# Using Student Modeling to Estimate Student Knowledge Retention

Yutao Wang
Worcester Polytechnic Institute
100 Institute RD.
Worcester, MA, USA
yutaowang@wpi.edu

Joseph E. Beck
Worcester Polytechnic Institute
100 Institute RD.
Worcester, MA, USA
josephbeck@wpi.edu

## ABSTRACT

The goal of predicting student behavior on the immediate next action has been investigated by researchers for many years. However, a fair question is whether this research question is worth all of the attention it has received. This paper investigates predicting student performance after a delay of 5 to 10 days, to determine whether, and when, the student will *retain* the material seen. Although this change in focus sounds minor, two aspects make it interesting. First, the factors influencing retention are different than those influencing short-term performance. Specifically, we found that the number of student correct and incorrect responses were not reliable predictors of long-term performance. This result is in contrast to most student-modeling efforts on predicting performance on the next response. Second, we argue that answering the question of whether a student will retain a skill is more useful for guiding decision making of intelligent tutoring systems (ITS) than predicting correctness of next response. We introduce an architecture that identifies two research topics that are meaningful for ITS decision making. Our experiments found one feature in particular that was relevant for student retention: the number of distinct days in which a student practiced a skill. This result provides additional evidence for the spaced practice effect, and suggests our models need to be aware of features known to impact retention.

## Keywords

Educational data mining, intelligent tutoring systems, student modeling, knowledge retention.

## 1. INTRODUCTION

The field of the educational data mining (EDM) has been focusing on predict correctness of the next student response for many years (e.g. .[2]). Very little work has been done with respect to longer-term prediction. Two common approaches for student modeling are knowledge tracing [5] and performance factors analysis [10]. Both of these approaches focus on examining past student performances, and predicting whether the student's next response will be correct or incorrect. The source of power for both of these techniques is the student's pattern of correct and incorrect responses. In fact, that input is the only piece of information knowledge tracing (KT) uses (beyond which skill the problem is associated with). KT observes whether the student responds correctly or not, and uses its performance parameters, guess and slip, to update its estimate of the student's knowledge. KT takes the form of a dynamic Bayesian network, where each time slice represents an item the student is working on.

Performance factors analysis (PFA) works similarly, and keeps track of the number of correct and incorrect responses the student has made on this skill. In addition, some versions of PFA also make use of an item difficulty parameter to account for item complexity. PFA takes the form of a logistic regression model, predicts whether the student will respond to an item correctly, and estimates coefficients for the number of correct and incorrect responses that maximize predictive accuracy.

A connection with student modeling is mastery learning. In a mastery learning framework, a student continues to practice a skill until it is "mastered." The exact definition of mastery varies, but typically involves recent student performance. For example, the KT framework suggests that the probability a student knows a skill exceeds 0.95, then the student has mastered the skill. The ASSISTments project (www.assistments.org) uses a simpler heuristic of three consecutive correct responses to indicate mastery.

However, there is evidence that strictly local measures of student correctness are not sufficient. Specifically, students do not always retain what they have learned. Aside from the psychology literature (e.g. [1,4,7,8]) there has been work within student modeling that demonstrated students were likely to forget some material after a delay. Qiu et al. [12] extended the Knowledge Tracing model, to take into account that students exhibit forgetting when a day elapses between problems in the tutor.

Researchers in the ITS field are currently using short-term retention as an indicator for mastery learning. However, for a cumulative subject like mathematics, we are more concerned with the ability of the students to remember the knowledge they learned for a long period of time. Pavlik and Anderson [9] studied alternative models of practice and forgetting, and confirmed the standard spacing effect in various conditions and showed that wide spacing of practice provides increasing benefit as practice accumulates, and less forgetting afterwards as well, which is consistent with classic cognitive science results [3].

## 2. PROBLEM AND APPROACH

Although the fields of student modeling and EDM have focused on short-term student performance, there is nothing inherent in student modeling or in EDM that requires such a focus. Conceptually, it is possible to construct models that predict other constructs of interest, such as whether the student will remember a skill after a period of time. Why would we want to construct such a model? We argue that whether a student will not only respond correctly on an item right away, the mastery approach used by KT, but whether the student will remember enough to respond correctly, after taking a break from working with the tutor, is a better definition of mastery. At best, it is unclear how to apply a short-term model such as KT or PFA for such a decision-making task. However, if we could build such a detector, we could deploy it in a computer tutor and use it to decide when to stop presenting items to students. Perhaps a student who gets several items correct in a row, and masters the skill by traditional definition, will be predicted to not retain the skill and should receive additional practice.

The approach we use is straightforward: rather than attempting to predict every next student performance, instead we focus on student performances that occur after a long delay. In this way, even though we are not explicitly modeling the forgetting process, our student modeling approach captures aspects of performance that relate to student long-term retention of the material. It is reasonable that the field of student modeling did not start with long-term retention, as only a small minority of student practice opportunities take place after a long delay. Therefore, such restrictions would result in a too-small data set to train the student model parameters. However, with the advent of large educational sets, such restrictions become less relevant.

The data used in this analysis came from the ASSISTments system, a freely available web-based tutoring system for 4th through 10th grade mathematics (approximately 9 through 16 years of age). The system is mainly used in urban school districts of the Northeast United States. Students use it in lab classes that they attend periodically, or for doing homework at night.

We collected data from school year September 2010 to September 2011, which consisted of 15931 students who solved at least 20 problems within ASSISTments. We filtered out skills that have fewer than 50 students. As a result, we have 2,363,982 data records. Each data record is recorded right after a student answered a problem, and logged relevant information including the identity of the student, the problem identity and skills required to solve it, the correctness of the student's first response to this problem, the duration the student spent on this problem, and the timestamp when the student start and finish solving this problem.

For this task, we defined a student as retaining a skill if he was able to respond correctly after approximately a week. We instantiated a week as any duration between 5 and 10 days. and choose the time interval of 5-10 days as our objects to analyze. We randomly selected one fourth of the students as training data, which result in 27,468 final data records. Note that less than 5% of the data are relevant for training a model of student retention. Thus, this problem requires large data sets.

# 3. STUDENT RETENTION ANALYSIS

## 3.1 RQ1: Is student retention predictable?

To answer this question, we built a logistic regression model, using the 27,468 data points with delayed practice opportunities described previously. The dependent variable is whether the student responded correctly on this delayed outcome. We used user identity (user_id) and skill identity (skill_id) as factors (fixed effets) in this model. We used the following features as covariates, treating incorrect responses as a 0 and correct responses as a 1:

- *n_correct*: the number of prior student correct responses on this skill; This feature along with *n_incorrect*, the number of prior incorrect responses on this skill are both used in PFA models;
- *n_day_seen*: the number of distinct days on which students practiced this skill. This feature distinguishes the students who practiced more days with fewer opportunities each day from those who practiced fewer days but more intensely, and allow us to evaluate the difference between these two situations. This feature was designed to capture certain spaced practice effect in students data;
- *g_mean_performance*: the geometric mean of students' previous performances, using a decay of 0.7. For a given student and a given skill, use *opp* to represent the opportunity count the student has on this skill, we compute the geometric mean of students' previous performance using formula: *g_mean_performance(opp) = g_mean_performance(opp-1)*0.7 + correctness(opp)*0.3*. The geometric mean method allows us to examine current status with a decaying memory of history data. The number 0.7 is selected based on experimenting with different values.
- *g_mean_time*: the geometric mean of students' previous response time, using a decay of 0.7. Similar with *g_mean_performance*, for a given student and a given skill, the formula of the geometric mean of students' previous response time is: *g_mean_time(opp) = g_mean_time(opp-1)*0.7 + response_time(opp)*0.3*;
- *slope_3*: the slope of students' most recent three performances. The slope information helps capture the influence of recent trends of student performance;
- *delay_since_last*: the number of days since the student last saw the skill. This feature was designed to account for a gradual forgetting of information by the student;
- *problem_difficulty*: the difficulty of the problem. The problem_difficulty term is actually the problem easiness in our model, since it is represented using the percent correct for this problem across all students. The higher this value is, the more likely the problem can be answered correctly.

It is important to note that the features were computed across all of the data, not just the items on which the student had not practiced the skill for 5 to 10 days. For example, the n_correct feature is computed across *all* of the student practices on the skill, not just those practices with a 5 to 10 day delay interval. However, we only create a row in our data set for such delayed retention items (thus there are 27,468 rows). After training the model on the ASSISTments data, we got a $R^2$ of 0.25. Since this model fit represents training-data fit, it is optimistic. But the model fit is at least strong enough to conclude that student retention appears to be predictable.

The Beta coefficient values and p-values for each covariate are shown in Table 1.

In this table, the positive B values mean the larger the covariate is, the more likely the student respond to this problem correctly. To our surprise, the influence of the *n_correct* and the *n_incorrect* features are not reliably different than 0. The features *n_day_seen* and *g_mean_performance,* on the contrary, are reliable predictors of student retention. In other words, for predicting long-term retention, the number of days on which the student practiced the skill is important, as is his recent performance. This result is consistent with cognitive "spaced practice effect" result [11]. The raw number of correct and incorrect responses is not a meaningful predictor. We expected that response time would be relevant to retention, due to its connection to automaticity and mastery [1].

| Covariate | B | p-value |
|---|---|---|
| ***n_correct*** | -0.003 | .330 |
| ***n_incorrect*** | -0.005 | .245 |
| ***n_day_seen*** | 0.055 | .000 |
| ***g_mean_performance*** | 0.813 | .000 |
| ***g_mean_time*** | 0.073 | .043 |
| ***slope_3*** | -0.033 | .444 |
| ***delay_since_last*** | -0.015 | .182 |
| ***problem_difficulty*** | 5.926 | .000 |

Table 1. Parameter table of covariates in Model1.

From the likelihood ratio tests of the training set, we found that the *skill_id* and *user_id* are also both important features in this model. This indicates that student performance on retention items varies by skill and by student. It is tempting to claim that retention varies by student, but this claim is premature as the user_id factor models student performance on retention items. However, such performance is composed of how well the student learned the material as well as how much of that knowledge was retained. A student could have a strong memory, but if he never learned the material his user_id factor would be low. Therefore, user_id does not solely represent retention.

To strengthen the results, we built test set to validate the model. Since the users of testing set are different from those of the training set, we cannot look up user parameters directly for users in the testing set. Instead, we use the mean value of user parameters of the model as an approximation of the user parameter in the testing set. We also did the same thing for the skills that only appear in the testing set.

The R2 of this model on the testing set is 0.17, indicating a reasonable model fit in-line with other attempts at using PFA

## 3.2 RQ2: Does forgetting vary by student?

We would like to separate the impact of the user_id feature into student knowledge and student retention. To accomplish this task, first, we started from the logistic regression model that we used in section 3.1, removed the factor *user_id* and substituted a covariate *non_1st_pcorrect*. The feature *non_1st_pcorrect* is the percent of a student's non-first attempts of the day that are correct. The intuition is that a student's first attempt on a skill each day is the one that is most affected by retention. By considering the student's overall performance, but excluding these items, we are estimating the student's proficiency in the domain in a way that is less contaminated with forgetting, and is thus a purer estimate of the student's knowledge. We trained this model on the same data as the previous model. The feature *non_1st_pcorrect* has an estimated Beta coefficient of 3.878, with a p-value 0.000. We got an $R^2$ of 0.210 on the data, which is a reasonable model fit. The difference in model fit is caused by the substituting the percent correct, on non-first encounters, for user_id.

We were curious as to the cause of this difference in model fits, and investigated the residuals from our model. The question is whether the residual was systematic, and could be predicted by user_id. We fit a general linear model with *user_id* as a random effect, and the residual as the dependent variable. The $R^2$ of this model is 0.235. Thus, the residual in our model, after accounting for student overall percent correct in contexts where forgetting was minimal, does vary systematically by user_id. Thus it appears that there is some construct beyond performance, such as forgetting, that varies by student.

Although it is tempting to claim this term represents student forgetting, it is necessary to validate the construct [6] we have modeled. To test whether we have modeled retention, we first extracted the student random effects from our GLM. We then computed the correlation between that term, and each student's difference in performance between the first and second question on a skill that occurs each day. Our belief is that this difference in performance is related to student forgetting, since a large increase in performance from the first to the second item suggests the student is remembering old information. Unfortunately, the correlation between these terms was negligible, so we are still searching for what our per-student effect is actually modeling.

# 4. CONTRIBUTIONS

This paper makes three main contributions. First, the mastery learning notion is expanded to take into account the long-term effect of learning. In comparison to the traditional view that Corbett and Anderson brought up in their seminal work [5], which looks at only the immediate knowledge, this paper looks at broader notion of knowing a skill.

The second contribution this paper makes is extending the PFA model [10] with features that are likely to be relevant for retention. Most prior work has focused on concepts such as item difficulty or amount of assistance required to solve a problem. However, those features focus on student performance and properties of items, not on broad characterizations of performance. Our study confirmed that the long-term knowledge appears to vary by skill, and possibly by student. In addition, the number of days on which a student practiced a skill is relevant, and could be an important feature in directing ITS decision making to enhance retention. This result confirms the spaced practice effects in a larger scope; also we found that the number of correct responses seem to be not so important in predicting knowledge retention.

The third contribution this paper makes is on discovering a new problem that is actionable by ITS. Previous student models focus on estimating student current knowledge, which is powerful for EDM, and an efficient use of data for testing a model, but provides limited guidance for tutorial decision making. This paper proposed a diagram of ITS action cycle that can be used to discover new problems in the EDM field that can lead to higher mastery learning rate in ITS systems.

One goal of EDM is to address questions that are relevant for tutorial decision making of ITS. Currently, many ITS simply present a sequence of problems and evaluate student performance right after the student finished these problems to see if the student mastered the given skill. This process does not have the mechanism for the system to review students' knowledge after a time period, nor know about students' long term performance. It is dangerous for ITS to promote a student on the basis of short term performance. We propose the follows diagram shown in

Figure 1, which allows ITS to aim for students long-term mastery learning.

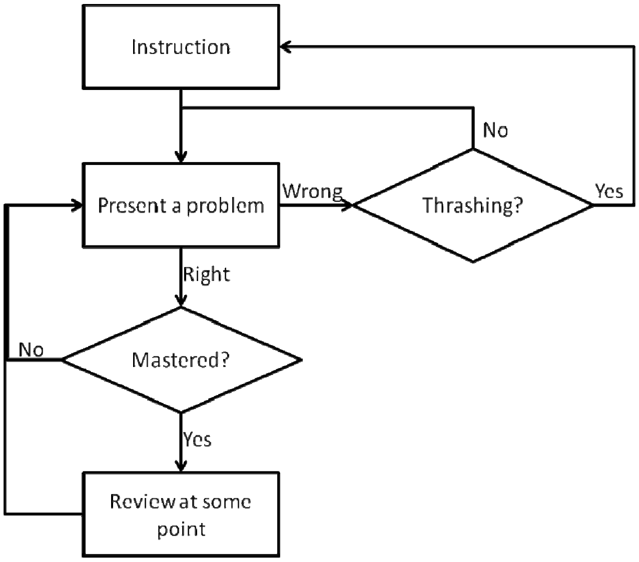

Figure 1. Enhanced ITS mastery learning cycle

This paper focuses on the diamond on the left side, whether the student has mastered a skill. Rather than using local criteria to decide whether mastery has occurred, we trained a model to decide mastery based on predicted performance. Beyond this EDM work, some review mechanism for ITS seems warranted, as for cumulative domains, such as mathematics, ensuring that student's have retained their knowledge is critical. Correspondingly, we have added a review mechanism to ASSISTments [13].

Another interesting EDM problem is the diamond on the right: when is a student likely to fail to master a skill in a timely manner and (statistically) exhibit negative behaviors such as gaming? We have made progress on this problem as well, and dubbed the phenomenon "thrashing." If a student is unlikely to master a skill via problem solving, it is essential to do something else, such as peer tutoring, having the teacher intervene, or providing instruction within the ITS.

What we like about both of these problems is that they are rich challenge problems for EDM, and provide actionable information for ITS to make use of in their decision making. If a student appears likely to retain a skill, it is probably not necessary to keep presenting items. If a student is likely to not master a skill, it is probably not productive to keep presenting problems.

# 5. FUTURE WORK AND CONCLUSIONS

There are three questions that we are interested in exploring. First, do students vary in how quickly they forget? Our first attempt at teasing apart the user_id factor gave inconclusive results, but this area is important enough to warrant further study. Another issue that we are interested in addressing is what are additional features that relate to forgetting? The field of psychology is rich in ideas, but there has been little existing work in student modeling.

Finally, we would like to deploy this model to a working ITS in the field. On one hand, this can help verify the model; on the other hand, this could be used to improve the ITS systems to help student achieving long-term mastery learning.

This paper present an ITS mastery learning diagram, which brings up useful problems in EDM that needs more work. In this paper we concentrate on estimating student knowledge retention and discovered some useful features for this task. Also, we were able to conclude student long-term performance is predictable, even when a student's ability to remember a skill comes in to play.

# 6. REFERENCES

[1] Anderson, J.R., Rules of the Mind. Lawrence Erlbaum (1993).

[2] Beck, J.E., et al. Predicting student help-request behavior in an intelligent tutor for reading. Ninth International Conference on User Modeling, (2003), Johnstown, PA.

[3] Cain, L.F. and Willey, R.D.V, The effect of spaced learning on the curve of retention. Journal of Experimental Psychology, Vol 25(2), (Aug 1939), 209-214. doi: 10.1037/h0054640.

[4] Cepeda, N.J., et al. Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psychological Bulletin, Vol 132(3), (May 2006), 354-380. doi: 10.1037/0033-2909.132.3.354.

[5] Corbett, A.T. and J.R. Anderson, Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, (1995), 4: p. 253-278.

[6] Crocker, L. and J. Algina, Introduction to Classical and Modern Test Theory. (1986), Fort Worth: Harcourt Brace Jovanovich College Publishers. 482.

[7] Ebbinghaus, H., Memory: A Contribution to Experimental Psychology. (1885), New York: Teachers College, Columbia University.

[8] George B.S. and John A.E., Knowledge Taught in School: What Is Remembered?, Review of Educational Research Summer, (1994) vol. 64 no. 2 253-286

[9] Pavlik, P.I. and J.R. Anderson, Practice and Forgetting Effects on Vocabulary Memory: An Activation-Based Model of the Spacing Effect. Cognitive Science, (2005). 29(4): p. 559-586.

[10] Pavlik, P.I., H. Cen, and K.R. Koedinger, Performance factors analysis—A new alternative to knowledge tracing., in Proceedings of the 14th International Conference on Artificial Intelligence in Education, V. Dimitrova and R. Mizoguchi, Editors. (2009): Brighton, England.

[11] Perruchet, P. (1989), The effect of spaced practice on explicit and implicit memory. British Journal of Psychology, 80: 113–130. doi: 10.1111/j.2044-8295.1989.tb02306.x

[12] Qiu, Y., et al. Does Time Matter? Modeling the Effect of Time with Bayesian Knowledge Tracing. in Fourth International Conference on Educational Data Mining. 2011.

[13] Wang, Y. and N.T. Heffernan. Towards Modeling Forgetting and Relearning in ITS: Preliminary Analysis of ARRS Data. in Fourth International Conference on Educational Data Mining. 2011.