Towards Detecting Wheel-Spinning: Future Failure in Mastery Learning

Yue Gong
Facebook
yuegong@fb.com

Joseph E. Beck
Worcester Polytechnic Institute
josephbeck@wpi.edu

ABSTRACT

Wheel-spinning refers to a phenomenon in which a student has spent a considerable amount of time practicing a skill, yet displays little or no progress towards mastery. Wheel-spinning has been shown to be a common problem affecting a significant number of students in different tutoring systems and is negatively associated with learning. In this study, we construct a model of wheel-spinning, using generic features easily calculated from most tutoring systems. We show that for two different systems' data, the model generalizes to future students very well and can detect wheel-spinning in an early stage with high accuracy. We also refine the scope of the wheel-spinning problem in two systems using the model's predictions.

Keywords
Student modeling; mastery learning; wheel-spinning; behavioral detector; prediction;

INTRODUCTION

Mastery learning has been implemented and applied in intelligent tutoring systems (ITS) in a variety of contexts. One common foundation builds on the ACT-R theory, which assumes that procedural knowledge of a skill can be acquired through problem solving of what is initially declarative knowledge [1]. The rationale of mastery learning is also well supported by the theory of "learning-by-doing," which refers to the capability of learners to improve their efficiency by regularly repeating the same type of action via practice [2]. The use of mastery learning is driven by the desire to provide students efficient practice, by avoiding giving too many problems to solve, which could waste valuable learning time [3] and possibly jeopardize student motivation to learn, but simultaneously ensuring there are not too few practice problems, which might leave students poorly prepared for learning future content [4] due to the lack of mastery.

An application of mastery learning is that students are presented as many problems as needed to master the skill. Consequently, the system keeps giving the student more problems to practice in the hope that he might utilize these new opportunities to master the skill. The student however could keep failing to learn the skill, which triggers the system to present even more problems to the student. Thus, the student can possibly become trapped in the mastery learning cycle if he fails to achieve mastery. We term this phenomenon "wheel-spinning", analogous to a car stuck in mud or snow; its wheels are spinning rapidly, but it is not going anywhere. Similarly, students are being presented many problems, but they are not making progress towards mastery.

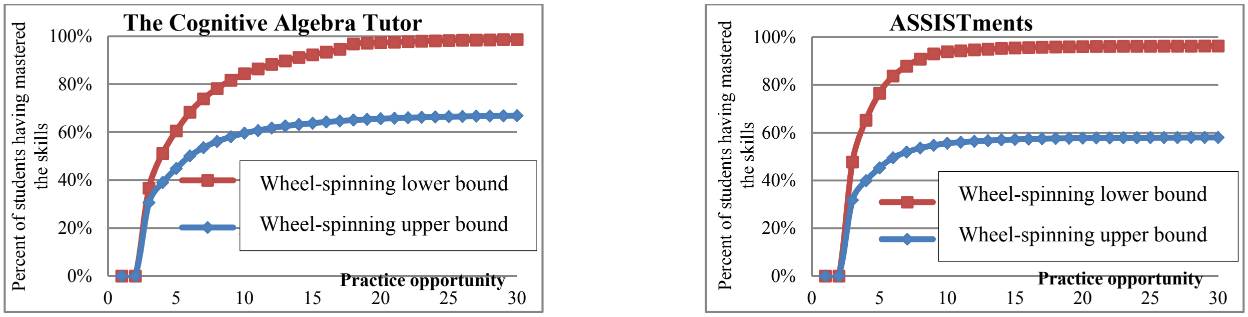
PRIOR WORK: THE SCOPE OF WHEEL-SPINNING

In our prior work [5], we investigated the scope of the wheel-spinning problem for two tutoring systems: the Cognitive Algebra Tutor (the CAT) and ASSISTments, a math tutor. To provide a background for this work, we briefly describe what we have found previously.

We defined mastery of a skill by a student correctly responding to three consecutive questions on the skill on his first attempt. This criterion is a lower bar for determining mastery compared to what was used in some system (e.g., [6]). We purposefully used a weaker mastery criterion as we did not want to artificially amplify the wheel-spinning problem by picking an overly strict criterion of mastery which is hard to achieve by students. We used real data to examine student progression to mastery. For the CAT, there are 146,479 problems done by 575 students for 111 algebra skills, forming 23,517 student-skill instances (i.e., a student working on a skill). For ASSISTments, there are 220,539 problems done by 5,997 students for 190 Math skills, forming 45,787 student-skill instances.

In Figure 1, the x-axis represents the number of problems and the y-axis represents the percent of student-skill instances that have demonstrated mastery. As we define wheel spinning as problem solving without making progress towards mastery, we set a limit of 15 practice attempts in the CAT and 10 practice attempts in ASSISTments before declaring wheel spinning. Similar to mastery, wheel-spinning exists in a context of students and skills, meaning it is only meaningful to say a student masters or wheel-spins on a skill. In addition to mastering a skill or wheel spinning, there is a third possible outcome:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions fromPermissions@acm.org.
L@S 2015, March 14–18, 2015, Vancouver, BC, Canada.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3411-2/15/03…$15.00
http://dx.doi.org/10.1145/2724660.2724673



(a) **(b)**

Figure 1 Wheel-spinning's lower bound and upper bound for the CAT (a) and ASSISTments (b).

what if a student who only practices a small number of problems on a skill? Consider a student using the CAT who solves 4 problems, but fails to master the skill – how should we categorize this student's performance? We call this third performance type "indeterminate," as we are unable to categorize it as either wheel spinning or mastery.

We leveraged indeterminate instances to estimate the scope of the wheel-spinning problem. We consider both an optimistic view that assumes all indeterminate instances would end with the student mastering the skill, and a pessimistic view that assumes that students would instead wheel spin in these instances. These two views establish an estimate of the scope of the wheel-spinning problem.

In Figure 1 (a) and (b), the top curves correspond to the optimistic view of mastery. It includes both direct observations of mastery, and imputed mastery from assuming all indeterminate cases would result in mastery. Since mastery and wheel-spinning are complementary, these curves also represent the lower bound of the wheel-spinning problem, which is the least severe case. Namely, that wheel spinning occurs approximately 16% of the time in the CAT and 6% of the time in ASSISTments. The bottom curves represent the pessimistic view of mastery, including observed wheel-spinning instances and imputing indeterminate instances as leading to wheel spinning. These curves represent the upper bound of the wheel-spinning problem, the worst case of wheel-spinning, which is about 40% in the CAT and 44% in ASSISTments. The difference between the upper and lower estimates of wheel spinning (16% to 40% and 6% to 44%) are large, and warrant further analysis to refine the scope the wheel spinning problem. At minimum, 16% and 6% rates of wheel spinning are substantial, and suggest wheel spinning should be investigated and addressed in computer tutors.

A GENERIC MODEL OF WHEEL-SPINNING

We wanted to understand whether wheel-spinning occurs randomly or it is associated with some factors that allow us to model it, and hopefully detect it at an early stage. Early detection would enable early intervention so as to not waste a learner's time. Our goal in this study is to build a generic model. Here, we are not over-claiming generalizability that one set of coefficient estimates can work for data from various systems. Rather, we want to find one set of features, which are computable based on data commonly stored by most ITSs, and by using this set of features, wheel-spinning can be modeled well in those ITSs.

We used the same criterion to determine mastery and wheel-spinning as in prior work. Three correct responses in consecutive questions of a skill is used as the mastery criterion. We chose 15 practice opportunities for the CAT and 10 for ASSISTments as the wheel-spinning determination checkpoints. A student could wheel-spin on one skill, while mastering another. Thus, the unit of declaring wheel-spinning is a student-skill pair.

Feature Engineering

Feature engineering is crucial in student modeling. Since learning and problem-solving are complex cognitive and affective processes, many student models succeed due to using extracted features (e.g. [4], [7]). We selected features to model wheel-spinning from three aspects: student in-tutor performance, the seriousness of the learner, and general factors.

Student In-tutor performance (on this skill)

Intuitively, whether a student is able to master a skill has much to do with how well the student understands the skill. We constructed five features to represent student in-tutor performance. These features are specific to the skill. Observations associated with other skills will not affect the calculations of feature values associated with this skill.

"Correct_Response_Count" - The prior number of problems solved correctly by the student on his first attempt. This feature is used in the Performance Factors Analysis model (PFA) [8], and found very useful in prediction [9].

"Correct_Response_In_A_Row_Count" - The current streak of correct responses on the first attempt. Since we define mastery as 3 correct responses in a row, the student's current streak is important. For example, if the student has got two problems right in a row, for the third problem, the value of this feature is 2. If his or her answer to the third problem happens to be incorrect, for the fourth problem, the value of this feature is reset to 0.

"Exp_Mean_Response_Time_Z-Score" - This feature is derived based on response times of prior questions of the same skill. Response time is the length of time spent by the student after being presented a problem and before taking the first action. Anderson et al. reported that according to ACT-R, fast speed of responding to a question often comes after high accuracy [10]. In practice, the winner of the KDD cup 2010 also applied this feature [7]. Instead of using an absolute value of time, we used exponential mean of the z-scores of response times as follows. For each problem, we collected all response times across all students to that problem. Then, we calculated the mean and standard deviation of the response times and used those to compute z-scores for all of the response times. Next, we aggregated all prior problems done by the student for the skill together and organized them in chronological order. On this series of z-scores, we used the following formula to calculate the exponential mean, γ * *prior_average* + *(1- γ)* * *new_observation*, with $\gamma = 0.7$. The exponential mean is a method of summarizing sequential data, but provides lesser weight to older observations, as prior observations are decayed by γ at each time step.

"Prior_Problem_Count_With_Hint_Request" - The number of prior problems on which the student requested a hint.

"Prior_Problem_Count_With_At_Least_5_Hint_Requests" - The number of prior problems of the skill for which the student requested 5 or more hints.

Many computer tutors provide multiple hints, with the final, or "bottom out" hint providing students the answer. Bottom-out hinting has a wide usage in student modeling as a predictor of student performance [11-13]. We were unable to compute from the provided log files whether CAT hints were bottom-out, so resorted to using two measures of hint usage as a proxy.

The Seriousness of the Learner (across skills)

Given the rich literatures on the connection between learning and the seriousness of the learner (e.g., [12-14]), we think of the seriousness of the learner as another type of factor possibly related to wheel-spinning. We found two aspects which seem reasonable to reflect the seriousness of the learner's attitude: the speed of responding to a problem and asking for hints in consecutive problems. Similar features capturing these two aspects were used in gaming detectors (e.g., [12-14]).

We designed 6 features reflecting response speed, and 2 features for consecutive hinting. Since these features are intended to represent the seriousness of the learner, they are calculated *across all* (in contrast to the student in-tutor performance features) skills, as we assume student attitude tends to be temporally stable across skills. The 6 features reflecting response speed are based on Z-score data described previously. This approach is appropriate as 25 seconds might be a fast response for a complex problem, but a slow response for a simple one. As we did not have access to all the details of the content presented, we relied on population statistics for that item to define relatively fast and slow responses. Fast responses were response times more than 1 standard deviation below the mean for that item. Slow responses were more than 1 standard deviation above the mean for that item, while normal responses were within 1 standard deviation of the mean for that item.

"Prior_Problem_Count_Fast_Correct" - the number of prior problems the student solved correctly and more than 1 standard deviation faster than the mean response time for that item.

"Prior_Problem_Count_Normal_Correct" - the number of prior problems the student solved correctly and within 1 standard deviation of the mean response time for that item.

"Prior_Problem_Count_Slow_Correct" - the number of prior problems the student solved correctly and more than 1 standard deviation more than the mean response time for that item.

"Prior_Problem_Count_Fast_Incorrect", "Prior_Problem_Count_Normal_Incorrect", "Prior_Problem_Count_Slow_Incorrect" – these three features are similar as the above three, yet for problems solved incorrectly.

We also computed 2 features based on hint usage. We used the same feature definition as we used for student in-tutor performance, with the tweak that rather than just looking at student performance on the skill being practiced, we look at performance across all skills. As a rationale, consider a student who asks for considerable help on just one skill. The likely explanation is the student is struggling with that skill. In contrast, a student who asks for considerable help on every skill is either confused about every skill, or is bored with the homework and is trying to get through the problems with minimal effort. The latter category is what we are trying to model with features capturing the seriousness of the learner.

"Prior_Problem_Count_With_Hint_Request_In_a_Row" - The number of consecutive prior problems across skills for each of which the student requested at least 1 hint.

"Prior_Problem_Count_With_At_Least_5_Hint_Requests_In_a_Row" - The number of consecutive prior problems across skills for each of which the student requested at least 5 hints.

General Features

We used two independent variables to track general information.

"Prior_Problem_Count" - The number of prior problems that have been practiced by the student for this skill. This feature is used in the Learning Factors Analysis model (LFA) [15]. We treated this feature as a factor as we did not see a strong reason to assume that there is a linear relationship between this variable and our target, wheel-spinning. Therefore, for each possible value (0 to 14 for the CAT and 0 to 9 for ASSISTments) our model estimates a unique value, indicating the effect to wheel-spinning given that many practices opportunities have passed.

"Skill_ID" - skill identification. This feature reflects skills may have different effects on wheel-spinning. It is treated as factor, and the parameter estimate for each skill could conceptually be interpreted as how difficult the skill is. Skill difficulty is a widely used feature in almost all types of student models. Research found that skill difficulty is very useful [16] in predicting student performance.

MODEL EVALUATION

Data

We applied three rules of data pre-processing. First, we removed indeterminate problems, as the wheel-spinning status is not certain and thus the value of the dependent variable is missing. Second, we removed student practice after reaching mastery on a skill, as in this case the status this student-skill pair was already determined. Third, for the similar reason, we removed all problems for a student after he reached the wheel-spinning cutoff (10 problems for ASSISTments and 15 for the CAT).

After data pre-processing, the CAT data set contains 96,919 problems done by 567 students. There are 76,684 mastery problems and 20,235 wheel-spinning problems. The ratio of the two classes is 3.8: 1. The ASSISTments data set contains 133,061 problems done by 5,103 students. There are 104,961 mastery problems and 28,100 wheel-spinning problems. The ratio of the two classes is 3.7: 1.

There are three types of predictions in student modeling [17]. In this paper, we examined type 1 prediction, how unknown students will perform on the observed problems. We conducted a three-fold cross-validation at the student level. As data were split at the student level, in each fold, students in the test data set have no intersection with those in the training data set. Therefore, to the model, these students serve as future, unseen, students, and the task for the model is to predict whether the student will master a problem or wheel-spin on it. All results are from averaging performance across all three folds.

Model Fitting

We trained separate logistic regression models to fit the CAT and ASSISTments data. The dependent variable is represented by a binary value, indicating mastery or wheel-spinning. We found that two sets of coefficient estimates are highly correlated to each other, suggesting that most features' effects on predicting wheel-spinning are consistent across the two systems. We do not include the estimates due to limited space. The interpretations below apply to the two models.

Among the features reflecting student in-tutor performance, "Correct_Response_In_A_Row_Count" is the most important feature and is negatively associated with wheel-spinning. The two features representing hinting are positively associated with wheel-spinning. Both of these results are in the expected direction.

In the features for the seriousness of the learner, in general, more prior incorrect responses is positively associated with wheel-spinning, which is worsened when combined with too fast responses. The two features representing "consecutive bottom-out hinting" are estimated as being positively associated with wheel-spinning.

In the general factors, we found that there is a clear, yet unexpected, linear trend that the more problems a student has attempted on a skill without mastery, the more likely he is to wheel-spin.

Model accuracy

We evaluated our wheel-spinning model on the training and test data sets for the CAT and ASSISTments. We present results in Table 1. We used three metrics, percent correct, AUC and R^2 to measure model accuracy.

Percent Correct is simply the percent of correct model predictions, interpreting model predictions above 0.5 as indicating wheel spinning and below 0.5 as indicating mastery. AUC is area under curve of the ROC. AUC measures how well the model is able to classify an instance of a binary category. The possible range of this metric is from 0.5 to 1. 0.5 means a random classification, and 1 means a perfect classification. Both Percent Correct and AUC show high performance on both the training and test sets. So we have confidence that our model generalizes to unseen students, and are able to interpret our parameter estimates as there is no evidence of overfitting.

Table 1 Model performance on the training and test data sets of the CAT and ASSISTments

	The CAT		**ASSISTments**	
	Training	Test	Training	Test
Percent Correct	86%	85%	87%	86%
AUC	0.89	0.88	0.89	0.88
R^2	0.40	0.38	0.42	0.40

The R^2 values are slightly more complicated, and are calculated based on the residuals between the actual value of the dependent variable, a binary categorical value of 0 or 1, and the predicted value of wheel-spinning, a decimal value ranging from 0 to 1. This metric, compared to the above two metrics, focuses less on classification ability, but more on the magnitude difference between the predicted values and true values. The R^2 value ranges between 0 and with 1 indicating a perfect fit. The measurements are around 0.4 for the training data and slightly lower for the test data. The values of this metric appear to be less satisfying than AUC at the first glance. However, given that the general performances of other major behavioral models, where the measurements of this metric is normally around

0.1 on test data [18, 19], our wheel-spinning model shows good performance, especially in its strong ability to generalize.

Model misclassifications

Typically, it is necessary to examine a classifier's misclassifications for each class of the dependent variable. This is particularly important for data where one class is of more interest than the others, or with imbalanced classes, one or more classes dominate the data set. In our context, we are particularly interested in investigating how the model performs for the wheel-spinning instances, as the ultimate goal is to address the question whether we can rely on the classifier to detect wheel-spinning and drive tutor interventions to prevent it. The two data sets also have imbalanced class distributions with many more instances of mastery than wheel spinning. The accuracy measure, which is used to compare classifiers, may not be well suited for evaluating models derived from imbalanced data sets. Therefore, we present the confusion matrices in Table 2 and Table 3 to help evaluate the models. The correct classifications are in the left top cell and the bottom right cell. The misclassifications are in the right top cell and the left bottom cell. We highlighted the numbers of misclassifications in italic. We also show the percent of each category in parentheses.

Table 2 Confusion matrix for test data on the CAT

		Predicted Category	
		Mastery	Wheel-spinning
Actual Category	Mastery	24,128 (74.7%)	*1,433 (4.4%)*
	Wheel-spinning	*3,172 (9.8%)*	3,572 (11.0%)

Table 3 Confusion matrix for test data on ASSISTments

		Predicted Category	
		Mastery	Wheel-spinning
Actual Category	Mastery	33,469 (75.5%)	*1,518 (3.4%)*
	Wheel-spinning	*4,390 (9.9 %)*	4976 (11.2%)

Table 2 and Table 3 have similar trends: more correct classifications on mastery than on wheel-spinning Most misclassifications occurred as false negatives (bottom left cell), while false positives (top right cell) is a minor problem. That is, the classifier is more likely to overlook a wheel spinning case than incorrectly assert wheel spinning when a student goes on to master a skill.

To focus on wheel-spinning, we used two additional metrics, precision and recall, which are useful when one of the classes is considered more important. Precision determines the fraction of instances that are actually wheel-spinning problems and the classifier has declared as a wheel-spinning problem. Recall measures the fraction of wheel-spinning instances correctly detected by the The precisions and recall results in Table 4 show moderate abilities of our models for the minority category: wheel-spinning. The precision rates are good, showing that over 70% declared wheel-spinning instances are accurately classified. The recall rates are lower in both models. The numbers are slightly above 50%, indicating that the model's ability to distinguish the wheel-spinning problems out of mastery problems is weaker. Thus, our model overlooks nearly half of the wheel spinning cases.

Table 4 Precision and recall on wheel-spinning on test data of the CAT and ASSISTments models

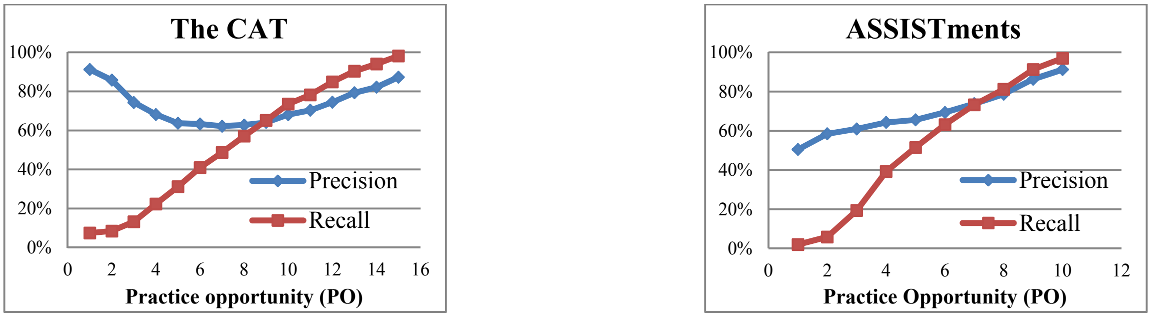
	Precision	**Recall**
The CAT	71.1%	52.7%
ASSISTments	76.6%	53.1%

Typically, there is a trade-off between precision and recall. We argue that compared to precision, a low recall rate is more acceptable. It is better to miss some students who will wheel spin than to incorrectly trigger interventions for students who will not struggle. This point is especially true as one of the first interventions we plan on attempting is informing the actual teacher that a student is likely to struggle on a particular skill. Given the limits on a teacher's time, allocating that effort efficiently is key.

Speed at detecting wheel-spinning

Unlike most behavioral detectors aimed at detecting a student's current behavior, a wheel-spinning detector is meant to detect something that happens over a longer period of time. Although wheel spinning is not flagged until problem 10 or 15, a student starts to struggle on a skill from the outset. Therefore, an important characteristic of this kind of detectors is its ability to detect the construct of interest quickly. We did fine-grained measures of detector performance at each practice opportunity, and plotted the evolutions in Figure 2. The x-axis is practice opportunity (PO). The y-axis represents both precision and recall.

It is particularly worth pointing out that due to the way we did data pre-processing, both the size and balance of the data change with PO. Removing excessive problems after mastery causes the data sets to shrink at each PO with fewer mastery instances left. For example, suppose we used 10 problems as the threshold to determine wheel-spinning. There are only two students, Ann and Mary, and one skill in the data set. Ann mastered the skill after practicing 4 problems. Mary practiced 12 problems in total without mastery, so she was wheel-spinning. We formed 10 sub data sets with each corresponding to a PO. In each of the sub data sets corresponding to PO1-PO4, there were 2 problems. One is from Ann, and the other is from Mary. The size and balance of the sub data sets of PO1-PO4 remain the same: the size is 2 and the ratio of mastery vs. wheel-spinning is 1:1.



(a) (b)

Figure 2 Precision and recall on wheel-spinning, broken down by practice opportunity for the CAT (a) and ASSISTments (b)

Starting from PO5, every sub data sets contains only 1 problem, which is done by Mary, as Ann has mastered the skill and no more problems of hers are in the original data set. The size of the sub data sets from PO5 to the last PO is 1, and the ratio of mastery vs. wheel-spinning is 0:1. Therefore, mastery becomes rarer over POs and wheel-spinning remains constant during the entire time.

Based on the quantitative relationship between mastery and wheel-spinning, we categorize POs into three chronological phases. Phase 1 [PO1-PO3] is the initiation phase in which the number of mastery student-skill pairs remains constant, so the ratio of mastery to wheel spinning is constant too. The ratio is quite high (10:1 for the CAT and 8:1 for ASSISTments). Phase 2 [PO4-PO10] for the CAT and [PO4-PO7] for ASSISTments, is the evolution phase in which the ratio of master to wheel spinning gradually decreases to roughly 1. Phase 3 is the termination phase, in which mastery becomes the minority class.

Figure 2 shows precision and recall for the CAT and ASSISTments on wheel-spinning by PO. The cold start problem is obvious and hurts recall badly. In the initiation phase, data balance at this time is poor as described above. Consequently, for PO1, the detectors have a near-zero recall rate. However, the detector is able to adapt quickly. At the end of the initiation phase, at PO3, when the detectors gathered information from only two prior problems, the recall rate of the CAT model increases to around 15%, and the recall rate of the ASSISTments model increases to 20%. Note that during this phase, neither data size nor balance changes; therefore the model should be credited for the improvements. In the following two phases, the recall rate rises gradually.

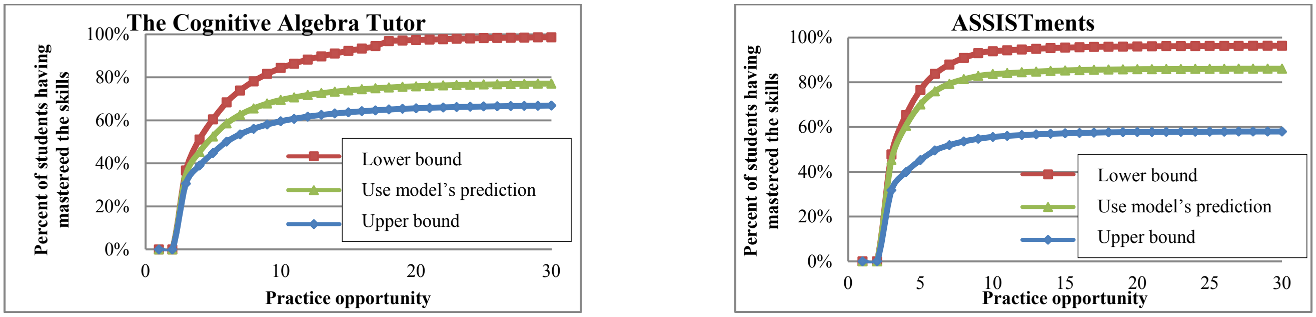
Precision looks very surprising to us. First, it is not hurt as badly as recall by the cold start problem. For ASSISTments, the initial measurement of precision is around 50%. This could be explained by the fact that the model is also informed by features across skills, and despite the first problem for the student-skill pair, the model has collected information from prior problems of other skills solved by the student.

More surprisingly, precision is around 90% at PO1 for the CAT model. It has a decreasing trend in the initiation phase and the first half of the evolution phase. We found this divergence is due to a dominant skill in the CAT through fine examinations.

Specifically, features representing student in-tutor performance have no effect at PO1, as their values at this point are 0. Then, we found no clear difference between the CAT and ASSISTments models in model estimates of the across-skills features. For the first general feature, Prior_Problem_Count, the parameter estimates of the two models at PO1 are both biases towards predicting mastery instead of wheel-spinning. Therefore, the only feature that could possibly be responsible for this divergence in the CAT vs. ASSISTments is “Skill_id”.

We found that in the ASSISTments data set, among all declared wheel-spinning problems at PO1, the skills are distributed randomly. On the other hand, in the CAT data set, at PO1, 79% declared wheel-spinning problems are associated with one skill. For some reason, many students failed that skill, resulting in wheel-spinning. The model captures this point, yielding an estimate for the skill representing a very low probability of mastering the skill. 79% declared wheel-spinning problems are associated with this skill, and they are correctly classified. This results in a high precision at PO1. The proportion of the declared wheel-spinning problems associated with this particular skill decreases over POs. With the lost of domination, precision drops after PO1 until PO7.

Both precision and recall continue to improve after the middle of the evolution phase. When the detectors enter the termination phase, at PO10 for the CAT and PO7 for ASSISTments, we see reliable performances. Both precision and recall are close to or above 70%. Note that at this point, the data sets have a good balance where two classes have roughly equal instances. There are still a couple of POs left before the wheel-spinning is certified, which means we can still accurately capture a good number of students and intervene to possibly prevent wheel-spinning.



(a) The CAT

(b) ASSISTments

Figure 3 The bounding and estimated curves (a) of the CAT (b) of ASSISTments

REFINING THE SCOPE OF WHEEL-SPINNING

In our prior work, we noted the difference in the upper- and lower-bounds warranted a closer investigation. We now utilize the wheel-spinning model we trained to estimate the true extent of the wheel-spinning problem. We elaborate the approach by using a concrete example. Suppose there are only two student-skill pairs in a data set, Ann-Addition and Mary-Subtraction. Ann did 5 problems for Addition and mastered the skill at the 5th PO. Mary did 5 problems for Subtraction and did not master the skill. We use 10 problems as the wheel-spinning checkpoint. Ann-Addition is a determinate instance, as mastery has been achieved. Mary-Subtraction is an indeterminate instance due to insufficient observations to determine wheel-spinning. The pessimistic upper bound on wheel spinning is computed by assuming Mary-Subtraction remains unmastered. The optimistic lower bound on wheel spinning is computed by assuming Mary-Subtraction is mastered. We now discuss how we use our model to estimate the true amount of wheel spinning within this envelope.

To impute indeterminate cases, we used our model's prediction for the last problem in the student-skill sequence. Using the above example, we would use the model's prediction for Mary-Subtraction after observing the fifth completed problem. We use the last datapoint since the model's accuracy will be highest. Suppose the prediction of the last problem of Mary-Subtraction is 0.3, meaning a 30% chance to demonstrate mastery. We would therefore split this case between mastered and unmastered, allocating 0.3 units towards the number of mastered cases and 0.7 units towards the wheel spinning cases. By fractionally allocating cases as mastery or wheel spinning according to our model's predictions, we can get a more accurate estimate of the rate of wheel spinning.

Figure 3 demonstrates how students make progress towards mastery as a function of PO. These graphs are identical to Figure 1, except the indeterminate cases are allocated according to our model's predictions as represented by the middle curve in both graphs. In Figure 3 (a), we see that the middle curve is closer to the bottom curve, the upper bound of the wheel-spinning problem. At PO 15, the wheel-spinning cutoff, the model's estimate of wheel spinning reaches about 75%. Thus, for the CAT, the wheel-spinning problem is estimated to affect approximately 25% of student-skill pairs. In Figure 3 (b), the wheel-spinning problem seems less severe in ASSISTments, as at PO10 (the wheel-spinning checkpoint) the estimate climbs up to around 82%, leaving 18% of the student-skill pairs wheel-spinning.

There are two points that we want to make in particular. First, 25% and 18% of student-skill pairs is a large number of lessons from which the learner gains nothing, but wasting learning time, and it could lead to frustration. Second, we speculate that wheel-spinning is underestimated by our model. Indeterminate student-skill instances have relatively few problems, which end at early POs. Take the ASSISTments model as example, at PO 3 in Figure 2, precision is around 60% and recall is around 20%; therefore, the true number of wheel-spinning problems is 3*n*, three times of the problems having been declared wheel-spinning. This systematic undercount of wheel spinning biases the model in favor of predicting mastery.

CONTRIBUTIONS, FUTURE WORK AND CONCLUSIONS

We made several contributions in this work. First, we deepened the understanding of the wheel-spinning problem by finding that it is not random. Instead, we showed that a set of features, computable from typically stored data by most ITSs, have consistent and robust associations with wheel-spinning across different learning environments and student populations. In student modeling, models often depend on detailed information specific to the system [4, 12, 13], which may expose the work to the issue of "most published research findings are false" due to barriers to replicate a study [20]. Building a generic model is important from a scientific point of view, as it makes replicating our results easier in other tutoring systems.

Second, we built a behavioral detector for wheel-spinning. We applied machine leaning techniques to generate the wheel-spinning model, which saves lots of human labor as needed in human-generated models, nor do we need human observers as in [21]. Unlike most behavioral detectors aiming to detect the student's state over a brief temporal interval [4, 12, 13], the wheel-spinning detector is designed to detect what happens over a long time period (10 - 15 problems away in the future). Furthermore, this work sheds light on research of behavioral detectors that require both high accuracy and fast speed of detection. Our detector has robust performances across different tutoring systems with over precision and recall.

Finally, we refined the scope of the wheel-spinning problem, which is important for understanding this generally-occurred problem, and showed it affected a substantial number of students using the CAT and ASSISTments. In addition, by using the model to estimate the amount of wheel-spinning, we presented an approach to evaluate the effectiveness of a tutoring system.

One interesting future work is to refine/design a wheel-spinning model by other machine learning techniques, especially for addressing the low recall problem. The most important open question is what we can do about wheel-spinning? The CAT and ASSISTments are effective tutoring systems and both have been developed to help students reach mastery. Nevertheless, wheel-spinning still hurts students in both systems. We are eager to know what interventions might be effective in reducing wheel spinning.

To sum up, in this work, we built a model with a set of generic features for wheel-spinning. We showed for the CAT and ASSISTments, the model can generalize very well to unseen students with satisfying accuracy and detection speed. Finally, we used model predictions to refine the scope of wheel-spinning, and showed that it affects a substantial proportion of students for both systems.

REFERENCES

1. Anderson, J.R., et al., Cognitive tutors: Lessons learned. The Journal of the Learning Sciences, 1995. 4: 167-207.
2. Wikipedia. Learning-by-doing (economics). http://en.wikipedia.org/wiki/Learning-by-doing_(economics).
3. Cen, H., K. Koedinger, and B. Junker. Is More Practice Necessary? - Improving Learning Efficiency with the Cognitive Tutor through Educational Data Mining. in International Conference on Artificial Intelligence in Education. 2007.
4. Baker, R.S., S. Gowda, and A. Corbett, Automatically Detecting a Students Preparation for Future Learning: Help Use is Key, in the 4th International Conference on Educational Data Mining, M. Pechenizkiy, et al., Editors. 2011. p. 179-188.
5. Gong, Y. and J.E. Beck, Wheel-spinning: An important problem with the mastery learning framework International Journal of Artificial Intelligence in Education, (in preparation).
6. Farrell, R.G., J.R. Anderson, and B.J. Reiser. An Interactive Computer-Based Tutor for LISP. in Fourth National Conference on Artificial Intelligence. 1984.
7. Yu, H.-F., et al., Feature engineering and classifier ensemble for KDD cup 2010. KDD Cup 2010: Improving Cognitive Models with Educational Data Mining, 2010.
8. Pavlik, P.I., H. Cen, and K. Koedinger, Performance Factors Analysis - A New Alternative to Knowledge Tracing, in the 14th International Conference on Artificial Intelligence in Education. 2009. p. 531-538.
9. Gong, Y. and J.E. Beck, Items, Skills, and Transfer Models: Which Really Matters for Student Modeling? , in the 4th International Conference on Educational Data Mining. 2011. p. 81-90.
10. Anderson, J.R., Rules of the mind. 1993, Hillsdale, NJ: Lawrence Erlbaum Associates.
11. Feng, M., et al. Addressing the assessment challenge in an Intelligent Tutoring System that tutors as it assesses. User Modeling and User-Adapted Interaction, 2009. 19: p. 243-266.
12. Gong, Y., et al., The impact of gaming (?) on learning at the fine-grained level. The 10th International Conf. on Intelligent Tutoring Systems. 2010. 194-203.
13. Baker, R.S.J.d., et al., Developing a Generalizable Detector of When Students Game the System. User Modeling and User-Adapted Interaction., 2008.
14. Arroyo, I. and B. Woolf, Inferring learning and attitudes from a Bayesian Network of log file data in the 12th International Conference on Artificial Intelligence in Education. 2005: Amsterdam.
15. Cen, H., K. Koedinger, and B. Junker. Learning Factors Analysis - A General Method for Cognitive Model Evaluation and Improvement. in Intelligent Tutoring Systems. 2006. Jhongli, Taiwan: Springer.
16. Gong, Y. and J.E. Beck, Looking Beyond Tranfer Models: Finding Other Sources of Power for Studnet Models, in The 19th International Conference on User Modeling, Adaptation and Personalization. 2011.
17. Chi, M., et al., Instructional Factors Analysis: A Cognitive Model For Multiple Instructional Interventions, in the 4th International Conference on Educational Data Mining, M. Pechenizkiy, et al., Editors. 2011. p. 61-70.
18. Gong, Y., J.E. Beck, and N.T. Heffernan, How to Construct More Accurate Student Models: Comparing and Optimizing Knowledge Tracing and Performance Factors Analysis. International Journal of Artificial Intelligence in Education, 2010.
19. Baker, R.S.J.d., et al., Ensembling Predictions of Student Knowledge within Intelligent Tutoring Systems, in The 19th International Conference on User Modeling, Adaptation, and Personalization. 2011. p. 13-24.
20. Ioannidis, J.P.A., Why Most Published Research Findings Are False. PLoS Med 2005. 2(8:e124).
21. Liu, Z., et al., Sequences of Frustration and Confusion, and Learning, in the 6th International Conference on Educational Data Mining, S. D'Mello, R. Calvo, and A. Olney, Editors. 2013: Memphis, TN. p. 114-120.