Square it up! How to model step duration when predicting student performance

Irene-Angelica Chounta
University of Tartu
Estonia
chounta@ut.ee

Paulo F. Carvalho
Carnegie Mellon University
USA
pcarvalh@cs.cmu.edu

ABSTRACT

In this paper, we explore how we can model students' response times to predict student performance in Intelligent Tutoring Systems. Related research suggests that response time can provide information with respect to correctness. However, time is not consistently used when modeling students' performance. Here, we build on previous work that indicated that the relationship between response time and student performance is non-linear. Based on this concept, we compare three models: a standard Additive Factors Analysis Model (AFM), an AFM model enhanced with a linear step duration parameter and an AFM model enhanced with a quadratic, step duration parameter. The results of this comparison show that the AFM model that is enhanced with the quadratic step duration parameter outperforms the other models over four different datasets and for most of the metrics we used to evaluate the models in cross validation and prediction.

CCS CONCEPTS

• **Applied computing~**Interactive learning environments • **Applied computing**~Computer-assisted instruction

KEYWORDS

student modeling, step duration, intelligent tutoring systems

ACM Reference format:

1 INTRODUCTION

The use of students' response time for modeling and predicting their performance in Intelligent Tutoring Systems (ITSs) has been, so far, a challenging task and an open question. For instance, even though response time can potentially be a good predictor of post-test scores, it does not always predict performance in individual learning steps [13].

Given these challenges, it is still not clear how to identify and model a "good" response time – a response time that indicates the student is knowledgeable about the task – or a "bad" response time – a response time that indicates that the student either is not interested in the activity or does not have the required background knowledge to address it. Being able to identify a "good" versus a "bad" response time is important because it allows us to provide help and feedback in a timely manner – when it is really necessary and needed – and consequently we can create better systems and gain a better understanding of how learning takes place.

In this paper, we propose a new modeling approach for predicting student performance using the student's response time. In particular, we build on the hypothesis that there is no linear relationship between student response time and correctness: a student who takes either too little time or too long to respond to a step (where a step can be either a tutor's question or task), will most likely be unsuccessful for this particular step. Therefore, we argue that modeling a student's response time as a quadratic factor - rather that a linear one – will result in more accurate and better performing student models [7].

Prior studies suggest that indeed the relationship between response time and student performance is non-linear [3,11]. On the one hand, a student needs a minimum amount of time in order to process the problem, retrieve appropriate information, and to construct a correct response. If the student attempts to respond too fast, this can mean that either they did not really process the task as required or that the student attempts to game the system. On the other hand, if the student takes too long to respond, this may indicate lack of background knowledge, failure to retrieve critical information, and inability to address the step.

Here, we compare three student models: a traditional cognitive student model, the Additive Factors Model (AFM) that predicts student performance with respect to prior student practice, an AFM-LT model that predicts student performance with respect to prior student practice *and* step duration – that is, the time a student takes to carry out a step of a learning activity – as a *linear function*, and an AFM-QT model that predicts student performance with respect to prior student practice *and* step duration as a *quadratic function*. We argue that the AFM-QT model will outperform both the AFM and the AFM-LT model with respect to the goodness of fit as well as with respect to prediction accuracy of student performance on unseen steps from seen students.

In the next sections, we provide an overview of the related work and describe the three student models used in this work. Then, we present the dataset used to train and test the student models and the method of study. Finally, we present the results and conclude with a discussion on the findings and future work.

2 STUDENT MODELING

2.1 Background about student models

Most student models developed for intelligent Tutoring Systems (ITSs) are based on the notion of mastery learning; that is, the student is asked to continue solving problems or answering questions on a concept until she has mastered it. Only then will the student be guided to move forward to other concepts [9,10]. Mastery learning is in line with the notion of learning curves that is, how many opportunities a student needs in order to master a skill or knowledge component. In order to assess mastery, ITSs use student models that predict the performance of students on various steps of a learning activity. Based on these predictions, the tutor chooses what kind of content or scaffolding to provide to students.

Typically, cognitive student models predict step outcomes – that is whether a student will carry out a step-task correctly or not – based on the skills involved in this step and student’s prior practice. For example, the Additive Factors Analysis Model (AFM) - introduced into ITS research by Cen et. al. [4,5] - predicts the likelihood of a student correctly completing a step as a linear function of student parameters (the student’s proficiency), knowledge components or skill parameters (the difficulty of the knowledge components or skill involved in certain questions or tasks) and the learning rates of skills. AFM considers the frequency of prior practice and exposure to skills. In addition to AFM, the Performance Factors Analysis Model (PFM) [16] considers whether prior practice was successful (that is, how many times a student answered correctly or incorrectly) and the Instructional Factors Analysis Model (IFM) [6] also considers the tells (that is, how many times the tutor gave away the answer of the next step directly instead of eliciting).

Even though there have been attempts to introduce time as a predictor for student’s performance (see e.g., [14]), to the best of our knowledge, as of now there is no student model that uses step duration to predict student success with consistent results used in ITSs.

2.2 Time and student performance

Previous work has focused on studying the relationship between time and outcome in terms of correctness, either directly or as a proxy of engagement, but also as a predictor for student performance. Xiong and Pardos [18] explored the use of response times for predicting future performance but they noted that they did not identify a clear trend between response time and correctness. Lin et. al. [13] incorporated response times in BKT models and explored whether this addition would lead to better performance in terms of prediction accuracy in next-step’s performance. This work indicated that response time can potentially be a good predictor for posttest scores but does not always support predicting performance on individual steps. Beck [2] proposed the use of response time as a proxy for students’ engagement and argued that setting time thresholds to students is counterproductive because this practice does not take into account students’ characteristics.

Other work suggested that there is a differentiation in response times with respect to the outcome’s correctness. In particular, Miller et al. [15] found that *on average*, the response time for correct answers was faster than for incorrect answers and that good background knowledge and high self-efficacy relates to fast response time.

Chounta and Carvalho [7] suggested that the relationship between step duration and student performance is non-linear and therefore it would not be accurate to model time as a linear factor. To conceptualize this, they defined the concept of the Zone of Interest, as “*a time frame defined by a minimum and a maximum step duration (dtmin and dtmax respectively) in which a student will likely provide the correct answer (and thus the error rate will be low in the same interval). Consequently, every step that lies outside this time frame will most likely be solved incorrectly or not solved, and the error rate will be high”.* Their results suggested that within the ZOI students tended to give more correct answers than outside the ZOI which indicated a non-linear relationship between time and performance. This concept is supported by related research that indicates that too little and too much time on task are indicators of unsuccessful practice and students’ attempts to game the system [3,11].

3 METHODOLOGY

3.1 Three student models: AFM, AFM-LT, AFM-QT

In this paper, we compared three nested student models. In particular, we used a standard AFM model as a basis. As aforementioned, the AFM model predicts student’s performance on a step that involves a skill KC based on the student’s prior practice regarding this skill. For the implementation of the AFM model, we followed Datashop’s proposed approach[1] shown in the regression formula (1):

(1) AFM = Outcome ~ Student + KC + KC:Opportunity

where:

Outcome is the result per step – correct or incorrect;

Student stands for the student id of the student who carries out this step;

KC is the skill involved in this step;

KC:Opportunity stands for the number of previous attempts a student had on this particular skill.

[1] https://pslcdatashop.web.cmu.edu/help?page=rSoftware

In order to include students' response time, we enhanced the standard AFM by adding step duration as a linear component to the original AFM model. This is depicted in the regression formula (2). We refer to this time-enhanced model as the AFM-LT (AFM-Linear Time) model.

(2) AFM-LT = Outcome ~ Student + KC + KC:Opportunity + + step_duration

where:

Outcome is the result per step – correct or incorrect;

Student stands for the student id of the student who carries out this step;

KC is the skill involved in this step;

KC:Opportunity is the number of previous attempts a student had on this particular skill.

step_duration is the time the student took to carry out this step (in seconds).

Next, we use another time-enhanced model - we refer to it as AFM-QT (AFM-Quadratic Time) model. Here, we enhanced the standard AFM by adding step duration as a quadratic component to the original AFM model step duration. This is depicted in regression formula (3).

(3) AFM-QT = Outcome ~ Student + KC + KC:Opportunity + + step_duration + $(\text{step\_duration})^2$

We argue that since the relationship between step duration and student performance is not linear, the quadratic model AFM-QT will be a better fit and provide more accurate predictions than both the AFM and the AFM-LT models.

3.2 Datasets

To further study our hypothesis, we used four datasets from four different STEM-related courses. All of the courses were supported by Intelligent Tutoring Systems. All datasets are shared via the online repository *Datashop* [12]. From these datasets we excluded steps that were not related to any skill (KC) and we also excluded steps that did not have any information about their duration. We used the following datasets:

- Fractions (domain: Math)[1]: the tasks of this dataset are focusing on identifying and constructing fractions using graphical representations. The data were collected in 2013. After preprocessing, dataset consisted of 77 students, 8003 steps and 20 KCs.
- Genetics [8]: the tasks of this dataset were complex problem-solving activities across a wide range of genetics topics. The data were collected in 2016. After initial preprocessing, it consisted of 124 students, 14458 steps and 53 KCs.
- Stoichiometry (domain: Chemistry) [14]: the tasks in this dataset are problem-solving activities and worked examples on Stoichiometry and the data were collected in academic year 2009-2010. After preprocessing, the dataset consisted of 55 students, 9564 steps and 36 KCs.
- Physics [17]: problem-solving tasks for Physics. Data were collected in academic year 2011-2012. After preprocessing, the dataset consisted of 314 students, 38499 steps and 102 KCs.

3.3 Study setup

In order to explore how different ways to model time would impact prediction of student performance, we compared the performance of the three student models – AFM, AFM-LT and AFM-QT – on the four datasets acquired from the aforementioned ITS-supported science courses.

For the comparison of the student models, we followed a two-step process:

As a first step, we fitted the three models on the whole dataset and we conducted a 10-fold cross validation. As measures of the models' quality and performance, we used 3 metrics that are commonly used for model selection: the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the Cross-Validation estimate of Accuracy (CV.ACC). For AIC and BIC, the lower the value the better the model. For CV accuracy, a higher value signifies a more accurate model. These metrics have been used in related work for choosing between parametric models with different numbers of parameters [4,14].

As a second step, we randomly split the dataset using the Pareto principle: 80% of the dataset was used to train our models and 20% of the dataset was used for testing – that is, for prediction. The training-testing process was repeated for all the four courses. The aim was to study how well the models would predict unseen steps. To evaluate the results of the second step, we used as predictive accuracy metrics the Root Mean Square Error (RMSE) - depicted in equation (4). RMSE has been used in related work to evaluate predictive accuracy [6].

$$(4) \quad RMSE = \sqrt{\frac{\sum_1^N (p_i - o_i)^2}{N}}$$

4 RESULTS

To explore our research hypothesis, we studied whether the use of step duration as a predicting feature improves the performance of a well-established student model. Then, we studied whether modeling step duration as a quadratic function would provide us with a better fitting model than modeling step duration as a linear function. We compared the three student models that we described in section 3.1: an AFM student model, an AFM with a Linear Step Duration feature and an AFM with a Quadratic Step Duration feature. As a first step, we trained the three models on each and every-one of the four courses' datasets and we carried out 10-fold cross-validation. Then, we compared the three models using the following metrics: AIC, BIC and cross validation estimate of accuracy (for 10-fold cross validation). For both AIC and BIC, lower values indicate models that are closer to the truth and consequently, more accurate fits. On the contrary, higher cross-validation estimates of accuracy indicate better fits.

4.1 Model comparison - AFM, AFM-LT & AFM-QT

The results of the first-step of the analysis are presented in Table 1. The results show that the AMF-QT models - that is the AFM enhanced with the step duration quadratic factor - are better fits over all the four datasets. In all cases, the AFM-QT models exhibit lower AIC and BIC than both the AFM and the AFM-LT models. In turn, the AFM-LT models - that is the AFM models enhanced with the linear step duration factor - have lower AIC and BICs from the traditional AFM models.

Table 1. Comparison between the three models AFM, AFM-LT and AFM-QT using AIC, BIC and cross-validation accuracy CV ACC

Course: Fractions			
	AIC	BIC	CV ACC
AFM	8483.4	8839.8	0.732
AFM-LT	7416.7	7780	0.784
AFM-QT	7270.8	7641.1	0.778

Course: Genetics			
	AIC	BIC	CV ACC
AFM	15385.55	17121.14	0.737
AFM-LT	14762.96	16506.13	0.756
AFM-QT	14486.7	16237.45	0.762

Course: Chemistry			
	AIC	BIC	CV ACC
AFM	11124.45	12027.34	0.770
AFM-LT	10579.93	11489.99	0.769
AFM-QT	10531.49	11448.7	0.769

Course: Physics			
	AIC	BIC	CV ACC
AFM	24985.21	29409.89	0.876
AFM-LT	24385.14	28818.39	0.877
AFM-QT	23863.51	28305.31	0.878

The results were similar regarding the cross-validation estimate of accuracy (for a 10-fold cross validation). The accuracy of the AFM-QT models in the cross validation was higher in comparison to the AFM models for three out of four datasets. For the chemistry dataset, the cross-validation accuracy was almost the same for all three models. In comparison to the AFM-LT models, the AFM-QT have either similar or higher accuracy.

We confirmed the results regarding the goodness of fit using the ANOVA Chi-square (likelihood ratio) test. For all datasets, the AFM-QT models significantly outperformed both the AFMs and the AFM-LT models ($p < .0001$). This finding suggests that, on the one hand, adding time - in this context as the step duration- as a predictive feature can improve the predictive power of a student model. On the other hand, the relation between step duration and performance seems to be non-linear. In our case, it was suggested that using a quadratic function for modeling students' response times can be a more appropriate way than using a linear function.

4.2 Models' comparison for unseen step prediction - AFM, AFM-LT & AFM-QT

Next, we compared the three models with respect to unseen step prediction. As described in section 3.3, we used the each of the three models to predict the outcome of unseen steps with respect to correctness. To evaluate the accuracy of prediction, we used the Root Mean Square Error (RMSE). The results are presented in Table 2.

The RMSE was lower for the AFM-QT models than the traditional AFM models for all four cases. Furthermore, the AFM-LT model was also more accurate than the AFM models. When comparing the AFM-QT to the AFM-LT models, the results suggest that the AFM-QT model is better than the AFM-LT model for all the datasets as well.

In one case, for the Physics course, the RMSE difference between the three models – and especially between the AFM-LT and the AFM-QT – is very small suggesting that the predictive accuracy is almost similar. This is also depicted in the cross-validation results. One possible explanation might be that the accuracy of the model is negatively affected by the big number of KCs. The dataset coming from the Physics course has twice as many (and even more) KCs than the other courses.

Table 2. Comparison between the three models AFM, AFM-LT and AFM-QT using RMSE for predicting performance on unseen steps

	RMSE (prediction)		
Dataset	AFM	AFM-LT	AFM-QT
Fractions	0.404	0.385	0.378
Genetics	0.435	0.428	0.421
Stoichiometry	0.446	0.437	0.433
Physics	0.303	0.299	0.298

5 CONCLUSIONS

In this paper, we investigated the use of step duration as a predictor of student performance in two ways: modeling step duration as a linear parameter (AFM-LT) and modeling step duration as a quadratic parameter (AFM-QT). To explore the effect of these two approaches, we compared the two models with a standard cognitive model (AFM) over four different STEM, ITS-supported courses.

We determined that including step duration as a quadratic parameter improves the model's performance both with respect to goodness of fit and with respect to predictive accuracy on unseen steps. The AFM-QT model outperformed the standard AFM and the AFM-LT models in all cases over all the performance metrics, except one. In this one case, the AFM-LT model performed slightly better with respect to cross-validation accuracy (AFM-LT = 0.784, AFM-QT = 0.778) but the AFM-QT performed better with respect to all other metrics. The results also showed that the student models tend to perform similarly when the number of KCs increases. This may indicate that increasing the number of KCs affects negatively the accuracy of the model in general.

The contribution of this approach is two-fold: first, the results suggest that in this way we can use step duration towards improving the performance of student models; second, it offers insight with respect to the relationship between response time and student performance. The results of this work have implications for designing intelligent tutoring systems, for providing timely feedback, and potentially for designing personalized learning and assessment activities. On the one hand, designing student models that take into account students' response times can support us in providing accurate predictions of student performance. On the other hand, being able to relate student performance to response times will allow us to provide focused and timely feedback. For example, using the outcome of the student model we can advise a student to use more time in order to think carefully a hasty answer or to provide a hint to a student who takes too long to carry out a step.

For future work, we plan to use this approach in combination with the Performance Factors Analysis Model (PFM). We envision this is an important step because PFM differentiates between correct and incorrect steps and thus we can model step duration separately for correct and incorrect outcomes. Furthermore, we plan to explore how this approach may impact student learning outcomes.

ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation grants #ACI-1443068 and BCS-1824257.

REFERENCES

1. Vincent Aleven. 2013. Classroom study 2013, Dataset 669 in Datashop. *Find it at: https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=669*.
2. Joseph E. Beck. 2004. Using response times to model student disengagement. In *Proceedings of the ITS2004 Workshop on Social and Emotional Intelligence in Learning Environments*, 20.
3. Paulo F. Carvalho, Min Gao, Benjamin A. Motz, and Kenneth R. Koedinger. 2018. Analyzing the relative learning benefits of completing required activities and optional readings in online courses. *Methods* 34: 68.
4. Hao Cen, Kenneth Koedinger, and Brian Junker. 2006. Learning factors analysis–a general method for cognitive model evaluation and improvement. In *International Conference on Intelligent Tutoring Systems*, 164–175.
5. Hao Cen, Kenneth Koedinger, and Brian Junker. 2008. Comparing two IRT models for conjunctive skills. In *International Conference on Intelligent Tutoring Systems*, 796–798.
6. Min Chi, Kenneth R. Koedinger, Geoffrey J. Gordon, Pamela Jordon, and Kurt VanLahn. 2011. Instructional factors analysis: A cognitive model for multiple instructional interventions.
7. Irene Angelica Chounta and Paulo Carvalho. 2018. Will time tell? Exploring the relationship between step duration and student performance. In *Rethinking Learning in the Digital Age. Makeing the Learning Sciences Count*, 993–996.
8. Albert T. Corbett. 2016. REAL CCHS Genetics Study Data 2016. Dataset 1566 in Datashop. *Find it at: https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=1566*.
9. Albert T. Corbett and John R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4: 253–278.
10. Albert T. Corbett, Kenneth R. Koedinger, and John R. Anderson. 1997. Intelligent tutoring systems. *Handbook of human-computer interaction* 5: 849–874.
11. David B. Daniel and John Broida. 2004. Using web-based quizzing to improve exam performance: Lessons learned. *Teaching of Psychology* 31, 3: 207–208.
12. Kenneth R. Koedinger, Ryan SJd Baker, K. Cunningham, A. Skogsholm, B. Leber, and John Stamper. 2010. A Data Repository for the EDM community: The PSLC DataShop. In *Handbook of Educational Data Mining*.
13. Chen Lin, Shitian Shen, and Min Chi. 2016. Incorporating Student Response Time and Tutor Instructional Interventions into Student Modeling. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, 157–161.
14. Bruce M McLaren. 2009. Pittsburgh_Science_of_Learning_Center_Stoichiometry_Study_5. Dataset 268 in Datashop. *Find it at:https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=268*.
15. Kelly Miller, Nathaniel Lasry, Brian Lukoff, Julie Schell, and Eric Mazur. 2014. Conceptual question response times in peer instruction classrooms. *Physical Review Special Topics-Physics Education Research* 10, 2: 020113.
16. Philip I. Pavlik Jr, Hao Cen, and Kenneth R. Koedinger. 2009. Performance Factors Analysis–A New Alternative to Knowledge Tracing. *Online Submission*.
17. Kurt VanLehn. 2011. OSU, Honors Physics: Mechanics, Fall 2011. Dataset 577 in Datashop. *Find it at: https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=577*.
18. XIAOLU Xiong and Zachary A. Pardos. 2011. An analysis of response time data for improving student performance prediction.