How Long Must We Spin Our Wheels? Analysis of Student Time and Classifier Inaccuracy

Yue Gong
Facebook
Facebook 1 Hacker Way
Menlo Park, CA, 94025
yuegong@fb.com

Yan Wang, Joseph Beck
Worcester Polytechnic Institute
Worcester, MA 01609
{ywang14, josephbeck} @wpi.edu

This section has been submitted to:

Gong, Y., Wang, Y., & Beck, J. (2016). How Long Must We Spin Our Wheels? Analysis of Student Time and Classifier Inaccuracy. *The 9th International Conference on Educational Data Mining.* ACM.

ABSTRACT

Wheel-spinning is the phenomenon where students, in spite of repeated practice, make no progress towards mastering a skill. Prior research has shown that a considerable number of students can get stuck in the mastery learning cycle--unable to master the skill despite the affordances of the educational software. In such situations, the tutor's promise of "infinite practice" via mastery learning becomes more a curse than a blessing. Prior research on wheel spinning overlooks two aspects: how much time is spent wheel spinning and the problem of imbalanced data. This work provides an estimate of the amount of time students spend wheel spinning. A first-cut approximation is that 24% of student time in the ASSISTments system is spent wheel spinning. However, the data used to train the wheel spinning model were imbalanced, resulting in a bias in the model's predictions causing it to undercount wheel spinning. We identify this misprediction as an issue for model extrapolation as a general issue within EDM, provide an algebraic workaround to modify the detector's predictions to better accord to reality, and show that students spend approximately 28% of their time wheel spinning in ASSISTments.

Keywords

Wheel-spinning; Precision; Recall; Intelligent Tutoring Systems

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 repeated problem solving of what is initially declarative knowledge, causing it to compile into production rules for a procedural representation [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 them 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 and there is the illusion of progress, but it is not going anywhere. Similarly, the tutor is presenting students with many problems to solve and there is the appearance of productive work, but the students are not making progress towards mastery.

Prior work [5] introduced the concept of wheel spinning, which describes the phenomena that students can not master a skill in a timely manner. Using data from two ITS called the Cognitive Algebra Tutor [13] and ASSISTments [14], they analyzed the severity of wheel-spinning, and build a logistic model to predict students wheel spinning. In general, the model provided good prediction accuracy with an AUC of 0.88 [6]. However, since the model was trained based on imbalanced data (most students master a skill rather than wheel spinning), the model has high false negative rate, which means wheel spinning cases are relatively more likely to be mispredicted as mastery cases. Therefore, when we apply this model to indeterminate cases (which we can not label wheel spinning or mastery based on the given data), the estimated rate of wheel spinning is likely an undercount. This paper addresses the undercount, and further estimates how much time students spend wheel spinning.

DATA SET

In this paper, we used the similar data set used in [6] from ASSISTments. ASSISTments is a web-based computer tutor, primarily used for middle-school math education (approximate ages 12 to 15). This data set contains information from 5997 students chosen at random, who used ASSISTments during the time period of September 2010 to July 2011. The students completed a total of 208,328 math problems during this time period. These students were primarily from the northeast United States. We have student self-reported ages, and 75% of the students asserted they were 12 to 15 years of age on January 1, 2011. Since the students spread across a wide range of grades, they solved problems including a large range of skills as well. The problems cover 190 math skills, such as Equation-Solving-More-

Than-Two-Steps, Area-Irregular-Figure, etc. Since we have access to the ASSISTments system's database, we can reach fine-grained information, such as every action the student made while he was solving the problem. This allows us to analyze the relationship between wheel-spinning and non-productive "learning" behaviors induced by these fine-grained data.

This work retains the initial definition of wheel spinning [5] of failing to master a skill within 10 practice opportunities. We define mastery as getting three problems correct in a row. This threshold of mastery is rather low, and so these results are a lower bound on wheel spinning. Some students practiced fewer than 10 problems without reaching mastery. It is not obvious whether these students would master the skill or not, and we categorize them as "indeterminate." Table 1 shows the number of student-skill pairs in each category.

Note that a student could wheel spin on adding fractions but master multiplying decimals. Therefore, we speak of wheel spinning or mastering a particular skill by a student. Thus, when characterizing the amount of wheel spinning, our analysis is in terms of student-skill pairs.

Table 4. Breakdown of student performance by mastery type

Category	Mastery	Indeterminate	Wheel-spin
Number of student-skill pairs	25449 (55.6%)	17528 (38.3%)	2810 (6.1%)

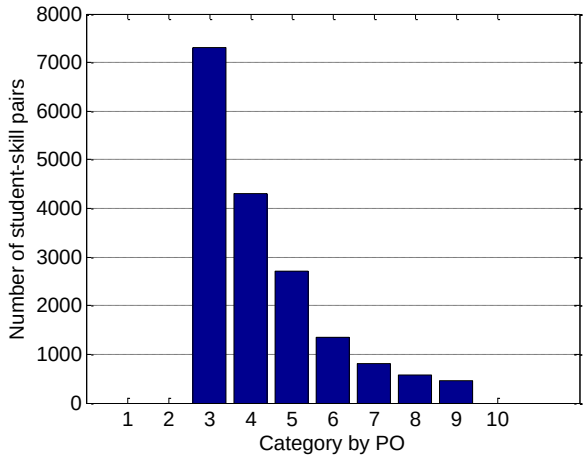

Figure 3. Number of indeterminate student-skill pairs at each PO

Since wheel spinning is trivial to predict for cases where we can observe either wheel spinning or mastery, we are more interested in the distribution of indeterminate cases. Figure 1 shows frequencies of student-skill pairs at a certain number of practice opportunities of indeterminate cases. Clearly, the larger the PO is, the fewer observations we have. Students in the indeterminate group tend to have fewer PO; the majority of students did no more than 5 problems. It is interesting that students seem to give up relatively rapidly on a problem set.

Overall, there is a large imbalance of more mastery cases than wheel spinning cases. However, this imbalance interacts with the number of practice opportunities (PO) a student has had on a skill, as shown in Figure 2. The number of student-skill pairs considered wheel-spinning does not change with PO, since by definition a student must reach PO 10 in order to be categorized as wheel spinning. The reason the number of wheel spinning cases is constant is that when we observe a sequence as either wheel spinning or mastery, we label all PO in the sequences with that label. Since 10 PO are required for wheel spinning, all 10 bins have the same quantity. However, students can master a skill after 3 PO. Therefore, the number of student-skill pairs still working towards mastery decreases rapidly as PO increases.

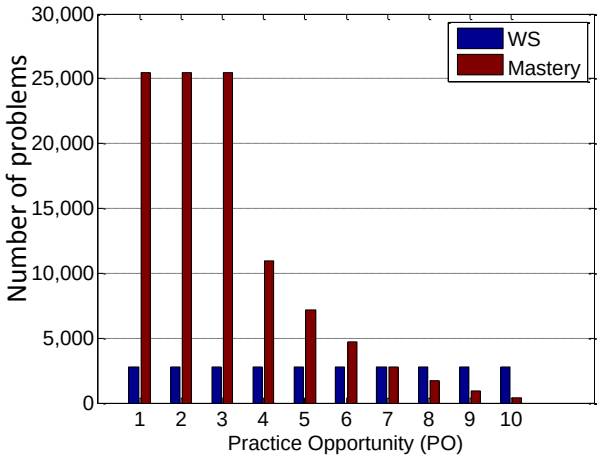

Figure 4. Number of wheel spinning and mastery problems at each PO

REVISIT THE WHEEL SPINNING PREDICTIVE MODEL

Model Performance Metrics

In this paper, we reused the model provided in [7]. The model is trained based on determinate cases (mastery and wheel-spinning cases), and then it is applied to indeterminate cases to make predictions and estimate the rate of wheel spinning. The model was trained using three fold cross validation. This model has strong performance statistics on the test set of unseen students: R2 of 0.4 and AUC of 0.88. However, its precision and recall are reasonable but less strong: 0.76 and 0.53, respectively. We now develop an argument to show as a consequence of the precision and recall statistics, the predictive model undercounts the amount of wheel spinning on the indeterminate cases.

Evaluation of the Model with Precision and Recall

In a classification model, the precision of a model, P, is the number of true positives, TP, divided by the total number of cases predicted as positive, PP. A model's recall, R, is the number of true positives divided by the total number of cases that are actually positive, +. As a consequence, we have the formulas

$$\mathbf{P} = \mathbf{TP} / \mathbf{PP} \quad (1)$$

$$\mathbf{R} = \mathbf{TP} / + \quad (2)$$

A model's precision is how selective it is. When it predicts the category will occur, how often is it right? Recall measures how comprehensive a classifier is. Of the actual cases, how many can it detect? Clearly, there is trade off between precision and recall. A classifier could be very cautious and only make a positive prediction when it was very certain, resulting in a high precision but low recall. Conversely, a classifier could categorize everything as an instance of the category, achieving perfect recall

but (presumably) low precision. The precision and recall of wheel spinning and mastery are shown in Table 2.

Table 5. Precision and recall for Mastery and wheel spinning

Category	Mastery	Wheel spinning
Precision	88.3%	75.6%
Recall	95.3%	52.5%

This model has a high precision and recall for predicting mastery. However, the precision and recall of wheel spinning is relatively low. Wheel spinning's precision of 75.6% means that about one out of four of the cases that is predicted as wheel spinning is actually mastery. Recall of 52.5% means that the model can only capture successfully about half of the WS cases.

The low recall of WS is not surprising if we look at the distribution of data set shown in Table 1. Mastery cases occupy a large portion. Under such a circumstance, it is understandable that the model tends to predict cases as mastery to reduce the prediction error—the goal of the model fitting process.

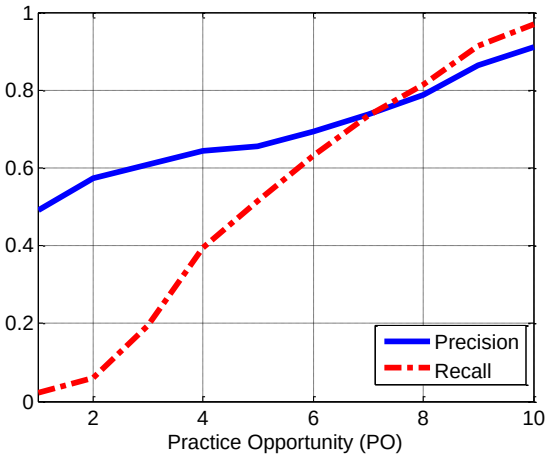

Figure 5. Precision and recall for Wheel Spinning prediction

More specifically, we analyzed precision and recall at different POs. Figure 3 and Figure 4 show the precision and recall of wheel-spinning and mastery of the wheel spinning prediction model, both disaggregated by PO. Interestingly, precision and recall both improve for problems in the wheel spinning category as the model observes the student making more practice opportunities on the skill. This explanation makes intuitive sense: as the model acquires more data, it is better able to detect when a student will wheel spin. Interestingly, precision and recall of the Mastery category both decrease with additional observations of the student performing the skill. At first, this situation seems paradoxical, until one consider the distribution of Mastery vs. Wheel Spinning in Figure 2. Initially, Mastery is the majority class. Its relative advantage begins to slip after PO 3, and by PO 7 it has achieved numerical parity with Wheel Spinning. After PO 7, Wheel Spinning is the majority class. As Mastery becomes less and less dominant in the data set, its predictive accuracy decreases.

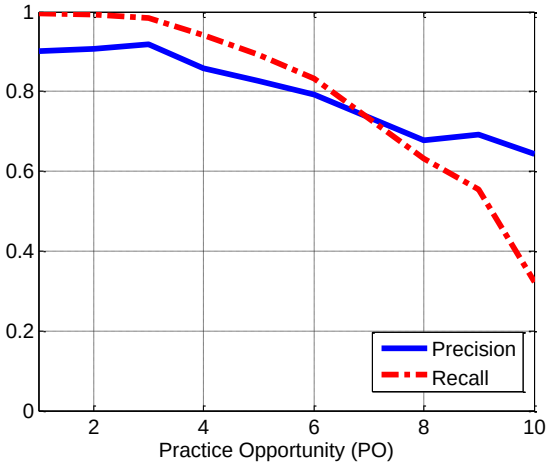

Figure 6. Precision and recall for Mastery prediction

Implications of imbalances in classifier accuracy

Consider the relationship between the precision and recall and the number of true positives. From a standpoint of precision, the number predictions made multiplied by the precision is equal to the number of correct predictions. That is:

P * **PP** = **TP** (3)

Conversely, we can define the number of true positives using recall. Specifically, the number of actual occurrences of a category, multiplied by the model's recall, provides the number correct predictions of that category. That is:

R * + = **TP** (4)

Since equations 3 and 4 both have the number of true positives on their right-hand side, we can set them equal to each other:

R * + = **P** * **PP** (5)

Dividing both sides by R and rearranging we get:

+ = **PP** * (**P** / **R**) (6)

In other words, the number of positive examples in a data set is equal to the number of predicted positives, multiplied by the precision over recall. A few points of discussion. First, it may seem conceptually odd to need to compute the number of positive examples in a data set, as it is normally countable directly from the data. However, for our problem we have a large number of indeterminate cases where we are unable to observe what their true label would be, and we need to infer it. More broadly, applying behavioral classifiers outside of the labeled training data encounters this same problem: how many instances are there really in the data set? Such a situation would arise when attempting to apply a model trained on one system to a second system. The second observation is that the (P/R) term in Equation 6 can be thought of as a normalizing constant for reweighting the data. The number of instances predicted to be positive is adjusted by P/R. Sometimes this adjustment will increase the number of instances and other times it will decrease the number of instances. In either case, *this adjusted number of instances is a better estimate of the number of positive examples in the data than the number of predicted positives from the classifier*.

An intuitive way to reweight the prediction results is to directly use the precision and recall ratios shown in **Table 5** to compute the P/R ratio. However, we have additional information in that

we know the relative counts of Wheel Spinning and Mastery change dramatically with PO. Therefore, rather than applying a global reweighting term of 0.756/0.525 for Wheel Spinning and 0.883/0.953 for Mastery, we instead create more fine-grained reweightings based on PO.

Figure 5 shows the P/R ratio for both categories broken down by PO. Note that for a low number of PO, the P/R ratio for wheel spinning is noticeably higher than 1. In other words, early on in the sequence many wheel spinning cases are miscategorized as Mastery by the classifier, and there is a systematic undercount in the number of Wheel Spinning students. In contrast, the Mastery category has a P/R ratio of approximately 1.0 throughout its range, only rising noticeably above 1.0 on PO 9 and 10. Thus, Mastery cases are undercounted late in the sequence of problem solving.

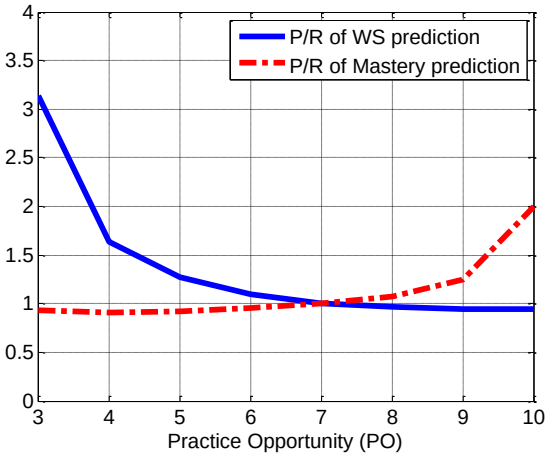

Figure 7. Ratio of precision/recall for Wheel Spinning and Mastery

REANALYSIS OF PREDICTION RESULTS FOR COMPUTING AMOUNT OF WHEEL SPINNING

We now turn our attention to first reestimating past results using the reweighted data. Then we focus on estimating the time spent wheel spinning using both the straightforward approach of using the classifier results as-is (i.e., the PP value) vs. using the reweighted PP * (P/R) value.

Estimating amount of mastery

By applying the predictive model to indeterminate cases, we can get predicted category of these cases. Since the ratio of precision and recall is not very large for Mastery prediction, the modification of those predictions is generally a small decrement. However, for Wheel Spinning predictions, the P/R ratio is generally higher than 1, causing an increase in the number of predicted cases of Wheel Spinning.

Figure 6 shows the number of indeterminate student-skill pairs predicted to result in Mastery. For each PO, the bar on the left represents the number of cases that will result in Mastery originally predicted by the model. The bar on the right for each PO represents the adjusted count by reweighting each student-skill pair by its corresponding P/R ratio. For problems at PO 3, the P/R ratio for Wheel Spinning predictions was over 3, so those cases are weighted 3 times as heavily. For Mastery problems, the P/R ratio was just under 1.0, so those counts are relatively unchanged. As a result of this reweighting, there is a noticeable drop in the estimated number of indeterminate students who will master the skill after 3 PO.

For PO3 through PO6, the reweighting is pessimistic and causes more student-skill pairs to be categorized as Wheel Spinning than the model predicts on its own. At PO 7, both categories have a P/R ratio of approximately 1.0, so the counts are (roughly) unchanged. For PO 8 and 9, since the P/R ratio of Mastery is larger than for Wheel Spinning, we see an increase in the expected number of students who Master the skill relative to the model's predictions.

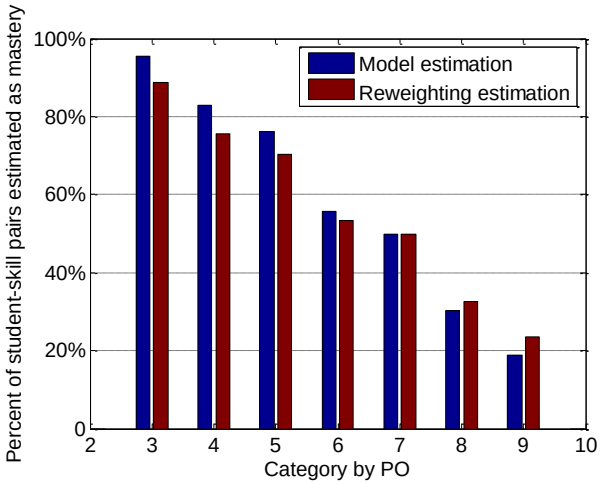

Figure 8. Original and reweighted proportions of student-skill pairs predicted as resulting in mastery

We now compute the cumulative percent of students who will master a skill, by assigning the indeterminate student-skill pairs to either Mastery or Wheel Spinning. Figure 7 shows the result of this process. The upper and lower lines are optimistic and pessimistic assumptions of student performance, and provide an absolute upper- and lower-bound on the percentage of student-skill pairs that will result in mastery. The upper-bound on mastery assumes all indeterminate students will master the skill. The lower-bound assumes all students will wheel spin. Our goal is to better estimate mastery within that range of possible values. The solid green line in the middle of the graph is the result of applying the model's predictions to the indeterminate data points (identical to the analysis in [6]). The dashed red line represents using the same model predictions, but reweighting them according to the P/R ratio provided in Figure 5. For example, if an indeterminate case was predicted as resulting in Wheel Spinning, we would count that as approximately 1.6 observations of Wheel Spinning, as that is the P/R ratio for that category for that number of practice opportunities. Overall, there is not a large change in the expected proportion of students-skill pairs reaching mastery. There is a slight decrease of 2% absolute in the expected amount of mastery, with about 16% (shown in Figure 7) of student-skill pairs expected to exhibit Wheel Spinning.

As another illustration of the impact of weighting the model's output, Table 3 shows the impact on the number of indeterminate cases counted as mastery or as wheel spinning. Note that no student-skill pair actually receives a different prediction as a result of the modification, the counts in the table change strictly as a result of reweighting the counts by P/R. Although we are able to obtain more accurate counts, we are not able to more accurately predict any individual case as Wheel Spinning or Mastery. Note

that the percentage of mastery in Table 3 (75%) differs from Figure 7 (84%) since Figure 7 refers to wheel spinning, mastery, and indeterminate cases, while Table 3 zooms in and considers only the indeterminate cases.

Table 6. Estimated number (percent) of indeterminate student-skill pairs predicted as each category

Category	Mastery	WS
Estimation	14028 (80%)	3500 (20%)
Modified Estimation	13086 (75%)	4442 (25%)

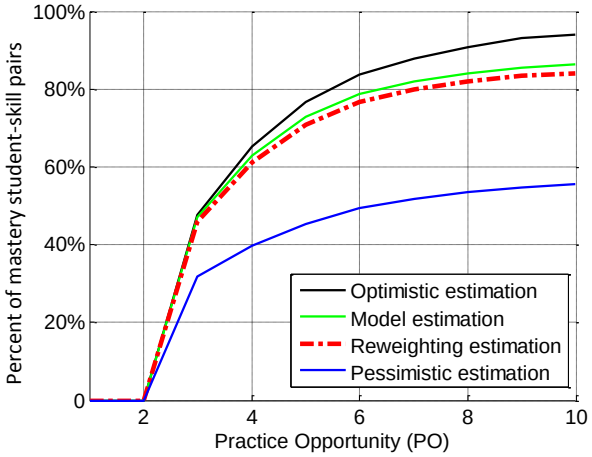

Figure 9. Cumulative percent of student mastering the skill by PO. Model estimate and reweighted estimate.

Estimating time spent wheel spinning

Our final analysis is to estimate the amount of time students spend in the wheel spinning state. First, we examined how long students spent solving a problem. Figure 8 shows the average number of seconds students spent on a problem, broken down by category (observed mastery, observed wheel spinning, or indeterminate), and plotted by PO. Several trends are evident. First, problems solved in skills where the student will wheel spin take approximately 25% longer to solve than problems solved in skills that the student eventually masters. The other observation is that there is a sharp drop in time to solve a problem from PO 1 to PO 2, presumably due to memory effects as students swap into working memory [7] the necessary procedures for solving problems of this type. After PO2, there is a slight decreasing trend in time spent per problem across indeterminate, mastery, and wheel spinning student-skill pairs. This interaction of time and PO illustrates the importance of using a P/R ratio conditioned by PO, as shown in Figure 5, as early values of PO, where the P/R ratio is greatest, take the greatest amount of time to solve.

The other thing to note is that wheel spinning students spend much longer on skills than students who master. First, wheel spinning students spend more time per problem (Figure 8). Second, wheel spinning students attempt many more problems on a skill than students who master it. Observed wheel spinning requires 10 observations. So we should expect the time spent wheel spinning to be substantially higher than the 16%, which the percent of student-skill pairs observed to exhibit wheel spinning.

To compute time spent wheel spinning, we treated student-skill pairs that resulted in either wheel spinning or mastery as time spent in the respective state. For indeterminate sequences, we compute the probability of Wheel Spinning according to the model for the last problem in the sequence. Presumably the final PO has the most information, and provides the best estimate of whether the student will wheel spin or not. We then use the P/R reweighting term for the final PO to reweight time spent in all of the problems for this student-skill pair. This approach maximizes information used in making the prediction, and uses the P/R ratio that is associated with that model's prediction. So if a student reaches PO 6 and is predicted to wheel spin, we use a ratio of approximately 1.1 (from Figure 5) to reweight the time spent in all 6 POs, and do not artificially inflate the time by using the P/R ratio from PO 1 through 5 for this student-skill pair.

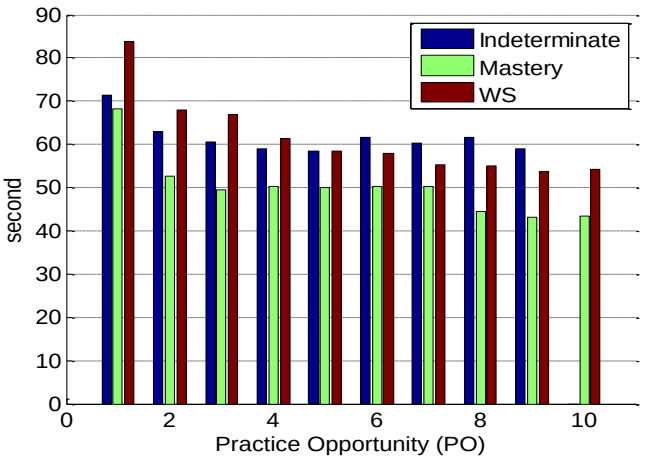

Figure 10. Average time spent on a problem

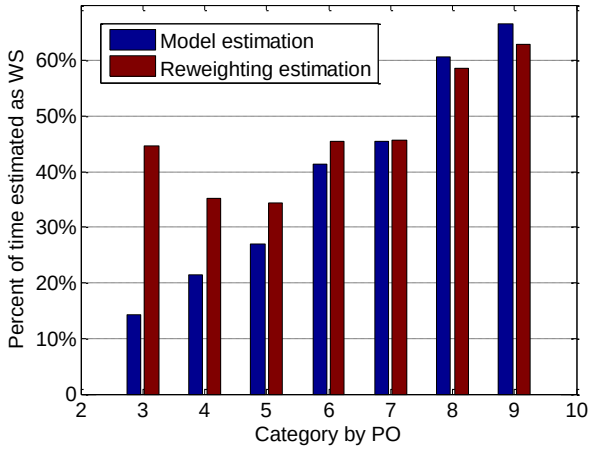

Figure 11. Estimated time spent wheel spinning for indeterminate cases

Figure 9 provides the amount of time students spend wheel spinning, broken down by PO. At PO 3, the reweighting results in a sharp increase in the amount of time estimated as spent wheel spinning. As the P/R ratio becomes closer to 1, the reweighted counts and model predictions become more similar to each other.

Table 7. Estimated time (in number of hours and as a percentage) spent Mastering and Wheel Spinning

Category	Optimistic	Model estimated	Reweighting data	Pessimistic
Mastery	2239	2035	1921	1481

	(84%)	(76%)	(72%)	(56%)
WS	422 (16%)	626 (24%)	740 (28%)	1181 (44%)

After reweighting the predicted amount of time spent Wheel Spinning or Mastering for each student-skill pair, we computed the total amount of time spent wheel spinning. Table 4 shows the time spent wheel spinning and mastering for our data set. Using the model's predictions as-is, we get that students spent 626 hours wheel spinning, or 24% of their time. Reweighting the data results in that amount increasing to 740 hours, or 28% of their time. Finding that over 600 hours of student time was wasted over a year is not a comforting thought. Using the reweighted estimate, over one-quarter of student time is spent in the wheel spinning state. This value is not a small number, and should be a focus of attention for improving the tutor.

CONTRIBUTIONS

This paper makes contributions to understanding wheel spinning and more broadly to the field of educational data mining. Within the context of wheel spinning, this paper extends prior work on estimating the amount of wheel spinning [6]. Given the prevalence and breadth of wheel spinning, approximately 26% in the Cognitive Algebra Tutor, 16% student-skill pairs in ASSISTments, and over one-third in a study of the cognitive tutor on a non-WEIRD population [8], efforts to better understand wheel spinning can have a broader impact than on other constructs commonly studied which are typically observed on many fewer students. Prior research [5, 6, 7] examined the total number of student-skill pairs that exhibit wheel spinning. Such analysis is informative, but neglects to consider the amount of time student spend spinning their wheels in the mastery learning cycle. The amount of time is particularly relevant given that problems where students are wheel spinning take somewhat longer to complete. Furthermore, students perform more problems in wheel spinning sequences than in sequences that end in mastery. Consequently, students in ASSISTments wheel spin on 16% of problem sequences, but spend 28% of their time in the wheel spinning state. The 28% would be even worse, except that some students who are likely to wheel spin stop doing the tutor's exercises and give up on the problem set. Realizing that much student time is being wasted by a commonly used computer tutor is surprising, and such analysis of time is rarely done, with a few exceptions [8].

The second contribution this paper makes is refining the understanding of a classifier for wheel spinning, and by extension, other classifiers used in educational data mining. The precision, recall, and AUC of the previously published predictive model of wheel spinning are quite good. However, looking at the performance in detail indicates there are systematic biases in its predictions, which should lead us to be cautious in interpreting its results.

The final contribution of this paper is in an interesting approach of correcting for imbalanced data in a classifier. The classifier is doing a good job for its role: minimize its prediction error, possibly extended with an asymmetric loss function to penalize certain types of mistakes more heavily. The classifier's job is not to make the most accurate extrapolation at a coarse grain size by correctly estimating the total number of times a certain behavior occurs. As a result, when a classifier is used to extrapolate to a new dataset and estimate the rate of occurrence of a phenomenon, there is a mismatch between that mission and its goal. As a simple example, for a problem with 99% positive examples, a very accurate classifier would categorize all examples as positive. It would not, however, be useful for extrapolating population statistics as it would claim that 100% of the data were positive examples when we know that is not true. Although we know the classifier is overpredicting the majority class, we are not sure *which specific instances* are being overcounted.

This work provides a means for reweighting the data to cause the classifier to better-align its predictions with known counts in the data. We are able to perform this reweighting by taking advantage of the relationship between precision, recall, and the known base rates. In addition, we leverage the strong relation between practice opportunity and precision/recall. Consequently, we are able to make better predictions about collections of data points, and better allocate student time between wheel spinning and mastery states. However, this algebraic trick does not allow to modify our prediction about any specific student-skill pair and increase the classification accuracy of the detector. This apparent conundrum, and separation of the roles of behavioral models into predictions of individuals and categorizing large numbers of trials is a contribution to the field of educational data mining[1]: simply extrapolating model predictions can lead to erroneous claims about the amount of a behavior or the time spent in that behavior. In fairness, the change for this study was moderate in scope: the amount of time spent wheel spinning is approximately 28% of total time rather than 24%. However, for detectors with weaker performance metrics, this difference could be much larger.

FUTURE WORK AND CONCLUSIONS

The most obvious line of future work is the creation of a stronger classifier for wheel spinning, as well as for other detectors of learner behavior and affect. The wheel spinning detector has strong performance metrics (on test-set data): AUC of 0.88, R2 of 0.4, precision of 0.76 and recall of 0.53 [6]. In spite of those solid metrics, there is a notable problem with extrapolation due to the skew between precision and recall. A naïve approach would be to simply alter the loss function [10] to balance precision and recall. However, this approach would reduce the predictive accuracy of the model, its sine qua non. Also, some algorithms in AI domain also provides possible solutions [11, 12], but those approaches modify the classifier's predictions, so there is a loss in accuracy of predictions. On the other hand, semi-supervised learning is also a technique we would like to try in the future. [15]

The second area is to analyze whether student characters that influence wheel-spinning between determinate cases and indeterminate cases are similar. In this paper, we assume that the model built on determinate cases also applies to indeterminate cases. However, whether this assumption holds should be validated. In the future, more data (previous information) about students in both determinate cases and indeterminate cases should be gathered, and analyzed for comparison of similarity between the two groups.

[1] We suspect we are not the first to reweight our data in this manner, but none of us are experts in information retrieval. The second author of the paper developed the idea independently while thinking about the classifier's performance metrics, and the first author developed an explanation for this paper and did a quick literature search to no avail. We would appreciate any pointers to the literature of making use of this approach to enable a model to better extrapolate.

The third area of research is to reduce the amount of time spent wheel spinning. Wheel spinning consumes a large amount of student time, typically in a block spent working on a particular topic. Beyond being ineffective for learning, it is presumably disengaging for learners as well. The problem is that most obvious interventions have been tried, as ITS designers attempt to construct systems from which students can learn. Analysis of how much wheel spinning could be reduced by ensuring students understood their prerequisite skills reveals a modest decrease [10]. Thus, there is a need for effective strategies for reducing wheel spinning. One possible strategy is a strong detector capable of quickly detecting that a student is likely to wheel spin, and simply stop providing her/him problems on the topic. This creation of an escape mechanism from the mastery learning cycle would reduce time spent wheel spinning, and couple with instruction by a human teacher or tutor, could possibly be an effective intervention.

The fourth area of future work is further thinking about the different uses of predictive models. This work examines two: predicting individual cases and extrapolating the model to an aggregate group, and identifies an issue with undercounting the minority class for analyzing the impact of a behavior. Are there other crucial differences between these two uses beyond the one noted in this paper? Is there a third type of use of models that has different properties entirely?

In conclusion, this paper extends what is known about wheel spinning. We have found that students spend approximately 28% of their time in a wheel spinning state. More interesting is how we calculated this number: reweighting the data to modify the impact of the model's predictions. Thus, this paper not only extends our understanding of the common and detrimental behavior of wheel spinning, but improves our methodological sophistication for understanding behavioral detectors.

ACKNOWLEDGMENTS

We acknowledge funding from NSF (# 1440753, 1316736, 1252297, 1109483, 1031398, 0742503), ONR's 'STEM Grand Challenges' and IES (# R305A120125 & R305C100024) grant for ASSISTments and support of the author.

REFERENCES

[1] Anderson, J.R., et al. 1995. Cognitive tutors: Lessons learned. The Journal of the Learning Sciences (April, 1995), 167-207.

[2] Wikipedia. Learning-by-doing (economics). Available from: http://en.wikipedia.org/wiki/Learning-by-doing_(economics) (March, 2016).

[3] Cen, H., Koedinger, K. and Junker. B. 2007. Is More Practice Necessary? - Improving Learning Efficiency with the Cognitive Tutor through Educational Data Mining. Frontiers in Artificial Intelligence and Applications, 158 (2007), 511.

[4] Baker, R.S., Gowda, S. and Corbett, A. 2011. Automatically Detecting a Students Preparation for Future Learning: Help Use is Key. In Proceedings of the International Conference on Educational Data Mining (Eindhoven, the Netherlands, July 06 - 08, 2011). EDM'11. ACM, New York, NY, 179-188.

[5] Gong, Y. and Beck, J. 2015. Towards Detecting Wheel-Spinning: Future Failure in Mastery Learning, in Proceedings of Conference on Learning @ Scale (Vancouver, Canada, March 14 - 18, 2015). L@S'15. ACM. New York, NY, 67 – 74.

[6] Beck, J. and M.M.T. 2014. Rodrigo, Understanding Wheel Spinning in the Context of Affective Factors, in Intelligent Tutoring Systems (2014), 162-167.

[7] Beck, J.E. and Gong, Y. 2013. Wheel-Spinning: Students Who Fail to Master a Skill, in Proceedings of International Conference on Artificial Intelligence in Education (Memphis, USA, July 09 – 13, 2013). AIED'13. 431-440.

[8] Mostow, J., et al. 2002. A la recherche du temps perdu, or as time goes by: Where does the time go in a Reading Tutor that listens? In Proceedings of the International Conference on Intelligent Tutoring Systems (Biarritz, France, 2002). ITS'2002. 383 – 390.

[9] Mitchell, T. 1997. Machine Learning. McGraw-Hill. 432.

[10] Wan, H. and Beck, J. B. 2015. Considering the influence of prerequisite performance on wheel spinning. In Proceedings of the International Conference on Educational Data Mining (Madrid, Spain, July 26 - 29, 2015). EDM'15. ACM, New York, NY.

[11] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16 (2002), 321-357.

[12] Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. 2003. SMOTEBoost: Improving prediction of the minority class in boosting. Knowledge Discovery in Databases: PKDD (Sep, 2003), 107-119.

[13] Koedinger, K. R., & Corbett, A. T. 2006. Cognitive tutors: Technology bringing learning science to the classroom. The Cambridge Handbook of the Learning Sciences. Cambridge University Press, Cambridge, MA.

[14] Heffernan, N. T., & Heffernan, C. L. 2014. The ASSISTments Ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. International Journal of Artificial Intelligence in Education, 24 (Dec, 2014), 470-497.

[15] Zhu, X., & Goldberg, A. B. 2009. Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning, 3 (Jun, 2009), 1-13.