Tracing Knowledge and Engagement in Parallel in an Intelligent Tutoring System

Sarah E Schultz
Worcester Polytechnic Institute
100 Institute Rd
Worcester, MA 01609
seschultz@wpi.edu

Ivon Arroyo
Worcester Polytechnic Institute
100 Institute Rd
Worcester, MA 01609
iarroyo@wpi.edu

ABSTRACT
Two of the major goals in Educational Data Mining are determining students' state of knowledge and determining whether students are affectively engaged with the task and in positive affective states. These two problems are usually examined separately and multiple methods have been proposed to solve each of them. However, little work has been done on tracing both of these states in parallel and the combined effect on a student's performance. In this work, we propose a model for tracing student engagement in parallel with knowledge as the student uses an Intelligent Tutoring System. We then compare this model to existing methods of tracing student knowledge and engagement.

Keywords
Knowledge tracing, engagement, performance, behavior, affect detection

1. INTRODUCTION
Intelligent Tutoring Systems are meant to adapt to a students' needs in order to better teach the student. In order to do this, they must have an estimation of student knowledge as the student progresses through the tutoring session. Systems might use their estimations of a student's mastery of the subject to decide whether to change the difficulty of problems given or progress to a new unit. These models may also be used by teachers and researchers to estimate students' mastery of skills or knowledge units. In the field of Educational Data Mining, the standard way to model and trace student knowledge is via knowledge tracing [1]. However, students often become disengaged as they use the software, as a result of boredom of frustration, confounding models which rely solely on performance data on individual questions to estimate knowledge, making it appear as though a student is forgetting.

The ability to detect affect is useful for Intelligent Tutors as it allows for the possibility for the tutor to intervene when a negative affective state is detected and help the student become engaged and motivated to learn. Some systems make use of sensor data to determine affect [7], but this is often impractical in a real-life learning scenario. Some researchers attempt to create sensor-less affect detectors using human coders who will observe students' apparent affective state during a session and then match these observations to behaviors that occur within the system at the same time in order to create a model, such as BROMP [11]. This is time-intensive, requiring a certain number of observations and highly trained coders.

While research has been done on tracing affective engagement without sensors or coders [3], little research has been done in modeling both knowledge and affect in parallel, attempting to account for these biases in knowledge estimation. In particular, a student's performance cannot be assumed to depend solely upon his or her knowledge of a skill, as how he or she is feeling will likely impact performance, as well. This is an area that is ripe for exploration.

Given a set of behaviors regarding correctness, timing and help seeking, some behaviors may be attributed to affective states, and some of them may be attributed to cognitive states [6, 7]. A Bayesian Hidden Markov Model (HMM) that attempts to trace knowledge and affect in parallel within the same model could potentially be able to discern between low affect and low knowledge, given a set of student correctness, timing and help seeking behaviors.

2. PREVIOUS WORK
The models explored in this work were inspired by previous successful Bayesian networks modeling students' knowledge and affect. The first of these is Knowledge Tracing, which has become a standard [1]. The second is the HMM-IRT model by Johns and Woolf [4], which took first steps towards modeling affect and knowledge in parallel.

2.1 Bayesian Knowledge Tracing
Corbett and Anderson's Bayesian Knowledge Tracing (BKT) [1] (Figure 1) is a hidden Markov model with two nodes at every time-step: the current (latent) knowledge state of the student and his or her performance on the current question (observed). Based on a student's correctness at answering questions at each time-step, the model estimates the probability that the student knows the current skill and then predicts the probability that the student will correctly answer the next question. The parameters for this model are $P(L_0)$, the probability that a student already knows the skill; P(T), the probability of learning the skill from one time-step to the next; P(G), the probability that a student who does not know the skill correctly guesses; and P(S), the probability that a student who does know the skill slips and gets the answer incorrect.

Traditionally, the KT model does not allow for forgetting (or unlearning) and this parameter is set to zero; this is in some way a quick fix, as when the model allows for forgetting, it is very sensitive to students "gaming the system" [9]. Consequently, estimates of knowledge mastery could quickly decline when students start behaving in these ways, such as hint abusing or quick guessing, and appear as if students are unlearning.

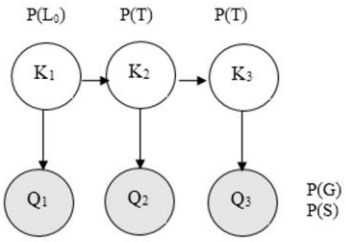

Figure 1- Bayesian Knowledge Tracing

2.2 HMM-IRT

Johns and Woolf [4] proposed another model, called the Hidden Markov Model-Item Response Theory (HMM-IRT) model. In this model, rather than using BKT, they use a hidden Markov model for tracing affect (what we call affective engagement in this paper), but pair it with a model for predicting student knowledge that relies on Item Response Theory for the estimation of conditional probabilities between specific question items and knowledge. Unlike BKT, this model estimates a single knowledge node. The HMM-IRT model allows the estimation of students' engagement at various time-steps (and relies on parameters of transitioning between affect/engagement states), but assumes a single mastery node, without learning or forgetting parameters.

The result of that research was that adding the affect/engagement component (top part of Figure 2) to the knowledge estimation model (bottom part of Figure 2) allowed for less of a decline in knowledge estimations after each question, which was apparently due to gaming behaviors and not due to unknowing.

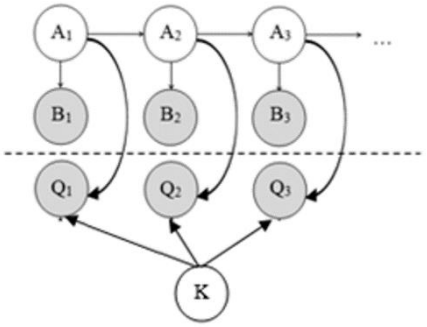

Figure 2- HMM-IRT Model

3. THE KAT MODEL

The Knowledge and Affect Tracing (KAT) model, shown in Figure 3, combines Knowledge Tracing with the HMM Affect Tracing portion of the HMM-IRT model, creating a model which allows for change in both students' knowledge and affective states. Both of these states influence question correctness.

The most important contributions, in our perspective, of both the HMM-IRT model and the KAT model, are the inclusion of transition probabilities between engaged states, in particular the probability of *becoming disengaged* in the next time step given that the student was previously engaged, and the probability of *becoming re-engaged* given that a student was previously disengaged. Knowing estimates of these probabilities for any learning system or for specific knowledge components should be very valuable to understand the impact of a learning system, or interventions. Similarly, it is valuable to know when estimates of engagement are low for personalization purposes, and knowing whether a student is likely game in the next problem or not.

The main drawback of the HMM-IRT model was that it did not include probabilities of acquisition or retention, but instead modeled students' knowledge as something stable and trait-like. Adding knowledge tracing to this model should enable researchers and systems to better predict both performance and behavior (gaming or not gaming) at the next step.

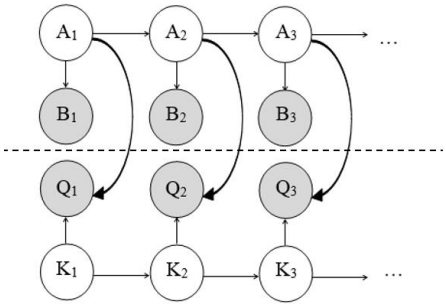

Figure 3- The KAT Model

The behaviors examined were the same as those used by Johns and Woolf [4]. These are quick guess (the student makes an attempt in less than four seconds), bottom out hint (the student uses all available hints), and normal (any other behavior). One additional behavior, called "many attempts", was also added for this work. This was defined as a student making more than three attempts at answering a problem. As multiple choice problems typically include only five possible answers, a student making more than three attempts has likely simply clicked on most choices. Baker, et al. have also shown relatively few attempts to be a good predictor of engaged concentration [8]. In preliminary tests of the KAT model, including "many attempts" as a possible behavior led to better fit than using only three behaviors in both datasets. The three behaviors not classified as normal are grouped as "gaming" behaviors in order to allow the models to predict whether a student will game at each opportunity. Although gaming is traditionally thought of as disengaged behavior, students could act in a way that is defined here as a gaming behavior even when they are engaged.

The new conditional probability tables of the observed nodes of the KAT model are shown in Tables 1 and 2. Knowing the skill (K), being affectively engaged (A), answering a question correctly (Q), and behaving normally (B) (i.e., not gaming) are indicated by "true" in their respective columns. The last column gives a name to these new probabilities to be estimated, which consist of guessing or slipping while being in a state of affective engagement or disengagement at the same time.

Table 1- CPT for Performance (Q) Nodes of KAT Model

Known (Latent)	Engaged (Latent)	Correct (Observed)	Probability
False	False	False	1-guess_not_eng
True	False	False	slip_not_eng
False	True	False	1-guess_engaged
True	True	False	slip_engaged
False	False	True	guess_not_eng
True	False	True	1-slip_not_eng
False	True	True	guess_engaged
True	True	True	1-slip_engaged

The probabilities associated to the Gaming Behavior nodes (B) are shown in table 2, and depend on affective engagement. These probabilities distinguish whether a student has gamed in a situation when he/she was actually truly engaged (some sort of an

'affective slip') corresponding to 'game_engaged' and its counterpart, where the student was actually affectively disengaged but apparently behaved normally this time (1-game_not_eng).

Table 2- CPT for Gaming Behavior Nodes (B) of KAT Model

Engaged (Latent)	Non-Gaming-Behavior (Observed)	Probability
False	False	game_not_eng
True	False	game_engaged
False	True	1-game_not_eng
True	True	1-game_engaged

San Pedro et al. showed that student knowledge of a skill is related to affect (for example, students who know a skill well are more likely to be engaged) [7], so a variation on the KAT model was created to take this into account. This model, KAT2, includes the link between knowledge and affect.

4. DATASETS

The data was gathered from student logs of two mathematics tutoring systems, ASSISTments [2] and Wayang Outpost [7], for middle and high school students. All problems in Wayang are multiple choice, while problems in ASSISTments generally, though not always, require students to type in their answer, instead.

The ASSISTments data used here is from the 2009-2010 school year. This data comes from a special type of problem in ASSISTments called "skill builders." In skill builders, students practice a specific skill until they get three problems correct in a row, in which case the skill is considered "mastered," or they reach a preset daily limit and are told to return later. The Wayang data set comes from the spring of 2009 and includes two hundred ninety five students in grades 7 through 10 from two rural-area schools in Massachusetts.

Five knowledge components were chosen from ASSISTments and four from Wayang to test the models as they are all limited to examining each knowledge component separately. Table 4 shows the breakdown of the data used by knowledge component.

Table 4- Knowledge Components Examined

Knowledge Component	System	Number Students	Total Number Opps	% Gaming
Box and Whisker	ASSISTments	505	2020	13
Circle Graph	ASSISTments	616	2487	30
Table	ASSISTments	713	2894	4
Pythagorean Theorem	ASSISTments	283	1290	10
Equations	ASSISTments	408	1598	35
Perimeter	Wayang	285	1422	15
Area	Wayang	279	1385	17
Angles	Wayang	274	1355	16
Triangles	Wayang	260	1267	20

5. METHODS

All models were built using Murphy's Bayes Net toolbox for MATLAB [5]. A student-level five-fold cross validation was run on all models, keeping folds consistent across models. Parameters were learned for the training data using expectation maximization and then tested on the test data. This was done five times for each knowledge component, where each time a different fold served as the test data while the other four served as training data. For all models, predictions of performance at the next step were compared with actual performance in order to calculate mean absolute error (MAE) and root mean squared error (RMSE). Additionally, for KAT and HMM-IRT, predictions of behavior were compared to actual behaviors. As struggling students will see more questions assessing the same knowledge component in both ASSISTments skill builders and Wayang Outpost, only the first five opportunities within each knowledge component are examined to avoid over-fitting to such students. Since these five opportunities are likely to be within one session, not allowing time for students to forget material, forgetting is still assumed to be zero. All data and code used can be found at the first author's webpage [10].

6. RESULTS

As both error metrics calculated, MAE and RMSE, resulted in patterns that were not significantly different, only RMSE is reported here.

Tables 5 and 6 show each model's predictive performance on the ASSISTments data and Tables 7 and 8 show how well the models did on the Wayang data. Tables 5 illustrates the RMSE for each model's prediction of students' performance, while 6 shows the error of prediction for students' behavior. These tables show the mean average of RMSEs across folds for each skill.

Table 5 – RMSE for Performance (Q)

Skill	KT	HMMIRT	KAT	KAT2
Box and Whisker	**0.426**	0.495	0.468	0.493
Circle Graph	**0.434**	0.524	0.507	0.512
Table	**0.467**	0.498	0.483	0.495
Pythagorean Theorem	**0.480**	0.498	0.484	0.503
Perimeter	**0.471**	0.476	0.476	0.476
Area	**0.455**	0.476	0.460	0.459
Angles	**0.454**	0.466	0.466	0.465
Triangles	**0.483**	0.487	0.4866	0.485

Table 6 – RMSE for Gaming Behavior (B)

Skill	HMMIRT	KAT	KAT2
Box and Whisker	0.350	0.326	**0.325**
Circle Graph	0.196	**0.178**	0.179
Table	0.462	**0.422**	0.433
Pythagorean Theorem	0.303	**0.295**	**0.295**
Perimeter	0.357	0.357	**0.356**
Area	0.377	0.362	**0.361**
Angles	0.377	0.359	**0.357**
Triangles	0.400	0.394	**0.392**

These tables show that BKT is the best predictor of student performance-- the correctness at answering future questions. The two KAT models also generally outperform HMM-IRT at predicting performance. The original KAT model was significantly better at predicting performance than the KAT2 model on the ASSISTments data (ttest $p<0.05$), except on the skill

"Table" (p=0.09), and although the KAT2 model performed slightly better on the Wayang data, this difference was not significant (p>0.1). Both KAT models are also significantly better at predicting behavior than the HMM-IRT model, except on the Wayang topic "Perimeter," where KAT2 is marginally better than HMM-IRT and KAT is marginally worse. The two KAT models were not significantly different with respect to predicting behavior, except for on the ASSISTments skill "Table," on which the original KAT model performed better.

7. DISCUSSION

While traditional BKT appears to be the best model for predicting student future correctness performance at math questions, KAT seems to be best at predicting performance and gaming behaviors simultaneously. KAT better predicts performance than HMM-IRT in eight of nine knowledge components tested and gaming behavior in all nine knowledge components, including six where there was a significant difference between KAT's predictions and HMM-IRT's. As KAT was significantly better than KAT2 at predicting student performance at math questions in one system, the KAT model appears to be a better choice for modeling students than the KAT2 variation.

The fact that KAT, which allows for student learning, was better able to predict performance means that it is quite likely that students are, in fact, learning while using these systems, so that the probability of acquisition and retention matter at the moment of predicting knowledge and performance in the next time slice. Assuming that a student's knowledge state does not change during the session, as in HMMIRT, leads to a poorer model fit.

It is interesting that KAT was also better at predicting behavior than HMMIRT, as both models use the same CPT for these nodes. When both affective transitions and learning are allowed, a change in performance can be attributed to either, or both, perhaps allowing a more accurate model of engagement, and therefore better predictions of gaming behavior.

8. CONTRIBUTIONS AND FUTURE WORK

This work introduced a new model, KAT, for tracing students' knowledge and engagement in parallel while using an ITS. While the traditional KT alone was slightly better at predicting performance than any of the other models, KAT was better at predicting student performance and behavior than the previously existing HMM-IRT model. A variation, the KAT2 model, was also explored and shown to be slightly weaker than the original KAT model.

While this work included the original form of the KAT model and one variation, many other variations could be valid. For example, research involving sensors and self-reports of affect has shown that performance on one question influences a student's affect at the next time-step [7]. This could be added to the KAT model to create another variation.

Future versions of the KAT model should also allow for more affective states, rather than measuring only engagement. For this study, it was useful to keep all variables binary in order to determine which model was best able to predict performance and behavior based on knowledge and engagement, but the KAT model is meant to be a model of knowledge and *affect* tracing. It is possible that allowing for more specific affective states could allow for better prediction of gaming. Perhaps being bored is more likely to lead to these behaviors than being frustrated, although both could fall under the category of "disengaged."

Additionally, allowing for forgetting would be an interesting avenue to explore in the future, looking at the knowledge predictions. It is possible KT will predict students are forgetting whereas knowledge estimations will not change in models allowing for gaming.

ACKNOWLEDGEMENTS

We would like to thank the U.S. Department of Education, #P200A120238, Fellowships in Computer Science to Support the Learning Sciences & Security, Heffernan (PI) and the National Science Foundation, #1324385, Cyberlearning DIP, Impact of Adaptive Interventions on Student Affect, Performance, and Learning, Arroyo, Woolf and Burleson. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.We are also grateful for the feedback and advice of Dr. Neil Heffernan, Dr. Joseph Beck, Trenton Tabor, Michael Wixon, and the entire EDM research group at Worcester Polytechnic Institute. We would also like to thank ASSISTments for making their dataset used within publicly available.

REFERENCES

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

[2] Mendicino, M., Razzaq, L., & Heffernan, N. T. (2009). A Comparison of Traditional Homework to Computer-Supported Homework. Journal of Research on Technology in Education, 41(3).

[3] Beck, J.E. "Engagement tracing: using response times to model student disengagement." *Proceedings of AIED conference*, 2005. p. 88-95. IOS Press

[4] Johns, J. and Woolf, B.P. "A Dynamic Mixture Model to Detect Student Motivation and Proficiency." *Proceedings of AAAI Conference*, 2006, 1, p. 163-168.

[5] Murphy, K. "The Bayes Net Toolbox for MATLAB", *Computing Science and Statistics*, 2002.

[6] San Pedro, M.O.Z., Baker, R.S.J.d., Gowda, S.M., Heffernan, N.T. "Towards an Understanding of Affect and Knowledge from Student Interaction with an Intelligent Tutoring System." In *Proceedings of the 16th International Conference on Artificial Intelligence and Education*. Memphis, TN, USA, July 9-13, 2013.

[7] Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., and Christopherson, R. "Emotion Sensors Go To School." In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*. Brighton, UK, July 6-10, 2009.

[8] Baker et al. "Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra." In *Proceedings of the 5th International Conference on Educational Data Mining.* Chania, Greece, June 19-21, 2012.

[9] Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z. (2004) Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game The System". In *Proceedings of ACM CHI 2004: Computer-Human Interaction*, 383-390.

[10] Schultz, S. Webpage: users.wpi.edu/~seschultz

[11] Ocumpaugh, J., Baker, R.S.J.d., Rodrigo, M.M.T. (2012) "Baker-Rodrigo Observation Method Protocol (BROMP)" *1.0. Training Manual version 1.0. Technical Report*. New York, NY: EdLab. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.