Worcester Polytechnic Institute
Digital WPI

Masters Theses (All Theses, All Years)

Electronic Theses and Dissertations

2015-01-27

Tracing Knowledge and Engagement in Parallel by Observing Behavior in Intelligent Tutoring Systems

Sarah E. Schultz
Worcester Polytechnic Institute

Follow this and additional works at: https://digitalcommons.wpi.edu/etd-theses

Repository Citation
Schultz, Sarah E., "*Tracing Knowledge and Engagement in Parallel by Observing Behavior in Intelligent Tutoring Systems*" (2015). *Masters Theses (All Theses, All Years)*. 140.
https://digitalcommons.wpi.edu/etd-theses/140

This thesis is brought to you for free and open access by Digital WPI. It has been accepted for inclusion in Masters Theses (All Theses, All Years) by an authorized administrator of Digital WPI. For more information, please contact wpi-etd@wpi.edu.

Tracing Knowledge and Engagement in Parallel by Observing Behavior in Intelligent Tutoring Systems

By

Sarah E Schultz

A Thesis

Submitted to the Faculty

Of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

In

Computer Science

January 2015

____

Prof. Ivon Arroyo, Advisor

____

Prof. Neil Heffernan, Co-advisor

____

Prof. Joseph Beck, Reader

____

Prof. Craig Wills, Department Head

Abstract

Two of the major goals in Educational Data Mining are determining students' state of knowledge and determining their affective state. It is useful to be able to determine whether a student is engaged with a tutor or task in order to adapt to his/her needs and necessary to have an idea of the students' knowledge state in order to provide material that is appropriately challenging. 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 examining both of these states in parallel and the combined effect on a student's performance. The work reported in this thesis explores ways to observe both behavior and performance in order to more fully understand student state.

Acknowledgements

I could not have completed this work alone. First of all, I’d like to thank my advisor, Dr. Ivon Arroyo, for all of her help, support, and advice. I’d also like to thank my co-advisor, Dr. Neil Heffernan III, and reader, Dr. Joseph Beck for their insights.

All of my colleagues in the Advanced Learning Technology and ASSISTments labs, especially my office-mate Wixon, and my good friend, Korinn Ostrow, for all of their help throughout iterations of the studies reported herein and for generally being awesome people.

I would also like to take this opportunity to thank the doctors, nurses, therapists, and staff at UMass Memorial Hospital and Whittier Rehabilitation, as well as the nurses and therapists at the Natick Visiting Nurses Association and Beth Israel Deaconess Hospital who helped me to recover after an accident last November. It is amazing that six months later, I am back at WPI finishing this thesis.

To my family for supporting me through everything. Without their continued guidance and love, I would never have been able to do any of this, and for that I am truly grateful. Especially to my mother, who raised me and my brother, and always believed in me.

And of course to my other half, Trenton Tabor, for everything he does for me, from suggestions on next steps in my work to attending my talks to being there when I was recovering and updating my professors and for always being willing to accompany me to anything I want to do. I am so glad to have such a caring person in my life.

Table of Contents

1. Introduction

Intelligent Tutoring Systems are computer programs meant to simulate the behaviors of a human tutor, and as such they must 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 adjust 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 Bayesian knowledge tracing [1]. However, students often become disengaged as they use the software, confounding models which rely solely on performance data to estimate knowledge. To these models, it might appear as though a student is forgetting or unlearning when she is simply no longer engaged in using the system. For example, Figure 1 suggests that this student was un-learning, while after looking at the logs in detail, it was clear that, after the 7th problem, the student was just clicking through all of the available multiple-choice answers without attempting to answer correctly. This type of behavior is defined by Baker et al as "gaming the system" [2] and is considered to be an indicator of disengagement or negative affect.

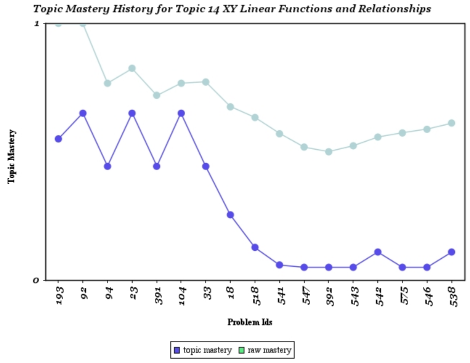


Figure 1- Bayesian Knowledge Estimation of a student on one skill (bottom line)

In this context, affect is defined as the current feeling or emotional state of the student, such as frustration, confusion, or engaged concentration. The ability to detect affect is useful for Intelligent Tutors as it allows the possibility for the tutor to intervene when a negative affective state is detected in order to help the student become engaged and motivated to learn. Some systems make use of sensor data to determine affect [10], but this is often impractical in a real-life learning scenario. If a student is assigned homework using a tutoring system, for example, researchers cannot expect that all students will have webcams, pressure mice, or posture sensors in their homes. Even in the classroom, except when researchers provide sensors for a specific study, it is unreasonable to expect to collect sensor data on every student. 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 [15]. While this has led to good results, it is time-intensive, requiring many and highly trained coders.

While some research has been done in tracing engagement without sensors or coders [3], little work 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. 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 [5, 9], such as in the example in Figure 1. A model that attempts to trace knowledge and affect in parallel could potentially be able to discern between low affect and low knowledge, given a set of student correctness, timing and help seeking behaviors.

2. Literature Review

The models explored in this work were inspired by previous successful Bayesian networks modeling students' knowledge and affect. The first of these is Bayesian Knowledge Tracing, which has become a standard in the field [1]. The second is the dynamic-mixture 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 2) is a hidden Markov model where at each time-step there is one latent node and one observed node. The observed node is the student's performance on the questions (correct or incorrect) and the latent is their knowledge state. Based on a student's performance at each time-step, the model estimates the probability that the student knows the skill or knowledge component s/he is practicing 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 guesses the correct answer; and P(S), the probability that a student who does know the skill slips and answers incorrectly.

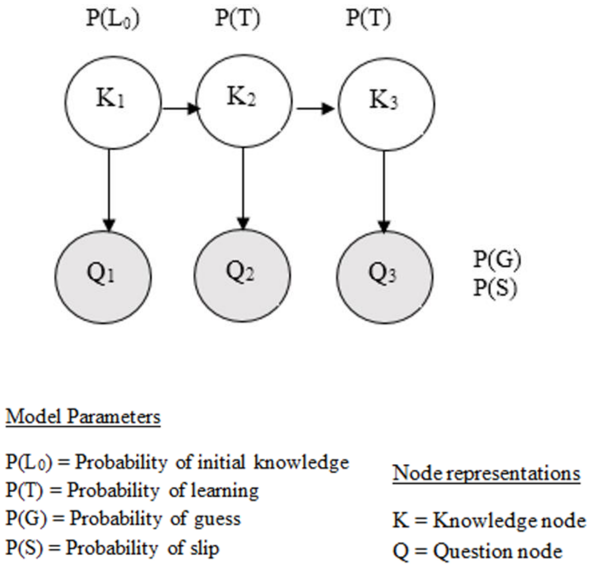

P(L0)
P(T)
P(T)
K1
K2
K3
Q1
Q2
Q3
P(G)
P(S)
Model Parameters
P(L0) = Probability of initial knowledge
P(T) = Probability of learning
P(G) = Probability of guess
P(S) = Probability of slip
Node representations
K = Knowledge node
Q = Question node

Figure 2- Bayesian Knowledge Tracing

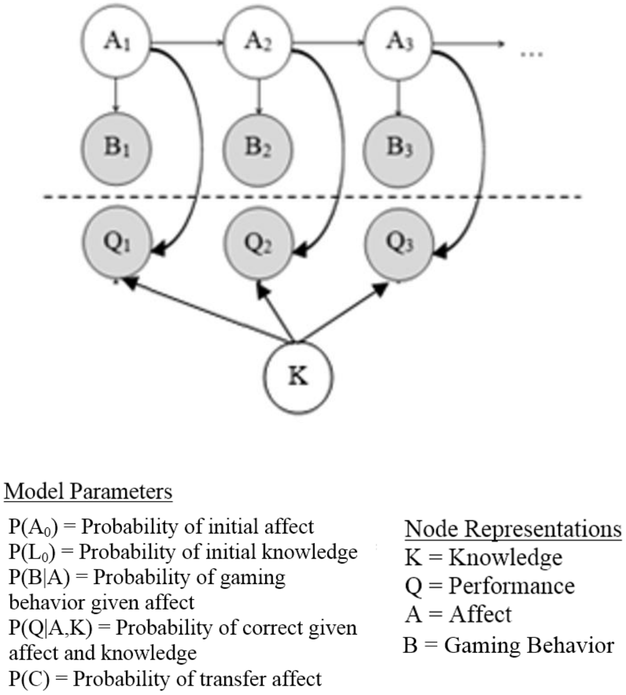
When Corbett and Anderson first published the Bayesian knowledge tracing model in 1995, they claimed that their goal was "to implement a simple student modeling process that would allow the tutor to […] tailor the sequence of practice exercises to the student's needs" [1]. While knowledge tracing is generally able to predict students' performance "quite well," it does not take into account the possibility of disengagement. Traditionally, knowledge tracing is used with the probability of transition from a learned to an unlearned state set at 0, so students are never presumed to be forgetting the skill. When the forgetting transition is allowed, models such as knowledge tracing can become confounded, mistaking disengagement for unlearning, as illustrated in Figure 1.

2.2 Dynamic Mixture Model

Johns and Woolf [4] proposed another model, called the Dynamic Mixture Model (DMM), or Hidden Markov Model- Item Response Theory (HMM-IRT) model. In this model, rather than using Bayesian Knowledge Tracing, they use a hidden Markov model for tracing affective engagement, but pair it with a model for predicting student knowledge that relies on Item

Response Theory for the estimation of conditional probabilities between question and knowledge. Unlike BKT, this model estimates a single knowledge node. The benefit of this is that all problems can be examined together with the single overall "knowledge" node, rather than separating them out by skill, which is necessary in BKT, as knowledge estimations for that model can vary between skills. The dynamic mixture model allows the estimation of students' engagement at various time-steps (and relies on parameters of transitioning between engagement states), but assumes a single mastery node, without learning or forgetting parameters.

The result of that research was that adding the engagement component (top part of Figure 3) to the knowledge estimation model (bottom part of Figure 3) allowed for less of a decline in knowledge estimations after each question than the simple IRT model, which was apparently due to gaming behaviors and not due to lack of knowledge.



Model Parameters

P(A_0) = Probability of initial affect
P(L_0) = Probability of initial knowledge
P(B|A) = Probability of gaming behavior given affect
P(Q|A,K) = Probability of correct given affect and knowledge
P(C) = Probability of transfer affect

Node Representations

K = Knowledge
Q = Performance
A = Affect
B = Gaming Behavior

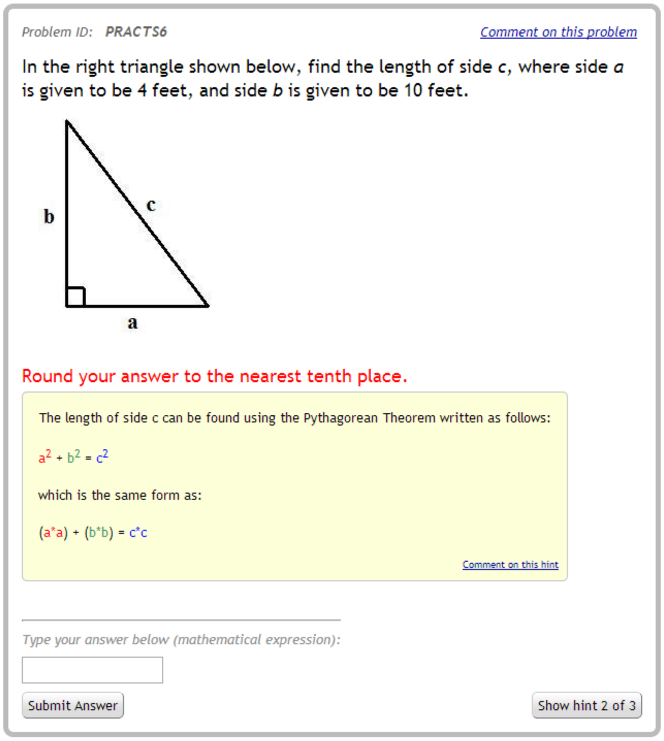
Figure 3- Dynamic Mixture Model

3. Data

The data used in this work was gathered from student logs of two mathematics tutoring systems, ASSISTments [6] and Wayang Outpost [5], for middle and high school students.

3.1 ASSISTments

ASSISTments is a tutor which allows teachers to create and assign problem sets, within which problems may be tagged with certain skills. Many problem sets will focus on a specific skill. Problems in ASSISTments generally require students to type in their answer, though some are multiple choice. Some questions include hints or scaffolding, which a student can ask for or they can be triggered after a student answers incorrectly. Figure 4 shows an example of an ASSISTments problem where the student has asked for one hint, which is shown in the yellow box.



Problem ID: ***PRACTS6***

In the right triangle shown below, find the length of side *c*, where side *a* is given to be 4 feet, and side *b* is given to be 10 feet.

Round your answer to the nearest tenth place.

The length of side c can be found using the Pythagorean Theorem written as follows:

$a^2 + b^2 = c^2$

which is the same form as:

(a*a) + (b*b) = c*c

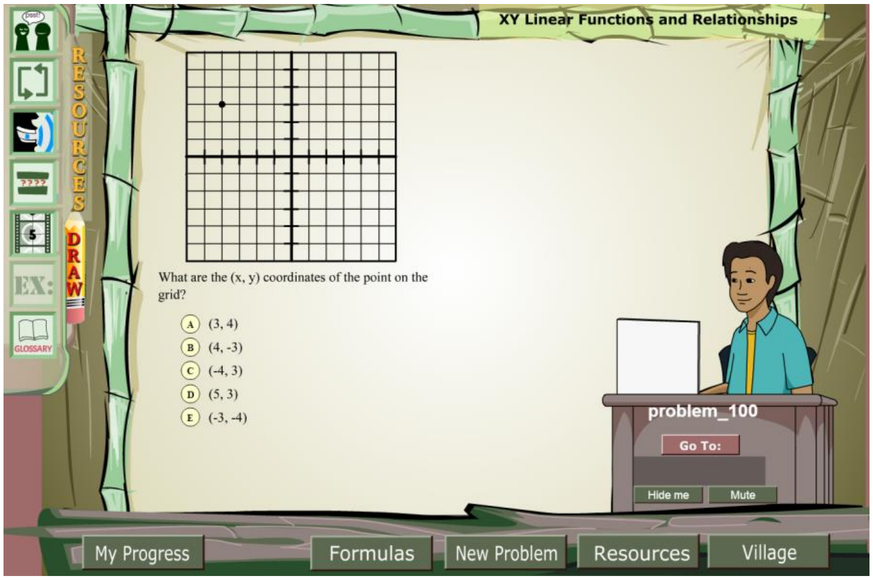
Type your answer below (mathematical expression):

Figure 4- ASSISTments problem with hint

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 answer three problems in a row correctly, in which case the skill is considered "mastered," or they reach a preset daily limit (usually ten questions) and are told to return later.

3.2 Wayang Outpost

In Wayang Outpost, problems are organized by topic. Teachers using Wayang may choose to turn off certain problems within a topic, but they cannot group problems in other ways. All problems in Wayang are multiple choice. Students may also ask for hints in Wayang, and the system includes a learning companion who will praise students for good effort or suggest that the student ask for hints when struggling. Figure 5 shows a question in the Wayang Outpost tutor.



What are the (x, y) coordinates of the point on the grid?

Ⓐ (3, 4)
Ⓑ (4, -3)
Ⓒ (-4, 3)
Ⓓ (5, 3)
Ⓔ (-3, -4)

Figure 5- Wayang Outpost

The Wayang data set used in this work 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.

3.3 Knowledge Components Used

Five knowledge components were chosen from ASSISTments and four from Wayang to explore, as the models examined are limited to examining each knowledge component separately. Table 1 shows the breakdown of the data used by knowledge component.

It was expected that students would game less in the ASSISTments data, as it comes from skill builders, in which a student's goal is to get three correct answers in a row in order to finish the problem set. Since incorrect answers mean more problems in this system, students might not be as tempted to game in order to get through the problem set. However, although three of the ASSISTments skills showed a lower incidence of gaming than any Wayang topic, the other two, "Circle Graph" and "Equations," had the highest amount of gaming of any knowledge component examined. Students gamed these skills approximately 30% and 35% of the time, respectively. The least gamed skill was "Table," where students exhibited gaming behavior only 4% of the time. The amount of gaming was more consistent in Wayang, ranging from 15 to 20 percent. Overall, the data examined includes a good range in the amount of gaming behaviors exhibited.

Table 1- Knowledge Components Examined

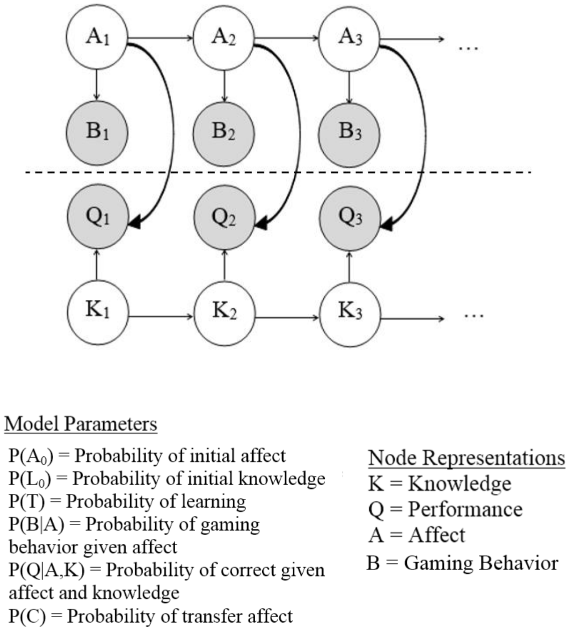
Knowledge Component	**System**	**Number Students**	**Total Number Opportunities**	**% Gaming**
Box and Whisker	ASSISTments	505	2020	13
Circle Graph	ASSISTments	616	2487	30
Table	ASSISTments	713	2894	4
Pythagorean Theorem	ASSISTments	283	1290	10
Equations	ASSISTments	408	1598	35
Perimeter	Wayang	285	1422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

4. Study 1- Knowledge and Affect Tracing Models

This study appeared in the Proceedings of the 7th International Conference on Educational Data Mining, London, UK, July 4-7, 2014.

4.1 The KAT Model

The first new model created and examined in this work is the Knowledge and Affect Tracing (KAT) model, shown in Figure 6. This model combines Knowledge Tracing with the affect tracing hidden Markov model portion of the dynamic mixture model, creating a model which allows for change in both students' knowledge and affective states. In this model, both of the knowledge and affect states influence performance. The dynamic mixture model does not allow for learning during the use of the tutor, but by modeling the students' current state in full, it should be possible to better predict performance and behavior (gaming or not gaming) at the next step.



Model Parameters

$P(A_0)$ = Probability of initial affect
$P(L_0)$ = Probability of initial knowledge
$P(T)$ = Probability of learning
$P(B|A)$ = Probability of gaming behavior given affect
$P(Q|A,K)$ = Probability of correct given affect and knowledge
$P(C)$ = Probability of transfer affect

Node Representations

K = Knowledge
Q = Performance
A = Affect
B = Gaming Behavior

Figure 6- The KAT Model

The behaviors examined when using this model were the same as those used by Johns and Woolf [4]. These are 1) quick guess (the student makes an attempt in less than four seconds), 2) bottom out hint (the student uses all available hints), and 3) normal (any other behavior). One additional behavior, 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 almost every choice in order to progress without solving the problem. Baker, et al, have also shown relatively few attempts to be a predictor of engaged concentration [11]. 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. Given that it is possible to skip questions in Wayang, doing this was also considered a "gaming" behavior, as the student does not solve the problem when doing this. The 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, it is possible that students could act in a way that is defined here as a gaming behavior even when they are engaged, just as they could possibly answer a question correctly even when they do not know the skill.

The conditional probability tables of the observed nodes of the KAT model are shown in Tables 2 and 3. Table 2 shows the CPT for the performance (Q) node. Knowing the skill, being engaged, answering a question correctly, and behaving normally (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 2- 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 3, and depend on affective engagement. These probabilities distinguish whether a student has behaved in a way considered gaming in a situation when s/he was actually engaged (some sort of an 'affective slip') corresponding to 'game_engaged' and its counterpart, where the student was actually disengaged but apparently behaved normally at this time-step (1-game_not_eng).

Table 3- CPT for Gaming Behavior (B) Nodes of KAT Model

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

4.2 The KAT2 Model

San Pedro et al. showed that student knowledge of a skill is related to affect (for example, students who do not know a skill well are more likely to be ~~engaged~~frustrated and become disengaged) [9], so a variation on the KAT model was created to take this into account. This model, KAT2, includes the link between knowledge and affect, except for at the first time-step where each simply contain a prior probability. The KAT2 model is shown in Figure 7. The conditional probability table for its affect nodes (except for the one at time 1) is shown in Table 4.

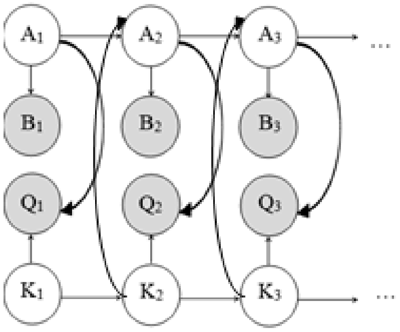


Figure 7- KAT2 Model

Table 4- CPT for Affect Nodes of KAT2

Knowledge (Latent)	**Previous Affect (Latent)**	**Current Affect (Latent)**	**Probability**
False	False	False	1-unknow_get_eng
True	False	False	1-know_get_eng
False	True	False	unknow_get_diseng
True	True	False	know_get_diseng
False	False	True	unknow_get_eng
True	False	True	know_get_eng
False	True	True	1-unknow_get_diseng
True	True	True	1-know_get_diseng

4.3 Methods

All models (BKT, DMM, and the two KAT models) were built using Murphy's Bayes Net toolbox for MATLAB [8]. A student-level five-fold cross validation [13] 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 all models except BKT, 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. In this first study, forgetting was not allowed. The reasoning for this was that all five opportunities are likely to occur in the same session, which would not allow time for material to be forgotten.

4.4 Results

The following tables show the average RMSE and MAE across folds for each knowledge component examined. For each row, the lowest (best) value is in italics. When that model was significantly better able to predict than the others (2-tailed paired t-test, p<0.05), the value is also bold.

Table 5 – RMSE Performance ASSISTments

Skill	**BKT**	**DMM**	**KAT**	**KAT2**
Box and Whisker	***0.426***	0.495	0.468	0.493
Circle Graph	***0.433***	0.524	0.507	0.512
Table	***0.467***	0.498	0.483	0.495
Pythagorean Theorem	***0.480***	0.498	0.484	0.504
Equations	*0.474*	0.498	0.484	0.495

Table 6- MAE Performance ASSISTments

Skill	**BKT**	**DMM**	**KAT**	**KAT2**
Box and Whisker	***0.363***	0.495	0.459	0.491
Circle Graph	***0.376***	0.522	0.506	0.512
Table	***0.436***	0.497	0.469	0.492
Pythagorean Theorem	***0.459***	0.498	0.480	0.502
Equations	***0.448***	0.497	0.472	0.492

Table 7 – RMSE Behavior ASSISTments

Skill	**DMM**	**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*
Equations	0.497	*0.451*	0.460

Table 8- MAE Behavior ASSISTments

Skill	**DMM**	**KAT**	**KAT2**
Box and Whisker	***0.134***	0.155	0.175
Circle Graph	0.052	***0.049***	0.060
Table	0.427	***0.336***	0.351
Pythagorean Theorem	*0.139*	0.145	0.150
Equations	0.410	***0.374***	0.387

Table 9- RMSE Performance Wayang

Topic	**BKT**	**DMM**	**KAT**	**KAT2**
Perimeter	*0.499*	0.510	0.507	0.501
Area	*0.490*	0.501	0.495	0.495
Angles	***0.484***	0.497	0.494	0.491
Triangles	*0.496*	0.505	0.505	0.502

Table 10- MAE Performance Wayang

Topic	**BKT**	**DMM**	**KAT**	**KAT2**
Perimeter	0.497	0.488	0.488	***0.484***
Area	*0.479*	0.501	0.483	0.484
Angles	*0.468*	0.484	0.470	0.471
Triangles	*0.491*	0.499	0.493	0.492

Table 11- RMSE Behavior Wayang

Topic	**DMM**	**KAT**	**KAT2**
Perimeter	*0.391*	*0.391*	*0.391*
Area	0.451	***0.4436***	0.444
Angles	0.375	0.369	*0.3688*
Triangles	0.434	0.429	*0.4289*

Table 12- MAE Behavior Wayang

Topic	**DMM**	**KAT**	**KAT2**
Perimeter	0.327	*0.326*	0.328
Area	0.433	*0.413*	*0.413*
Angles	0.302	*0.289*	0.290
Triangles	0.404	***0.393***	0.397

These tables show that BKT is overall the best predictor of student performance, despite the lack of knowledge about a student's behaviors or affective engagement, at least when forgetting is not allowed.   The two KAT models generally outperform DMM at predicting performance (the one exception is for KAT2 on the ASSISTments skill "Pythagorean Theorem," but this difference is not significant). The original KAT model was significantly better at predicting performance than the KAT2 model on the ASSISTments data (p<0.05), except for RMSE on the skill "Table" (p=0.09). For Wayang, the KAT2 model was significantly better at predicting performance than KAT for both error metrics in the "Perimeter" and "Triangles" topics. Both KAT models are also generally better at predicting behavior than the dynamic mixture model. The only exceptions are the MAEs for the ASSISTments skills "Box and Whisker" and "Pythagorean Theorem," where DMM did better, and the RMSE of the Wayang topic "Perimeter," where all three models performed about the same. The RMSEs of the two KAT models were not significantly different with respect to predicting behavior, except for on the ASSISTments skill "Table" and the Wayang topic "Area," on which the original KAT model performed better. The MAEs of the original KAT model were significantly lower than those of the KAT2 model on all ASSISTments skills and the Wayang topic "Triangles," and not significantly different on the other Wayang topics.

4.5 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 knowledge performance and gaming behaviors simultaneously.

The difference between the dynamic mixture model and the KAT model is that KAT allows for student learning. 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 DMM, leads to a poorer model fit.

5. Study 2-KTB

This study appeared in the Proceedings of the Workshop on Approaching 20 Years of Knowledge Tracing at the 7th International Conference on Educational Data Mining, London, UK, July 4-7, 2014.

5.1 The Knowledge Tracing with Behavior Model

Since the KAT models were not able to predict performance as well as BKT in study one, a different model for examining both behavior and performance was created in an attempt to meet the goal of predicting both. This model, the Knowledge Tracing with Behavior (KTB), model has only one latent node, which we call "knowledge"-- although in reality is a combination of both knowledge and engagement-- and two observables, performance and gaming behaviors. This model is shown in Figure 8. This model also has fewer parameters than the dynamic mixture model or KAT model, but still can predict both performance and gaming behavior of the students.

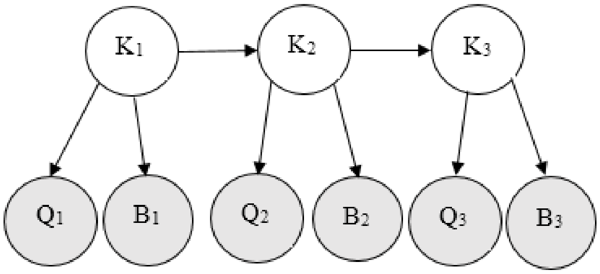


Figure 8- KTB Model

5.2. Bayesian Engagement Tracing

In study one, models were compared against BKT on their prediction of performance. However, a similar model for engagement and behavior was absent. To that end, a model of "Bayesian Engagement Tracing" (BET) is included in this work, which is the same as the HMM part of Johns and Woolf's model or the engagement piece of the KAT model, but not connected to any other model (top part of Figures 3 and 6). This is similar to the “gaming tracing” model proposed by Giguere, et al [17], except that they examined only response time as an observable. This model is useful for comparing to the more complicated models that include these nodes.

5.2 Methods

In general, the methods used in this study were the same as in study one. This time, however, forgetting was allowed in all models containing a knowledge node at each time-step, in order to examine whether combined models would do better than BKT when that model could confuse disengagement with forgetting. The BKT and BET models were compared against the Johns and Woolf Dynamic Mixture Model, the original KAT model from study one, and the KTB model.

5.3 Results

The results of this study are reported in the following tables, in the same manner as those for study one were presented in the previous chapter.

Table 13- RMSE of Performance Prediction for ASSISTments

Skill	BKT	KTB	KAT	DMM
Box and Whisker	0.427	*0.425*	0.468	0.495
Circle Graph	0.437	***0.432***	0.505	0.523
Table	0.469	*0.467*	0.483	0.498
Pythagorean Theorem	0.479	*0.476*	0.485	0.497
Equations	0.476	*0.472*	0.484	0.498

Table 14- MAE of Performance Prediction for ASSISTments

Skill	BKT	KTB	KAT	DMM
Box and Whisker	0.365	*0.364*	0.457	0.495
Circle Graph	0.382	***0.376***	0.504	0.522
Table	0.441	*0.439*	0.470	0.497
Pythagorean Theorem	0.458	***0.453***	0.480	0.498
Equations	0.452	*0.448*	0.472	0.498

Table 15- RMSE of Performance Prediction for Wayang

Topic	**KT**	**KTB**	**KAT**	**DMM**
Perimeter	*0.498*	0.499	0.507	0.505
Area	0.491	*0.490*	0.494	0.499
Angles	*0.487*	0.490	0.495	0.497
Triangles	*0.498*	0.499	0.505	0.504

Table 16- MAE of Performance Prediction for Wayang

Topic	**KT**	**KTB**	**KAT**	**DMM**
Perimeter	0.494	0.496	0.491	*0.489*
Area	0.480	*0.478*	0.483	0.499
Angles	0.473	0.476	***0.467***	0.484
Triangles	0.495	0.495	*0.494*	0.499

Table 17- RMSE of Behavior Prediction for ASSISTments

Skill	**BET**	**KTB**	**KAT**	**DMM**
Box and Whisker	***0.317***	0.322	0.326	0.350
Circle Graph	*0.177*	0.184	0.178	0.194
Table	*0.415*	0.421	0.423	0.463
Pythagorean Theorem	*0.287*	*0.287*	*0.287*	0.294
Equations	***0.442***	0.447	0.451	0.504

Table 18- MAE of Behavior Prediction for ASSISTments

Skill	**BET**	**KTB**	**KAT**	**DMM**
Box and Whisker	0.202	0.207	0.155	***0.134***
Circle Graph	0.064	0.071	*0.049*	0.052
Table	0.344	0.355	*0.337*	0.426
Pythagorean Theorem	0.169	0.170	0.142	*0.135*
Equations	0.391	0.400	*0.376*	0.414

Table 19- RMSE of Behavior Prediction for Wayang

Topic	BET	KTB	KAT	DMM
Perimeter	*0.433*	0.435	0.439	0.438
Area	*0.449*	0.450	0.457	0.464
Angles	*0.396*	0.398	0.403	0.413
Triangles	***0.433***	0.435	0.442	0.446

Table 20- MAE of Behavior Prediction for Wayang

Topic	BET	KTB	KAT	DMM
Perimeter	***0.374***	0.378	0.405	0.403
Area	*0.401*	*0.401*	0.434	0.452
Angles	***0.312***	0.315	0.345	0.367
Triangles	***0.373***	0.378	0.412	0.421

We can see from these tables that KTB is generally the best predictor of performance, although not always significantly so, while BET tends to be the best predictor of gaming behavior.

The following two charts show the average predictions for performance (question correctness) of BKT and KTB at each time step (dotted lines) against the actual average performance at that time step (solid line). One representative knowledge component from each system is shown here.

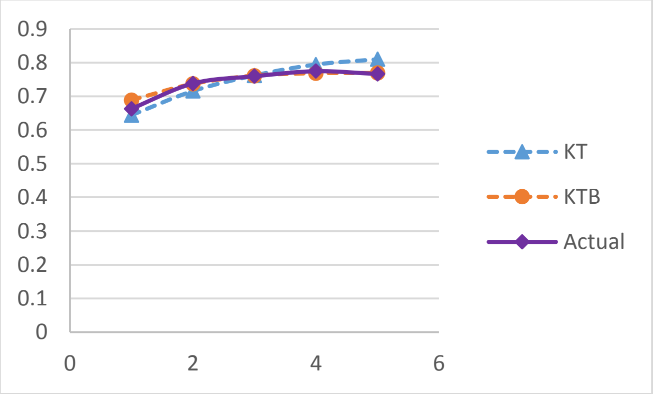


Figure 9- Predicted vs. Actual Performance on "Box and Whisker"

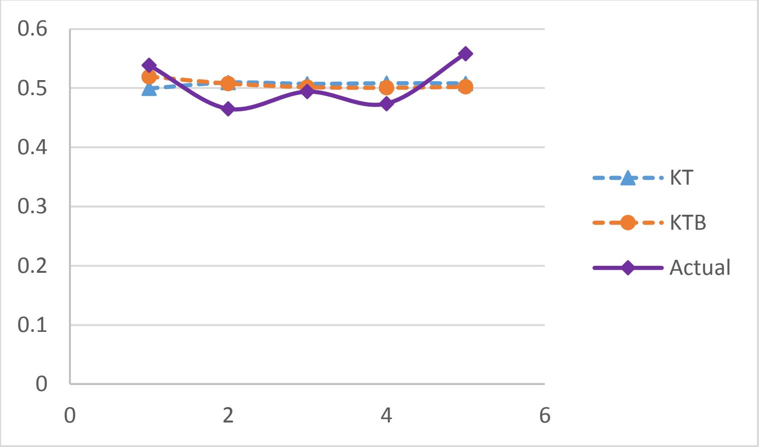


Figure 10- Predicted vs. Actual Performance on "Triangles"

Looking at these charts, the two models appear to have very similar predictions at each time step, although in Figure 9 the orange KTB line does appear to be slightly closer to the actual data.

5.4 Latent Values

Additionally, the predictions of the latent node for each model were examined. Figures 11 to 13 show the average latent value at each of the five time slices for a specific knowledge component. These three graphs show the three patterns that emerged when charting the latent values for all nine knowledge components.

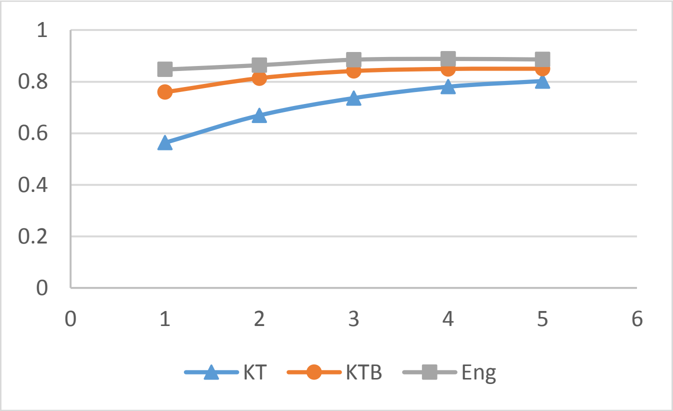


Figure 11- Latent Predictions of ASSISTments skill "Box and Whisker"

Figure 11 shows the latent predictions for the ASSISTments skill “Box and Whisker.” While knowledge appears to increase in both BKT and KTB, the prediction at the first opportunity (prior) is higher in KTB and less learning appears to take place over the five problems, whereas BKT starts with a lower prior and catches up. In an ASSISTments skill builder, students who “master” the skill quickly will drop off after they answer three questions correctly

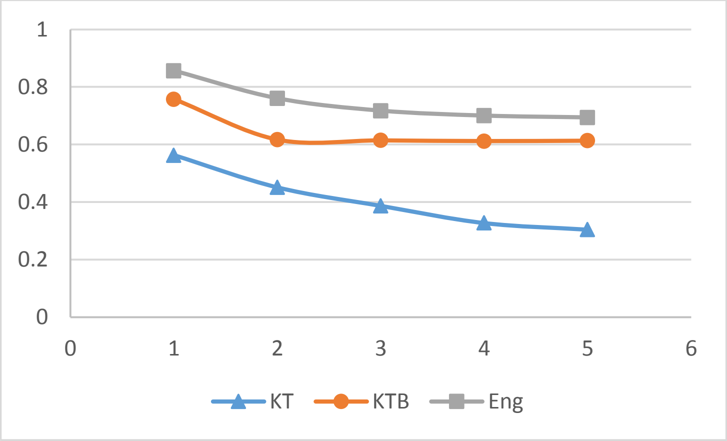


Figure 12- Latent Predictions of Wayang topic “Perimeter”

In the Wayang topic “Perimeter,” BKT predicts that students are “forgetting” throughout the session. However, by looking at the prediction of engagement, it is clear that they are becoming less engaged and this might be contributing to the appearance of unlearning. The KTB latent decreases from the first opportunity to the second, but then remains relatively flat, rather than giving the impression that students continue to forget.

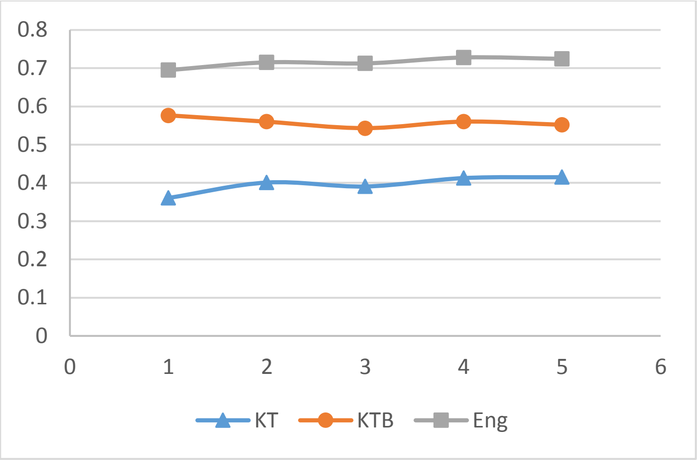


Figure 13- Latent Predictions of Wayang topic "Area"

Finally, the Wayang topic "Area" results in relatively flat curves for all three latent nodes. However, KTB predicts that the prior "knowledge" is higher than BKT does. It is important to note that Wayang Outpost has adaptive problem difficulty selection, which is likely affecting the flatness of the results compared to ASSISTments. This shows that, whenever problem difficulty adjustment is carried out, problem difficulty should be included in the model (probably as a different node); otherwise, knowledge tracing will believe that knowledge remains stagnant, when actually problems assigned to the student are getting harder to solve.

5.4 Discussion

In Figures 11-13, the KTB latent generally lies between the knowledge latent from BKT and the engagement latent from BET; this makes sense, since it is a combination of the two. It is interesting that in many cases this combined model predicts performance and behavior as well, or marginally better, than the two separate HMMs.

6. Study 3- All Models + KAT New

6.1 The KAT New Model

Based on the results of the first two studies, a new model was created and examined. This model, KAT New, is shown in Figure 14. Like the KTB model, it allows for knowledge to impact both performance and behavior, and it adds the latent engagement node back in, which now impact only behavior.

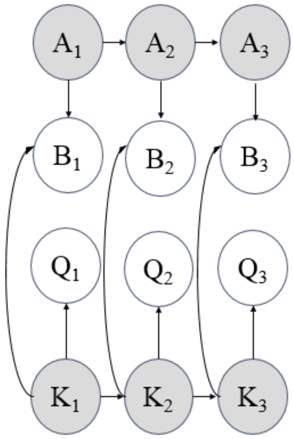


Figure 14- KAT New

6.2 KAT New 2

Since we originally suspected that performance would be impacted by affect, but the KAT New model does not include this link, we also created a version of the KAT New model that reintroduces it, KAT New 2. However, at this point the latent nodes appear identical, as both influence both observables, so we did not expect this model to perform well.

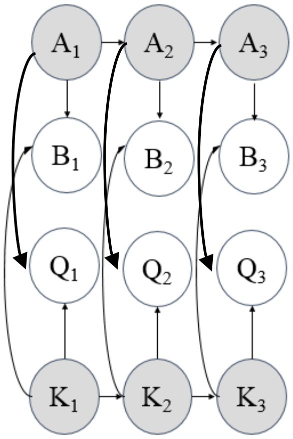


Figure 15- KAT New 2

6.3 Methods

As in the previous experiments, all models were created using the Bayes Net toolbox for MATLAB [8] and a student-level five-fold cross-validation was run. All models from the previous studies were examined, with the addition of the two new models. As in study two, forgetting was allowed. In addition to error metrics based on the probabilistic predictions of each model, we also binarized these predictions in order to calculate accuracy and kappa.

6.4 Results

While both error metrics based on the probabilistic predictions of the observables and accuracies and kappa values based on binary predictions were calculated, there were no knowledge components for which any one model was significantly better than all others, so we here focus on the accuracy and kappa values. Additionally, a baseline error rate was missing in the previous studies. Here, we use a majority class prediction as the baseline for accuracy.

Tables 21 and 22 show the accuracies of each model's prediction of performance compared to the accuracy of a majority class prediction method for each skill and topic.

Accuracies higher than majority class are marked in green and those lower than majority class are marked in red. The highest accuracy for each skill or topic is marked in bold.

Table 21- Accuracy of Each Model for Performance in ASSISTments

Skill	KT	KAT	KAT2	KAT New	KAT New 2	KTB	DMM	Maj. Class
Box and Whisker	0.734	0.734	0.392	0.734	0.324	**0.744**	0.734	0.734
Circle Graph	0.663	0.612	0.452	**0.668**	0.596	0.667	0.554	0.507
Table	0.718	0.381	0.282	0.736	0.275	**0.737**	0.282	0.718
Pythagorean Theorem	0.631	0.636	0.427	**0.643**	0.481	**0.643**	0.578	0.634
Equations	0.650	0.601	0.475	0.656	0.507	**0.660**	0.540	0.507

Table 22- Accuracy of Each Model for Performance in Wayang Outpost

Topic	KT	KAT	KAT2	KAT New	KAT New 2	KTB	DMM	Maj. Class
Perimeter	0.517	0.529	**0.533**	0.512	0.520	0.518	0.529	0.529
Area	0.570	**0.584**	0.576	0.560	0.541	0.560	0.527	0.544
Angles	**0.607**	0.596	0.599	0.603	0.481	0.602	0.572	0.572
Triangles	0.519	0.530	**0.539**	0.518	0.493	0.520	0.506	0.506

We can see from the above tables that in ASSISTments KTB is the only model that consistently has a higher accuracy in predicting performance than majority class while KAT New always does at least as well as majority class. On the other hand, in Wayang Outpost, KAT2 is the only model that consistently does better than majority class and KAT always does at least as well. In ASSISTments, we could actually do at least as well as majority class in three of five skills by predicting the opposite of KAT2's prediction, although other models would still do better than this "complement of KAT2" prediction. It is interesting that KAT2 does well in Wayang Outpost, since in ASSISTments it ~~does not predict as well as majority class~~would be better to predict the opposite. This may indicate that different models will be better suited to different systems, although using better initial parameters to seed expectation maximization, such as Dirichlet priors [18], could help with the complement issue. Overall, the models that seem to perform the best at predicting performance are KAT New and KTB, which each only have one knowledge component, "Perimeter", where they do not predict performance at least as well as majority class.

Tables 23 and 24 show the Cohen's kappa agreements between the predictions of each model and the actual performance. The highest kappa for each knowledge component is marked in bold.

Table 23- Kappa for Performance in ASSISTments

Skill	**KT**	**KAT**	**KAT2**	**KAT New**	**KAT New2**	**KTB**	**DMM**
Box and Whisker	0.000	0.000	-0.050	0.085	-0.096	**0.176**	0.000
Circle Graph	0.327	0.218	-0.089	**0.338**	0.188	0.334	0.112
Table	0.000	-0.136	0.000	0.148	-0.028	**0.149**	0.000
Pythagorean Theorem	0.010	0.062	-0.041	**0.078**	-0.008	0.074	-0.006
Equations	0.297	0.208	-0.062	0.310	0.000	**0.318**	0.079

Table 24- Kappa for Performance in Wayang Outpost

Topic	**KT**	**KAT**	**KAT2**	**KAT New**	**KAT New2**	**KTB**	**DMM**
Perimeter	0.019	0.000	0.011	-0.012	**0.072**	-0.004	0.000
Area	0.112	**0.188**	0.168	0.103	-0.004	0.103	0.024
Angles	**0.153**	0.105	0.107	0.131	0.045	0.129	0.000
Triangles	0.037	**0.052**	0.070	0.034	-0.002	0.037	0.000

These kappa values show similar results to the above accuracy values. KAT New and KTB are the only models with kappa above 0 for all ASSISTments skills, although they have small negative values for the Wayang topic "Perimeter," indicating that they are slightly worse than chance at predicting performance on this topic. Again, KAT2 does fairly well in Wayang, with all kappa values greater than 0, but does not perform better than chance on any ASSISTments skill. Traditional Bayesian Knowledge Tracing also achieves a kappa of at least 0 for every knowledge component.

In addition to how well each model predicted performance, we want to see how well they predict our other observable, gaming behavior. Once again, we compare the accuracies of each model against a majority class predictor. This is shown in Tables 25 and 26. Again, values greater than the majority class prediction are marked in green and those lower are red while the highest accuracy for each knowledge component is bold.

Table 25- Accuracies for Behavior in ASSISTments

Skill	**BET**	**KAT**	**KAT2**	**KAT New**	**KAT New 2**	**KTB**	**DMM**	**Maj. Class**
Box and Whisker	0.874	0.874	0.874	0.874	**0.875**	0.874	0.874	0.874
Circle Graph	**0.739**	0.724	0.650	0.727	0.676	0.728	0.697	0.697
Table	0.962	0.961	0.960	0.960	**0.962**	0.960	0.960	0.960
Pythagorean Theorem	0.905	0.905	0.905	0.905	0.900	0.905	0.905	0.905
Equations	**0.687**	0.670	0.619	0.678	0.646	0.678	0.646	0.646

Table 26- Accuracies for Behavior in Wayang Outpost

Skill	**BET**	**KAT**	**KAT2**	**KAT New**	**KAT New 2**	**KTB**	**DMM**	**Maj. Class**
Perimeter	0.745	0.745	0.742	0.745	0.674	0.745	0.745	0.745
Area	0.711	0.686	0.697	0.711	0.591	0.711	0.711	0.711
Angles	0.792	0.792	0.792	0.792	0.751	0.792	0.792	0.792
Triangles	0.740	0.740	0.740	0.740	0.694	0.740	0.740	0.740

We see from these tables that none of the models allow us to predict behavior better than majority class in Wayang, while Bayesian Engagement Tracing and KAT allow us to predict better than majority class in three of five ASSISTments skills, and as well in the other two, and KAT New and KTB allow us to predict better in two of five skills and as well in the other three. While KAT2 did well at predicting performance in Wayang, it actually performs worse than majority class at predicting behavior in two of four topics.

Tables 27 and 28 show the kappa values for the agreement between each model's prediction of behavior and the actual behaviors. The highest kappa for each knowledge component is again marked in bold.

Table 27- Kappa for Behavior in ASSISTments

Skill	**BET**	**KAT**	**KAT2**	**KAT New**	**KAT New 2**	**KTB**	**DMM**
Box and Whisker	0.000	0.000	0.000	0.000	**0.036**	0.000	0.000
Circle Graph	0.320	0.282	0.127	0.315	**0.347**	0.316	0.000
Table	0.203	0.164	0.000	0.000	**0.245**	0.000	0.000
Pythagorean Theorem	0.000	0.000	0.000	0.000	**0.028**	0.000	0.000
Equations	**0.273**	0.203	0.110	0.268	0.000	0.268	0.000

Table 28- Kappa for Behavior in Wayang Outpost

Topic	**BET**	**KAT**	**KAT2**	**KAT New**	**KAT New 2**	**KTB**	**DMM**
Perimeter	0.000	0.000	0.009	0.000	**0.082**	0.000	0.000
Area	0.000	0.020	0.026	0.000	**0.082**	0.000	0.000
Angles	0.000	0.000	0.000	0.000	**0.167**	0.000	0.000
Triangles	0.000	0.000	0.000	0.000	**0.137**	0.000	0.000

Interestingly, KAT New 2 generally has the highest kappa values for behavior even though there are only two knowledge components for which it predicts behavior with a higher accuracy than majority class. In most cases, the models are able to predict behavior at chance and only in a few cases can they predict better than chance.

6.5 Example Cases

We now examine some example students from the data. All predictions of latent nodes were saved along with the predictions of observables. These predictions are based on the model as fit to the four other folds. The students we will examine include one “good” student who gets answers correct and does not perform gaming behaviors, one “engaged struggling” student who is not able to get all answer correct but does not perform gaming behaviors, and one “gaming” student who consistently performs gaming behaviors.

The first student was a “good” student in the ASSISTments skills “Equations.” This student did not exhibit gaming behavior and only answered one question incorrectly, which was the third one. We would expect that this student is likely engaged and likely to know this skills, although perhaps estimation of knowledge would decrease after the third opportunity where s/he answers incorrectly. Figure 16 shows the knowledge estimation at each time step (before observing the current time step’s behavior and performance) for each model while Figure 17 shows the engagement estimations. In this case we include KTB’s latent as knowledge and the dynamic mixture model is shown only in the engagement estimation since it has only one knowledge node.

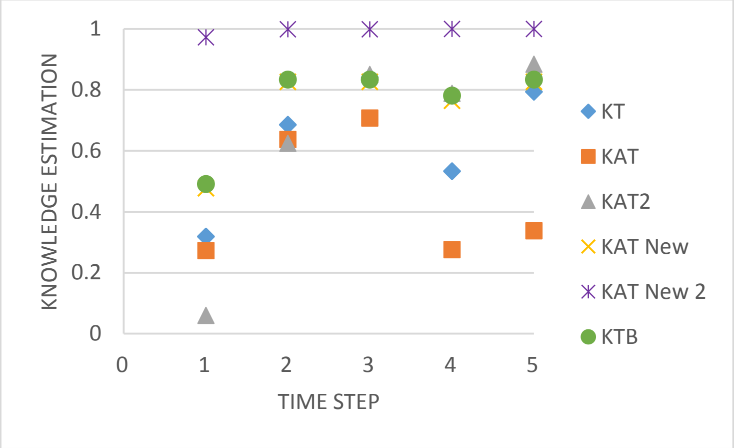


Figure 17- Knowledge Estimations of a "Good" Student

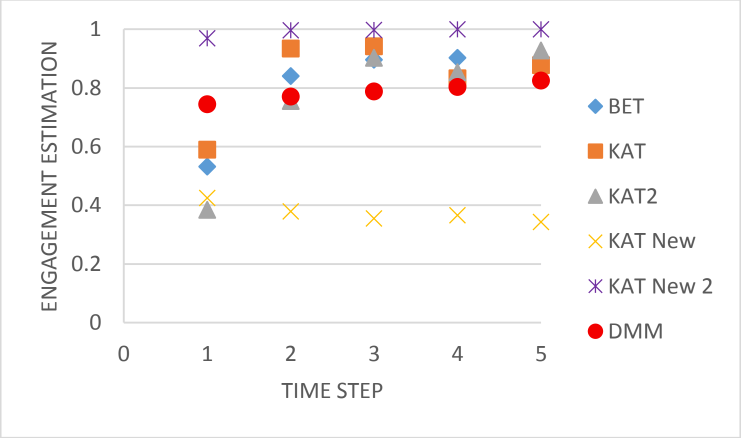


Figure 18- Engagement Estimations of a "Good" Student

We can see in Figure 17 that all of the models generally follow the expected trend in their prediction of knowledge. After the student gets the third question wrong, the estimated probability that s/he knows the skills dips but otherwise the trend is upward. KAT New 2, which has a very high prior, is the only model that does not clearly show a dip at time four. BKT and KAT appear to be strongly affected by the one incorrect answer while the other models' estimations only dip slightly.

Since the student is not gaming and is doing well, we expect that the engagement estimations will not decrease over time. Looking at Figure 18, KAT New is the only model that seems to do the opposite of what we expect overall while KAT and KAT2 show a dip in engagement estimation after the student gets an answer incorrect since it is possible in those models that being less engaged could cause a slip.

The next student examined is one who appears to be struggling in the Wayang topic "Area." S/he did not answer any questions correctly, but also did not exhibit any gaming behavior. We would expect that s/he is unlikely to know the skill, but likely to be engaged in attempting to answer the questions correctly. The knowledge and engagement estimations for this student are shown in Figure 19 and 20.

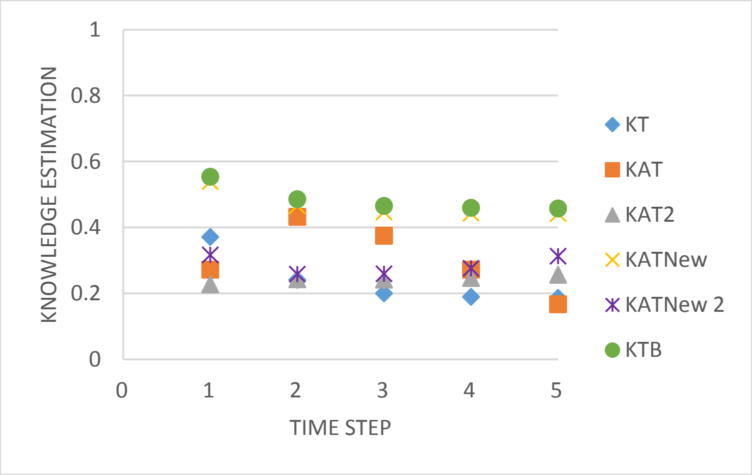


Figure 19- Knowledge Estimations of a Struggling Student

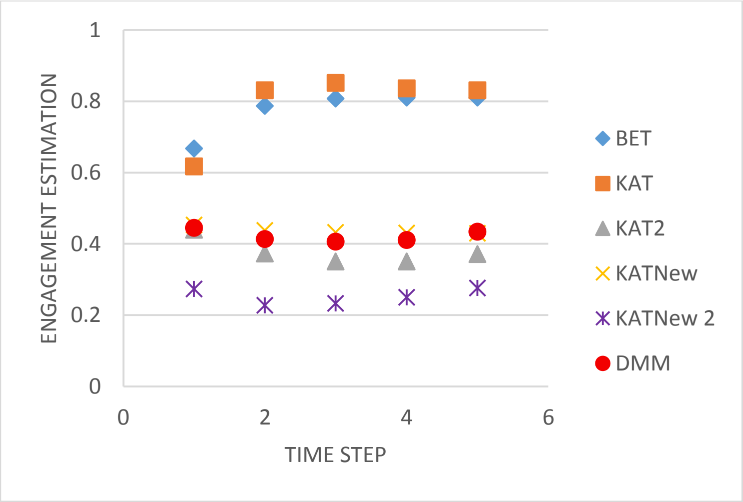


Figure 20- Engagement Estimations of a Struggling Student

We can see in Figure 19 that all models have a low and overall decreasing estimate of the student's knowledge, as expected. However, in Figure 20 we see that only Bayesian Engagement Tracing and KAT have an estimate of student engagement over 50% at any time.

The final student we examine is a student who is gaming in the ASSISTments skill "Circle Graph." This student did not answer any questions correctly and also exhibited gaming behavior on each question. We expect that engagement will be low and it will be difficult to accurately estimate knowledge. The knowledge and engagement estimations for this student are shown in Figure 21 and 22.

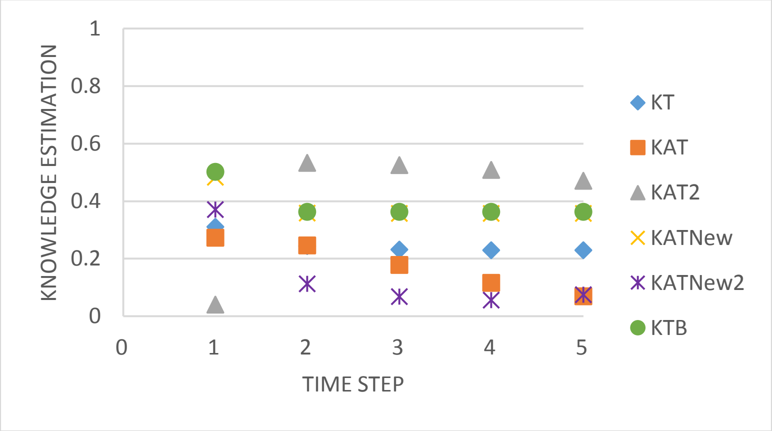


Figure 21- Knowledge Estimations for a Gaming Student

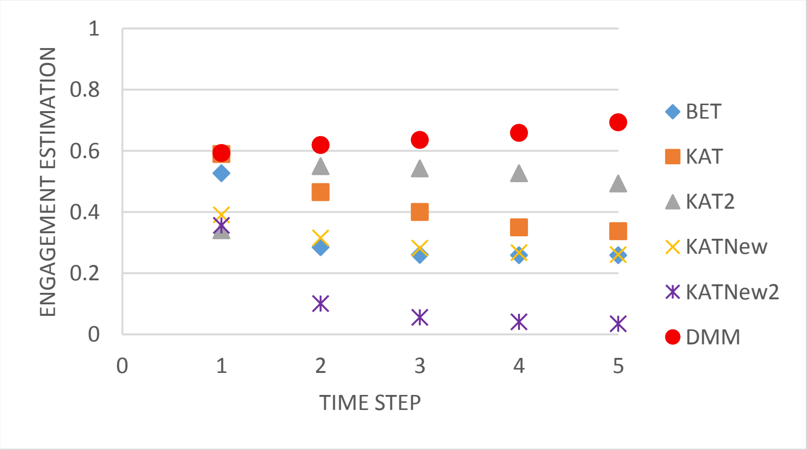


Figure 22- Engagement Estimations for a Gaming Student

Other than the dynamic mixture model, the models generally follow the expected trend of decreasing the engagement estimations at each time step. Most of the models show a slight decrease in knowledge estimate at each time step, since the student is getting the answers incorrect, but they tend not drop off as quickly as the engagement estimation since disengagement could account for some of the incorrect answers. Surprisingly, both of KAT2's estimations increase between time steps one and two even though the student gamed and answered the question incorrectly at time one. The other KAT models, KTB, and the separate hidden Markov models appear to follow the patterns we expect for this student.

Overall, many of the models examined appear to follow the expected patterns of knowledge and engagement estimations for these three students. It is interesting to note that the knowledge and engagement predictions tended to be almost identical in the case of KAT New 2, where both latent nodes impact both observables, indicating that separating out these latent variables in this way was not useful. The models that best follow the expected trends are the separate hidden Markov models (Bayesian Knowledge Tracing and Bayesian Engagement Tracing) and KTB. Each of the other models appears to make an unexpected estimation at least once.

6.6 Discussion

Examining the accuracy and kappa values of each model in predicting performance and behavior, it appears that KTB, KAT New, and the separate hidden Markov models have the highest accuracy and the same models have the highest kappa values in predicting performance, although KAT New 2 generally has the highest kappa values in predicting behavior. Overall, from this data we conclude that KTB, KAT New, and separate BKT and BET models all appear to predict the observables well, with KTB and KAT New perhaps being able to predict knowledge slightly better than KTB and BET perhaps able to predict behavior slightly better than KTB or KAT New.

In examining the latent estimations of specific students, however, KAT New surprisingly estimated that even the student who was answering most of the questions correctly and did not exhibit gaming behavior was likely to be disengaged. BET and KAT were the only models that predicted the struggling student was likely engaged despite the incorrect answers. KAT, however, increased its estimation of this student's knowledge after s/he answered the first question incorrectly. KTB's estimate of "knowledge," really some combination of knowledge and engagement, appear to follow accurate trends- overall high for the "good" student, dropping slightly then leveling off for the struggling student, and dropping quite a bit then leveling off for the gaming student. Overall, KTB and separate Markov models seem to be the best for both predicting observables and estimating the latent variables.

7. Conclusions & Discussion

The contribution of this work is a first step toward sensor-less engagement detection and the ability for an ITS to differentiate between a student who does not know, or has forgotten, a skill, and a student who has simply become disengaged.

Overall, it appears that the simpler models are better able to predict future performance and behavior than more complicated ones. In study one, standard Bayesian Knowledge Tracing was generally able to predict next question correctness than the two KAT models or DMM, while in study two the models with one latent node (BKT, BET, and KTB) tended to outperform the others. This appears to be true across systems, as well, as the error was smaller in both datasets.

In study three, KAT New also performed well in predicting performance and behavior, but did not appear to have estimates of its latent nodes that were consistent with what was expected, while BKT, BET, and KTB's estimates better aligned to what we observed.

It is unclear whether the KTB model, with its single latent, is a better predictor of future performance and behavior than two separate hidden Markov models (BKT and BET), as it does appear to do slightly better when predicting performance than BKT, but not quite as well at predicting behavior as BET. This should be studied further with additional data in order to make a clearer assertion as to which is best. Future studies should also be careful in balancing the folds, rather than selecting them completely randomly as was done herein. Since gaming behavior only occurs a certain percentage of the time, it is possible that the training and test data were not always balanced- for example if all of the gaming behavior occurred in one fold, then a model trained on the other four would likely not fit this fold well.

8. Future Work

In this work, all nodes were binary in order to be able to predict probabilities. However, rather than examining just engagement versus disengagement, it would be useful to look at more specific affective states, such as frustration or interest. This would likely require separate models for various affects and/or multiple parameters relating behaviors to affect.

There are also additional variations on the KAT model that could be investigated, for example, adding a link from performance at one time step to affect at the next. However, given the result that simpler models appear to perform better than more complicated ones, this seems less promising. It is also important to note that the gaming behaviors here examined and performance are not actually independent, as once a student asks for a hint or makes an incorrect attempt s/he will be marked as incorrect. Therefore, future models should take this into account. One possibility is to separate out the "gaming behavior" observable into two nodes, time to first action, and type of first action (attempt or hint). Asking for a hint would automatically mean the answer is incorrect, since that is how the system works, and a quick attempt would mean a different probability of correctness than a slow attempt.

It would also be useful to compare these sensor-less models to existing models for engagement detection that use sensors, such as in [10], or observations such as BROMP [15] in order to determine how well we can do without sensors or observers as compared to with them. If sensors or observations lead to significantly better results, it might be preferable to use these, when possible, whereas if we can do as well or almost as well with a sensor-less method, this might be preferred.

Once these models have been thoroughly explored, the next step is to integrate them into the system in order to provide more accurate interventions within the tutor. The Math Spring system (the new version of Wayang Outpost) will use the predictions of affective engagement from the model in order to adjust the difficulty of problems or intervene in a way intended to improve the student’s affective state. This way, the system will be better able to keep students engaged in order to help them to learn.

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] 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.

[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] Arroyo, I., Mehranian, H., & Woolf, B., "Effort-based tutoring: An empirical approach to intelligent tutoring." *Proceedings of the 3rd International Conference on Educational Data Mining,* 2010, pp. 1-10.

[6] Pardos, Z. A., Heffernan, N. T., Anderson, B., Heffernan, L. C. (2010). "Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks." Chapter in C. Romero, S. Ventura, S. R. Viola, M. Pechenizkiy and R. S. J. Baker. *Handbook of Educational Data Mining*. Boca Raton, Florida: Chapman & Hall/CRC Press.

[7] Efron, B. & Gong, G. (1983). "A leisurely look at the bootstrap, the jackknife, and cross-validation." In *American Statistician*, 37, 36-48.

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

[9] 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.

[10] 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.

[11] 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.

[12] S. D'Mello, S. Craig, B. Gholson, S. Franklin, R. Picard, and A. Graesser, "Integrating Affect Sensors in an Intelligent Tutoring System," Proc. *Computer in the Affective Loop Workshop* at 2005 Int'l Conf. Intelligent User Interfaces, pp. 7-13, 2005.

[13] Pardos, Z.A., Baker, R.S., San Pedro, M.O.Z., Gowda, S.M, and Gowda, S.M. "Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes." In *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge*, 2013, p. 117-124.

[14] Efron, B. & Gong, G. (1983). "A leisurely look at the bootstrap, the jackknife, and cross-validation." In *American Statistician*, 37, 36-48.

[15] 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.

[16] Arroyo, I., Ferguson, K., Johns, J., Dragon, T., Meheranian, H., Fisher, D., Barto, A., Mahadevan, S., Woolf, B.P. (2007) *Repairing Disengagement with Non-Invasive Interventions.* International Conference of AI in Education. IOS Press.

[17] Giguere, S., Beck, J., & Baker, R. (2010, January). Analyzing student gaming with bayesian networks. In Intelligent Tutoring Systems (pp. 321-323). Springer Berlin Heidelberg.

[18] Rai, D., Gong, Y., & Beck, J. E. (2009). Using Dirichlet Priors to Improve Model Parameter Plausibility. International Working Group on Educational Data Mining.