forgotten are more productive items to practice [3, 21]. Thus, more learning gains may be achieved on such trials, but those trials may also be more time-consuming than efficiency-based approaches [9, 16]. We describe these issues in detail below. Next, we describe our approach, in which we also use a theoretically driven learner model, and also account for the efficiency of practice when making pedagogical decisions with those predictions. For our adaptive learning system (ALS), we chose a logistic regression framework instead due to its greater simplicity and flexibility caused by the easy combination of multiple factors in the underlying regression, unlike methods like BKT, where the addition of additional factors greatly complexifies the optimization and implementation of the model. Using logistic regression allowed us to add features more easily with forgetting effects and spacing effects, crucial for modeling declarative memory. This flexibility also allowed us to include features to account for linguistic features of practice items and student differences in prior knowledge, learning rate, and motivation.

3 Logistic Regression Model

3.1 Student differences

Prior knowledge varies across students that use adaptive educational technology (e.g., [22]). This variability is part of the motivation for creating adaptive educational technology in the first place. Individual differences in prior knowledge are important because they determine the initial challenge of practice, which has ramifications for the efficacy of the practice and the student's subjective experience. Prior knowledge can also impact the rate at which the student can encode new information - if they already have a somewhat coherent mental model of the to-be-learned topic, they may better learn new related concepts [22]. Variability in student learning rates is an important topic and has been investigated to some extent already. Lee & Brunskill [23] demonstrated that practice scheduled by a BKT model and mastery criterion might be more efficient if the BKT model included individual student parameters. Yudelson, Koedinger, & Gordon [24] provided further evidence that accounting for student-level learning rates could improve model fits, perhaps more than student prior knowledge. Mis-estimating learning rates can also lead to systematic errors when adaptively sequencing practice [25], reducing practice efficiency. If a system consistently overestimates a students' learning, that student may be subjected to overly challenging practice content. This overestimation may have other unintended consequences on the student. For instance, overly difficult or easy tasks can lead to inattention [26]. In other words, a mismatch between the model predictions and the student can reduce a students' engagement with the task [27, 28] and perhaps ultimately increase the probability of student drop-out [29]. In sum, accounting for student-level differences is more important than simply improving model fit. Accounting for this variance can have tangible beneficial effects on the efficiency of the ALS.

3.2 Multi-level item differences

Existing ALS assume KCs vary in difficulty [30]. This difficulty can be accounted for by estimating different learning rates across KCs. One advantage of differences intrinsic to KCs is that these differences frequently generalize across students and improve prediction accuracy for new students. Additionally, most learning materials (and representative KCs) are taught repeatedly to many students. In other words, new students are more common than new content, and thus generalizing information about KCs can greatly enhance ALS predictive accuracy and usefulness. For example, in our research, we use cloze items as our primary trial type, providing learners with sentences with keywords omitted that they must fill in. We plan to eventually have difficulty measures for these items, treating them as KCs. However, currently, we do not estimate intercepts for these KCs. Instead, item/KC learning is tracked at three nested levels: the student-level, the fill-in-word (i.e., performance on prior attempts to recall that particular word across different cloze sentences), and the sentence-level (independent of which fill-in-word for the sentence is chosen).

3.3 Transitioning to a more comprehensive model

The goal of our model was not to prove cognitive theories but to apply them. Though more detailed than models like AFM, PFA, or BKT, the learning effects of our model are simpler than cognitive architectures such as ACT-R. We included the additional complexity to account for robust learning effects that these simple popular models do have mechanisms to capture. However, in many cases, we use basic mechanisms from these models and variants of these models, which we find to combine well using logistic regression as the inference method. For example, as can be seen below, we use mechanisms related to PFA variants such as R-PFA [31], PFA-Decay [32], and PFA difficulty [33], along with memory-based recency similar to ACT-R [34]. Below we describe this initial model used to track student learning and performance in Anatomy and Physiology across four semesters. Following that, we introduce new features that attempt to account for additional student and item differences. Some of these features are derived from student self-report data, while others are new tracking features designed to account for student individual differences in learning rate and prior knowledge.

3.4 Features in the model

These features are strongly informed by cognitive theories of learning, although to our knowledge have not previously been utilized together in a live adaptive learning system. The model was parameterized for the first semester using a highly detailed cloze learning experiment with basic statistics sentences (for a description of this dataset, see [34]). We refit this model using the first semester Fall data; those parameters were subsequently used for successive semesters. The choice of model terms was based on the need to produce a model that would behave consistently (parameters showing high correlation even though different) in all cases when presented with reasonable data,

Temporal recency. There is a strong effect of elapsed time since a previous practice on performance [35]. Skills and memories decay rapidly as a function of time, and the most important part of this decay is well represented by a power-law function of the times since last practice. We included recency features for both cloze-level KCs and the cloze fill-in answers themselves for the present model.

Long-term learning. While we might expect long-term forgetting, once recency is factored in and with the presence of the adaptive factors below, long-term learning is captured as permanent. Our model's long-term learning features were counts of correct and incorrect attempts, computed at the level of the cloze fill-in (KC) within-student. Counts of correctness were weighted by the difficulty of the attempt (operationalized as the predicted probability of correctness). An additional learning feature for correct counts was included in which the difficulty weighted counts were squared to represent the diminishing returns of additional correct attempts (PFA-Difficulty, [33]). This approach has been shown to perform similarly to the PFA approach. However, it has the important implication that there is an optimal level of difficulty, which is more well supported theoretically than a constant effect of practice difficulty. In other words, the implied null hypothesis in most models is that difficulty is unimportant to learning, which contravenes many research findings (e.g., [9, 16, 21]), and we think that is implausible. Weighting by predicted difficulty means that the contribution of a given attempt is specific to the particular student because the predictions are based on a model estimate of the student's difficulty using that student's unique practice history.

Tracking. As a student learns a concept, their performance should improve. Changes in performance over time can be used by computing a running average. However, recent trials are more informative of students' ability, and thus running averages should weight recent trials more heavily. Such predictors have been shown to improve model fit (e.g., R-PFA, [31]). We used a similar approach, but instead of using a ratio of correctness vs. incorrectness, we used the logarithm of the correct/incorrect ratio. This alternative approach has additional flexibility by not being restricted to a range of [0,1] and having the ability to indicate a decrease of performance expectation by using ghost success AND ghost failures [34].

Syllable hints. Providing hints to help students answer fill-in questions can increase the probability of success without reducing learning [36]. Thus providing hints may be productive in adaptive instructional systems. In order to quantify the correctness probability increase as a function of syllable hints, the model included separate intercepts for each level of hint syllable (0, 1, or 2 syllable hints).

Intercepts for each prior semester. Student aptitude, instructional approaches, and other external factors may change across semesters. To reduce the impact of these potential differences on the model, separate intercepts were included for each semester to improve model stability.

$$\begin{gathered}\text{Performance} = \\ \text{Recent performance of student} + \\ \text{Recent performance of student on the fill-in}+ \\ \text{Temporal recency of the sentence/fill-in pair} + \\ \text{Temporal recency of the fill-in} + \\ \text{Long-term effect of practice difficulty of success} + \\ \text{Long-term effect of count of failures} + \\ \text{Intercept for each of the 3 hint condition levels} + \\ \text{Intercept for each prior semester}\end{gathered} \tag{1}$$

In section 4, we look at the quantitative importance of each term to prediction. In section 5, we explore the qualitative results of the model applied to practice optimization. Equation 1 fit with $R^2 = .1093$, which is considerably lower than experimental results, but in line with other implemented systems [34]. Logistic regression using GLM showed that all predictors (except for semester intercepts) had coefficient Z-scores greater than 8. Since the model was fit for N = 93155 trials with only 16 parameters, there was no concern with overfitting within the context of our collected data, so we do not show cross-validation results for space reasons (but more complex models would certainly face this problem). To better understand the parameterization, note that the first four terms required non-linear decay parameters that were solved for using an R optimizer that recomputed the feature and the logistic regression model iteratively to arrive at optimal non-linear parameters [32, 34]. The exact values of these parameters are specific to the data and can be obtained from the authors as needed. It was interesting to note that recent memory was forgotten much more quickly for the sentence and fill-in word pair than the recency effect of the fill-in word across sentences, which persisted longer.

4 Participants and Data Screening

359 participants were included in the present analysis. Demographics surveyed from a subset of the participants (N=133) indicated that this population is approximately 87% female and has a median age of 33 years old (SD = 9.1 years). 63% of respondents reported being African American. Practice attempts were only included up to the fifth repetition for each sentence/fill-in pair. Otherwise, all data were included.

5 Model Fit

To illustrate the importance (or lack thereof) for each component of the model, we computed the subject mean R^2 differences for models that excluded 1 variable at a time from Equation 1 and for a dominance analysis used to compute the avg contribution of the parameter if used in all possible compositions of the 7 terms (excluding semester intercepts).

Table 1. Comparison of the effect of each factor.

Term	Δ McFadden's R^2 last term	Δ McFadden's R^2 avg. dom.	Δ Loglikelihood
Recent performance of student	0.0426	0.0495	2747.7
Recent performance of student on the fill-in	0.0057	0.0174	367.3
Temporal recency of the sentence/fill-in pair	0.0039	0.0061	254.4
Temporal recency of the fill-in	0.0127	0.0131	820.9
Long-term effect of practice difficulty of success	0.0025	0.0091	161.3
Long-term effect of count of failures	0.0006	0.0015	38.8
Intercept for each of the 3 hint condition levels	0.0083	0.0077	535.8

The results in Table 1 show that the importance of each component varies greatly. For example, the effect of failures on long-term learning is extremely small (slightly negative coefficient in this case indicates it is likely overfit or not significant). This term may be removed in future versions, despite some indication it was positive in our early results. This situation illustrates the complexity of the model since this term is almost certainly negative because it is multicollinear with all 4 recency terms (2 for performance and 2 for temporal recency), which means we cannot conclude that long-term learning is actually near 0 for failures. It is more likely that the recency terms capture the short-term effects of failure well enough and that the feedback effect of review may be forgotten quickly. This result agrees well with theories of forgetting that suggest that successful testing results in durable learning. In contrast, passive study from a review is forgotten more quickly (e.g. [37]).

6 Application Model to Pedagogical Optimality

Our project is guided by the assumption that there is an optimal sequence in which practice items should be practiced for a given learning task. This optimal sequence is also assumed to be unique to the student and varies dynamically according to their prior practice history and performance. Historically, this has been achieved by using prior practice history as an input to algorithms that make pedagogical decisions. For instance, Smallwood [1] used prior performance (e.g., correctness percentage, learning rate) on mathematics problems to decide whether students were ready for new (more challenging) problems, should continue on the current problem type, or perhaps needs to revisit prior content. Additionally, our system's design assumes that there is an optimal difficulty at which to practice for a given learning task. In this case, we are operationalizing difficulty as the probability of correctly answering a test question. The relevance of difficulty for practice optimality has been researched extensively in psychology (e.g. [38]). A general conclusion has been that imposing some difficulty benefits learning [39], but researchers disagree about how much difficulty to impose. Some have argued that imposing greater difficulty provides optimal learning benefits [3, 40]. One justification for this approach is known as Discrepancy Reduction Theory, in which whatever content is least known should be practiced because that item can provide the largest potential learning gains. There is truth to this idea; learning gains are higher with more difficulty [21], although the benefit is not universal (e.g.[41]). However, a frequently overlooked variable that consistently correlates with difficulty is practice time. As difficulty increases, the time to complete the task or recall the information also increases [42, 43]. There are at least two reasons for this consistent finding. First, better-learned information (in the form of skills or memories) is recalled more quickly [44, 45]. Second, as difficulty increases, so does the risk of failure. When failures occur, feedback is necessary for most practical contexts. This feedback is time-consuming and may not be necessary if the information is successfully recalled. Thus, feedback time is primarily associated with failures and can dramatically increase the time cost of practice. Note that despite costing more time, failures do not necessarily confer more learning [46]. In short, imposing more difficulty may become inefficient even if per-trial it appears to provide more learning (e.g., [9, 16]).

Considering the efficiency of pedagogical decisions can improve learning outcomes [9, 16]. Rather than schedule practice according to whichever item would provide the most gain, Pavlik & Anderson [16] instead had students practice whichever item provided the most gain per second. As a result, students’ practice time was more efficient, and they completed many more practice trials. Furthermore, students that practiced according to this algorithm had significantly higher memory retention than alternative algorithms based on discrepancy reduction theories [3] and other scheduling heuristics. In our A & P project, efficiency (utility) was determined by computing an optimal gain curve using Equation 2.

$$utility = \frac{p\Big(B_0\big(p-p^2\big)\Big)+(1-p)(B_1)}{p\big(B_3+B_4e^{-qlogis(p)}\big)+(1-p)(fixedcost)} \tag{2}$$

Confirming the efficacy of this method, Eglington and Pavlik [9] simulated student performance at various difficulties to determine an optimal difficulty empirically. Simulated student practice was estimated using a logistic regression model parameterized by fitting an existing dataset. They tested the simulation predictions by having students learn Japanese-English word pairs in which the same model scheduled practice at various difficulties. Practicing at relatively low difficulty was found to be most beneficial for memory retention in contrast with prevailing theories [20, 21]. Importantly, they found that a discrepancy reduction approach (focusing practice on more challenging items) was both significantly worse than practicing at a lower difficulty (higher efficiency) and also not significantly better than a non-adaptive control condition. It is important to note that the benefit of practicing at low difficulty is partially due to the learning materials - cloze learning trials are typically fast when answered correctly but relatively slow when incorrect due to the necessity of providing feedback. In short, the optimal difficulty may vary depending on the learning context.

In order to schedule practice efficiently, the time cost for correct vs. incorrect attempts must be known. Response time for successful recall is well fit by several models (e.g., [47, 48]). Incorrect response times are harder to model, but using median durations is effective [9]. We used prior data from students completing a similar cloze task to model correct and incorrect response times. Using this response time data, we computed the optimally efficient threshold (OET) to practice, with our goal to use that difficulty (operationalized as the correctness probability) to schedule practice within MoFaCTS. For example, if the OET were determined to be .6, then on each trial, the system would estimate the correctness probability for each potential cloze item (based on that students' practice history) and choose the item closest to .6. To compute the OET, we needed to compute learning gains as a function of correctness probability (difficulty) and divide those gains by the estimated time cost of practicing at that difficulty. We computed learning gain by fitting our logistic regression model, which estimates the long-term learning from practicing an item and being correct and practicing and being incorrect. There is a different time cost for each of these potential outcomes, the estimated trial duration for being correct or incorrect. Together, these computations give us an efficiency score (or utility) for each value of correctness probability (see Figure 2). In other words, the learning benefits of being correct or incorrect are linked to the time it takes for either outcome.

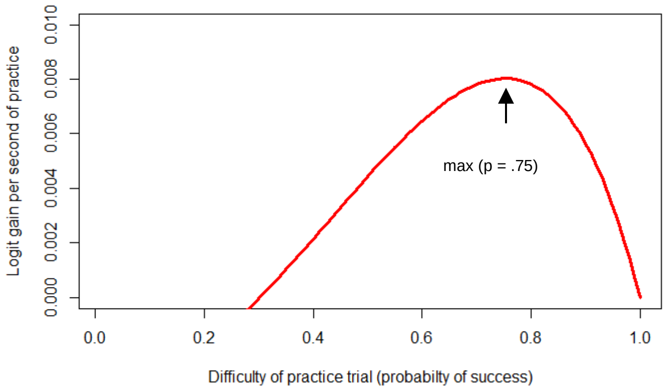
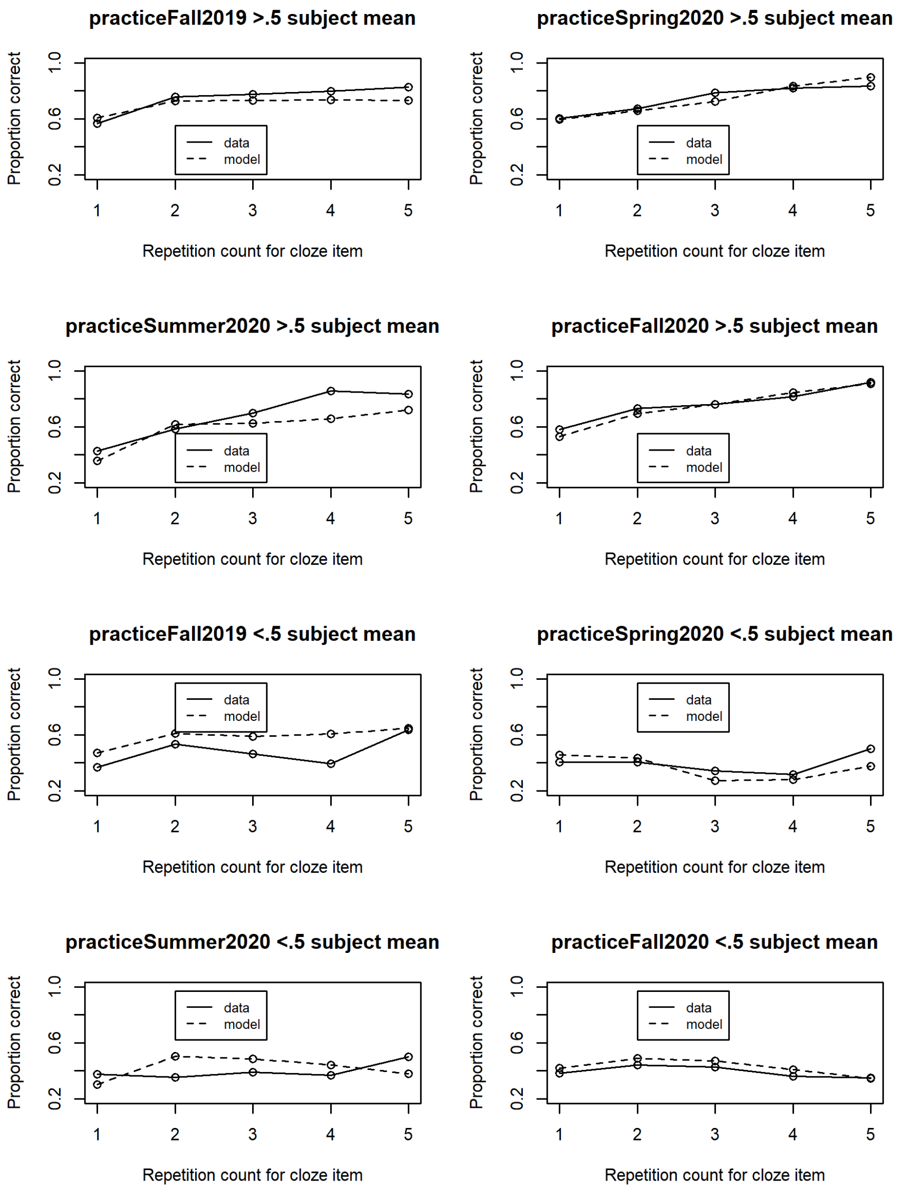


Figure 2. A visualization of computing Equation 1 to determine the optimal practice difficulty used to make pedagogical decisions in MoFaCTS. Given our current parameters, we find .75 to be the optimal level of practice.

7 Implementation

Fig. 3 Shows the fit of the model with 16 parameters characterizing the results across the four semesters for a split of the data into participants with mean performance greater than and mean performance less than 50%. We do not know the system's efficacy at this time since we have not begun efficacy testing. We see an encouraging correlation between system usage and performance of all other work in the courses, shown in Fig. 4. However, as Fig. 3 suggests, we are falling short of our goal of providing practice near the OET for participants performing at less than 50%, while participants with >50% performance are averaging near the 75% OET (which was estimated at 72% in prior semesters). These results clarify that we have improvements to make if we are to serve better the approximately half of our population (51%) that produced means less than 50%. We suspect that the problem may have been caused by our prior model, suggesting a much higher gain for failures than we found after introducing changes in the model structure after Fall 2020 (the model in this paper). Perhaps not coincidentally, Fall 2020 coincided with making the system mandatory, which likely also meant the bottom 50% mean students may have been less motivated, unlike the extra credit samples from the prior semesters.



practiceFall2019 >.5 subject mean

practiceSpring2020 >.5 subject mean

practiceSummer2020 >.5 subject mean

practiceFall2020 >.5 subject mean

practiceFall2019 <.5 subject mean

practiceSpring2020 <.5 subject mean

practiceSummer2020 <.5 subject mean

practiceFall2020 <.5 subject mean

Figure 3. Performance of the students with posthoc fit of 16 parameter model described in this paper. Above 50% and below 50% subsets. Student data was collected with various prior versions of this same model.

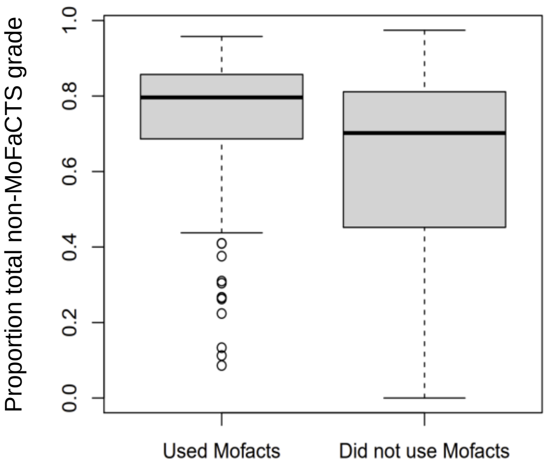


Figure 4. Encouraging correlation between usage and class performance.

8 Conclusion

Our system's scheduling algorithm extends beyond simpler scheduling flashcard paradigms (e.g., the Leitner method) and models (mastery learning with standard BKT) using a model that allows the spaced sequencing of content according to rich features of the student history. Such a practice system may be a useful component of many iTextbooks due to the importance of declarative and conceptual facts in many academic domains. In our development research, we use cloze items as our primary trial type, which entails providing learners with sentences with keywords omitted that they must fill in. Future work plans to begin including outside of practice factors from our student surveys to influence our model and reduce the cold start problem. In addition, we are currently developing new contextual and semantic features to add to the system. The contextual features include data like the class of the fill-in-response (e.g., content vs. connector word) and its importance in the content domain. Semantic features are conceptual connections between sentence items independent of the already tracked fill-in-word and represent new KCs that should influence the model.

9 Acknowledgments

This work was supported by the Institute of Education Sciences (IES; R305A190448). Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of IES. This work was also supported by the Learner Data Institute project (NSF; 1934745) and the University of Memphis Institute for Intelligent Systems.

References

1. Smallwood, R.D.: A Decision Structure for Teaching Machines. MIT Press, Cambridge (1962)
2. Borko, H.: A. A. Lumsdaine and Robert Glaser (Editors). Teaching Machines and Programmed Learning: A Source Book. Washington, D. C.: Department of Audio-Visual Instruction, National Education Association, 1960. 724 Pp. Behavioral Science: 7**,** 479-480. (1962)
3. Atkinson, R.C.: Optimizing the Learning of a Second-Language Vocabulary. Journal of experimental psychology: 96**,** 124-129. (1972)
4. Tatsuoka, K.K.: Rule Space: An Approach for Dealing with Misconceptions Based on Item Response Theory. Journal of Educational Measurement: 20**,** 345-354. (1983)
5. Skinner, B.F.: Teaching Machines. The Review of Economics and Statistics: 42**,** 189-191. (1960)
6. Corbett, A.T., Anderson, J.R.: Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. User Modeling and User-Adapted Interaction: 4**,** 253–278. (1994)
7. Olney, A.M., Pavlik, P.I., Maass, J.K.: Improving Reading Comprehension with Automatically Generated Cloze Item Practice. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.): Proceedings of Artificial Intelligence in Education: 18th International Conference, 262-273. Springer International Publishing, Wuhan, China (2017)
8. Pavlik Jr., P.I., Olney, A.M., Banker, A., Eglington, L., Yarbro, J.: The Mobile Fact and Concept Textbook System (Mofacts). 21st International Conference on Artificial Intelligence in Education (Aied 2020) Second Workshop on Intelligent Textbooks, 35–49. In CEUR workshop proceedings (Vol. 2674). (2020)
9. Eglington, L.G., Pavlik Jr, P.I.: Optimizing Practice Scheduling Requires Quantitative Tracking of Individual Item Performance. npj Science of Learning: 5**,** 15. (2020)
10. Pimsleur, P.: A Memory Schedule. The Modern Language Journal: 51**,** 73-75. (1967)
11. Yan, V.X., Schuetze, B.A., Eglington, L.G.: A Review of the Interleaving Effect: Theories and Lessons for Future Research. PsyArXiv**,** 1-39. (2020)
12. Lindsey, R.V., Shroyer, J.D., Pashler, H., Mozer, M.C.: Improving Students' Long-Term Knowledge Retention through Personalized Review. Psychological Science**,** 639-647. (2014)
13. Leitner, S.: So Lernt Man Lernen. Herder, Freiburg im Breisgau, Germany (1972)
14. Metzler-Baddeley, C., Baddeley, R.J.: Does Adaptive Training Work? Applied Cognitive Psychology: 23**,** 254-266. (2009)
15. Wozniak, P.A., Gorzelanczyk, E.J.: Optimization of Repetition Spacing in the Practice of Learning. Acta Neurobiologiae Experimentalis: 54**,** 59-62. (1994)
16. Pavlik Jr., P.I., Anderson, J.R.: Using a Model to Compute the Optimal Schedule of Practice. Journal of Experimental Psychology: Applied: 14**,** 101–117. (2008)
17. Heffernan, N.T., Heffernan, C.L.: The Assistments Ecosystem: Building a Platform That Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. International Journal of Artificial Intelligence in Education: 24**,** 470-497. (2014)
18. Kelly, K., Wang, Y., Thompson, T., Heffernan, N.: Defining Mastery: Knowledge Tracing Versus N-Consecutive Correct Responses. 8th International Conference on Educational Data Mining, 39-46. (2015)
19. Murphy, R., Roschelle, J., Feng, M., Mason, C.A.: Investigating Efficacy, Moderators and Mediators for an Online Mathematics Homework Intervention. Journal of Research on Educational Effectiveness: 13**,** 235-270. (2020)

20. Settles, B., Meeder, B.: A Trainable Spaced Repetition Model for Language Learning. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1848-1858. Berlin, Germany (2016)
21. Bjork, R.A.: In: Metcalfe, J. (ed.): Metacognition: Knowing About Knowing, 185-205. MIT Press (1994)
22. Liu, R., Koedinger, K.R.: Towards Reliable and Valid Measurement of Individualized Student Parameters. International Educational Data Mining Society. (2017)
23. Lee, J.I., Brunskill, E.: The Impact on Individualizing Student Models on Necessary Practice Opportunities. International Educational Data Mining Society. (2012)
24. Yudelson, M.V., Koedinger, K.R., Gordon, G.J.: Individualized Bayesian Knowledge Tracing Models. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.): Artificial Intelligence in Education, 171-180. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
25. Eglington, L.G., Pavlik Jr, P.I.: Systematic Prediction Errors from Individual Differences Are Inevitable and Treatable. (under review)
26. Seli, P., Risko, E.F., Smilek, D.: On the Necessity of Distinguishing between Unintentional and Intentional Mind Wandering. Psychological Science: 27**,** 685-691. (2016)
27. Kurzban, R., Duckworth, A., Kable, J.W., Myers, J.: An Opportunity Cost Model of Subjective Effort and Task Performance. Behav Brain Sci: 36**,** 661-679. (2013)
28. Seli, P., Konishi, M., Risko, E.F., Smilek, D.: The Role of Task Difficulty in Theoretical Accounts of Mind Wandering. Conscious Cogn: 65**,** 255-262. (2018)
29. Steyvers, M., Benjamin, A.S.: The Joint Contribution of Participation and Performance to Learning Functions: Exploring the Effects of Age in Large-Scale Data Sets. Behavior Research Methods: 51**,** 1531-1543. (2019)
30. Ritter, S., Harris, T.K., Nixon, T., Dickison, D., Murray, R.C., Towle, B.: Reducing the Knowledge Tracing Space. International Working Group on Educational Data Mining. (2009)
31. Galyardt, A., Goldin, I.: Move Your Lamp Post: Recent Data Reflects Learner Knowledge Better Than Older Data. Journal of Educational Data Mining: 7**,** 83-108. (2015)
32. Gong, Y., Beck, J.E., Heffernan, N.T.: How to Construct More Accurate Student Models: Comparing and Optimizing Knowledge Tracing and Performance Factor Analysis. International Journal of Artificial Intelligence in Education: 21**,** 27-46. (2011)
33. Cao, M., Pavlik Jr, P.I., Bidelman, G.M.: Incorporating Prior Practice Difficulty into Performance Factor Analysis to Model Mandarin Tone Learning. In: Lynch, C., Merceron, A., Desmarais, M., Nkambou, R. (eds.): Proceedings of the 11th International Conference on Educational Data Mining, 516-519. (2019)
34. Pavlik Jr, P.I., Eglington, L.G., Harrell-Williams, L.M.: Logistic Knowledge Tracing: A Constrained Framework for Learner Modeling. arXiv.org. (2021, preprint)
35. Wixted, J.T.: The Psychology and Neuroscience of Forgetting. Annual Review of Psychology: 55**,** 235-269. (2004)
36. Fiechter, J.L., Benjamin, A.S.: Techniques for Scaffolding Retrieval Practice: The Costs and Benefits of Adaptive Versus Diminishing Cues. Psychonomic Bulletin & Review: 26**,** 1666-1674. (2019)
37. Thompson, C.P., Wenger, S.K., Bartling, C.A.: How Recall Facilitates Subsequent Recall: A Reappraisal. Journal of Experimental Psychology: Human Learning & Memory: 4**,** 210-221. (1978)
38. Schmidt, R.A., Bjork, R.A.: New Conceptualizations of Practice: Common Principles in Three Paradigms Suggest New Concepts for Training. Psychological Science: 3**,** 207-217. (1992)

39. Koedinger, K.R., Aleven, V.: Exploring the Assistance Dilemma in Experiments with Cognitive Tutors. Educational Psychology Review: 19**,** 239-264. (2007)
40. Pashler, H., Zarow, G., Triplett, B.: Is Temporal Spacing of Tests Helpful Even When It Inflates Error Rates? Journal of Experimental Psychology: Learning, Memory, and Cognition: 29**,** 1051–1057. (2003)
41. Vaughn, K.E., Kornell, N.: How to Activate Students' Natural Desire to Test Themselves. Cognitive Research: Principles and Implications: 4**,** 35. (2019)
42. Rickard, T.C.: Bending the Power Law: A Cmpl Theory of Strategy Shifts and the Automatization of Cognitive Skills. Journal of Experimental Psychology: General: 126**,** 288-311. (1997)
43. Pyc, M.A., Rawson, K.A.: Testing the Retrieval Effort Hypothesis: Does Greater Difficulty Correctly Recalling Information Lead to Higher Levels of Memory? Journal of Memory and Language: 60**,** 437-447. (2009)
44. Peterson, L.R.: Paired-Associate Latencies after the Last Error. Psychonomic Science: 2**,** 167-168. (1965)
45. Schlag-Rey, M., Groen, G., Suppes, P.: Latencies on Last Error in Paired-Associate Learning. Psychonomic Science: 2**,** 15-16. (1965)
46. Vaughn, K.E., Hausman, H., Kornell, N.: Retrieval Attempts Enhance Learning Regardless of Time Spent Trying to Retrieve. Memory: 25**,** 298-316. (2017)
47. Anderson, J.R., Lebiere, C.: The Atomic Components of Thought. Lawrence Erlbaum Associates, Mahwah, NJ (1998)
48. Anderson, J.R., Fincham, J.M., Douglass, S.: The Role of Examples and Rules in the Acquisition of a Cognitive Skill. Journal of Experimental Psychology: Learning, Memory, & Cognition: 23**,** 932–945. (1997)