Lee, Y.-J. (2015). Analyzing Log Files to Predict Students' Problem Solving Performance in a Computer-Based Physics Tutor. *Educational Technology & Society, 18* (2), 225–236.

Analyzing Log Files to Predict Students' Problem Solving Performance in a Computer-Based Physics Tutor

Young-Jin Lee

Educational Technology Program, University of Kansas, 1122 West Campus Road, Room 413, Lawrence KS, 66045, USA // yjlee@ku.edu

ABSTRACT

This study investigates whether information saved in the log files of a computer-based tutor can be used to predict the problem solving performance of students. The log files of a computer-based physics tutoring environment called Andes Physics Tutor was analyzed to build a logistic regression model that predicted success and failure of students' problem solving from their past interactions with the computer-based tutor. The logistic regression model developed in this study was able to correctly identify about 70% of the observed problem solving performance. The 10-fold cross-validation and the Receiver Operating Characteristic (ROC) curve analyses suggest that the developed logistic regression model can predict students' problem solving performance on unseen new problems with a similar accuracy in the future.

Keywords

Log file analysis, Computer-based learning environment, Problem solving, Statistical modeling

Introduction

With the rapid development of computer and telecommunication technologies, computer-based learning environments, such as Massive Open Online Courses (MOOCs), are gaining popularity especially in higher education (Allen & Seaman, 2010). One of the benefits of computer-based learning environments is that they provide instructors with a new avenue for understanding the learning processes of their students; since students leave hidden trails in the log files recorded in the computer-based learning environment, instructors can quietly collect information relevant to the learning activities of their students (e.g., how much time students spend on a particular learning content, when they work during the day/week, etc.) without interrupting their learning processes.

As computer-based learning environments providing log files become more prevalent, Educational Data Mining (EDM) is emerging as a new, exciting research field. One of the active research areas in EDM is estimating academic performance of students who are using computer-based learning contents. For example, Hwang and Wang (2004) reported that the academic performance of college students taking a technology course was positively correlated with diligence and intensity of their learning in an e-learning environment. Lykourentzou, Giannoukos, Nikolopoulos, Mpardis, and Loumos (2009) proposed a data mining technique that can identify at-risk students who are likely to fail or drop an online course. They used a variety of data mining techniques, such as Feed-Forward Neural Networks, Support Vector Machines (SVM) and probabilistic ensemble simplified fuzzy ARTMAP, to build their prediction model from log files of an e-learning environment. Palazzo, Lee, Warnakulassoriya, and Pritchard (2010) developed a computational method that can detect students who submitted copied answers to their electronic homework problems by analyzing the log files saved in a Web-based tutoring system. They found that college freshmen who copied more than 30% of electronic homework problems scored up to 2 standard deviations lower in the final exam than their non-copier classmates who had comparable academic skills measured at the beginning of the semester (Palazzo, Lee, Warnakulassoriya, & Pritchard, 2010). Macfadyen and Dawson (2010) were able to build a logistic regression model that correctly identified 81% of students who received a failing grade by examining log files of their learning management system that documented learning activities of their students, such as the number of email messages sent, discussion messages posted and assessment modules completed by students.

This study seeks to extend the body of knowledge in EDM by focusing on *more fine-grained learning activities* occurring in a computerized learning environment, and how such information can be used to predict the learning outcomes of students for *a longer period of time*. While previous studies focused on whether or not students visited a particular Web page, this study examines more specific learning activities taking place in a computer-based learning environment (e.g., whether students were able to solve relevant problems with or without using instructional supports available in the computer-based learning environment). Moreover, while previous studies focused on predicting one academic performance measured at the end of the semester (e.g., final exam score or course failure), this study tries to predict students' academic performance throughout the semester (e.g., problem solving performance in weekly homework assignments). The objective of this study is to build a statistical model that can estimate the probability for students to successfully accomplish a problem solving step required to resolve a difficult physics problem without using hints available in the computer-based tutor, given their interaction history recorded in the log files of the learning environment (e.g., whether students were able to solve related problems, when they solved those problems, how well they did and how much time they spent in solving those problems, etc.).

Method

Andes physics tutor

In this study, the log files from Andes Physics Tutor (http://andestutor.org) were analyzed to build a statistical model that predicted how likely a student is to correctly solve a physics problem from his or her past interactions with a computer-based tutor. Andes Physics Tutor is an intelligent tutor program capable of providing appropriate instructional supports and guidance while students are solving difficult physics problems requiring an analytic solution. Unlike many computer-based instructional programs, which typically provide a single corrective feedback or hint on the *final answer* even though students have to go through multiple steps before arriving at their final answer, Andes Physics Tutor provides appropriate feedback and allows students to ask for hints on *each problem solving step*. This instructional approach has been found to be very effective in helping students learn difficult physics concepts and improve their problem solving skills (VanLehn et al., 2005).

Figure 1 shows an example of physics problems provided in Andes Physics Tutor. While solving a physics problem in Andes Physics Tutor, students need to define physical quantities and variables, such as mass, velocity, acceleration, external force, etc., and choose an appropriate coordinate system for the given problem as if they were solving it on paper. Also, students have to specify the steps they find necessary while solving a problem. Andes Physics Tutor watches students, just like a human tutor, and provides appropriate instructional supports and guidance when students make a mistake during their problem solving processes (e.g., incorrectly define a physical quantity, choose a step that is not required to solve the problem or fail to resolve any required step) or in response to hint requests from students.

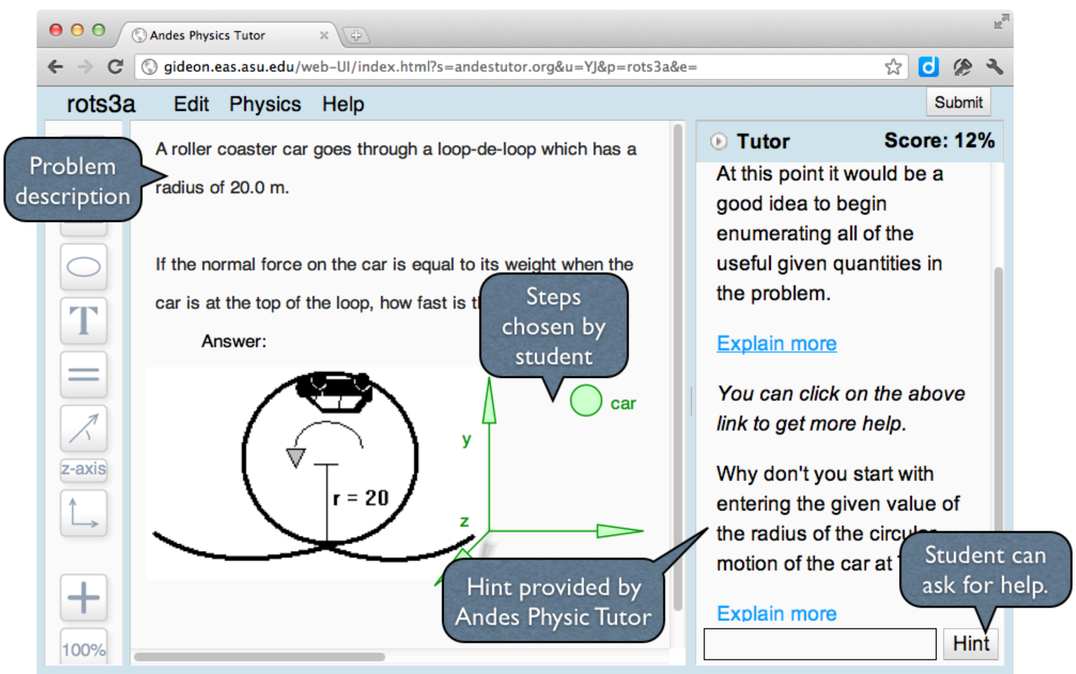

Figure 1. An example of Andes Physics Tutor problems

Data set

The data set analyzed in this study was obtained from the Pittsburgh Science of Learning Center (PSLC) (http://www.learnlab.org) which provides log files created from several computer-based learning environments they had developed in the past. Their DataShop Web service allows education researchers around the world to access log files depicting actions of students learning various subject matters, from foreign language to algebra and physics, in computer-based learning environments (Koedinger, Baker, Cunningham, Skogsholm, Leber, & Stamper, 2010). The data set analyzed in this study documents learning activities of 69 students enrolled in an introductory physics course at the United State Naval Academy (USNA) in Fall 2008 semester. The original data set obtained from the PSLC DataShop includes 175,267 database transactions where each transaction contains information about students, their learning activities and other information internally used in Andes Physics Tutor (Koedinger et al., 2010). Among these, seven variables relevant to the problems students solved and their problem solving processes were used to create explanatory variables that were employed in the statistical model developed in this study (see Table 1).

Table 1. Seven variables selected from the transaction record in the PSLC data set

Variables from PSLC data set	Description
Anonymized student ID	Anonymous student ID generated by PSLC's DataShop Web service
Time	Time at which the transaction occurred
Student response type	The type of attempt made by the student (e.g., ATTEMPT or HINT_REQUEST)
Problem name	The name of the problem associated with the transaction
Step name	The name of a problem solving step associated with the transaction
Outcome	The tutor's evaluation of the student's attempt (e.g., CORRECT, INCORRECT or HINT)
Knowledge Component (KC)	Knowledge Component associated with the transaction

Knowledge Component (KC) is defined as "an acquired unit of cognitive function or structure that can be inferred from performance on a set of related tasks" (Koedinger, Corbett, & Perfetti, 2010), which can be generalized to various cognitive constructs such as production rule (Newell, 1990; Anderson & Lebiere, 1998), schema (van Merriënboer & Sweller, 2005; Gick & Holyoak, 1983) or facet (Minstrell, 2001), and everyday terms such as concept, principle, fact or skill. In order to solve the physics problem shown in Figure 1, for example, students need to find out the acceleration of a car, which in turn requires an understanding of a few prerequisite physics concepts and problem solving skills such as (1) gravity and centripetal force, (2) setting up an appropriate coordinate system suitable for the problem, and (3) gravity as a centripetal force. These physics concepts and problem solving skills are the KCs required for calculating the acceleration of a car in the physics problem shown in Figure 1. PSLC researchers employed a statistical analysis technique called Learning Factor Analysis (LFA) to identify KCs for each problem solving step in all Andes Physics Tutor problems in the data set (Cen, Koedinger, & Junker, 2006).

Data pre-processing

Since each step in the Andes Physics Tutor problem is associated with one or more KCs, this study focused on students' interactions with the computer tutor from the problems that provided an opportunity to learn relevant KCs in the past, rather than looking at all past interactions saved in the log file. For the problem shown in Figure 1, for instance, which has three KCs (a centripetal force, how to set up an appropriate coordinate system, and the relationship between gravitational and centripetal forces), this study examined students' interactions with the computer tutor while they were trying to solve other problems requiring an understanding of at least one of these three KCs, instead of examining all of their past interactions recorded in the Andes Physics Tutor log file. In particular, this study looked at (1) how many times students had a chance to learn the required KCs in the past; (2) how long they spent before submitting their first answer; (3) whether their first attempt to solve relevant problems was correct or incorrect; (4) how many hints they requested; (5) how many correct answers they submitted and (6) how many mistakes they made while solving the problems requiring an understanding of at least one relevant KCs in the past.

Since all this information, except for the correctness of the submitted answer, is not directly available in the original data set obtained from the PSLC DataShop Web service (see Table 1), data pre-processing had to be performed. In order to facilitate data pre-processing, the downloaded data set, which was in a tab-delimited text format, was first imported into a MySQL database (http://www.mysql.com), and a data pre-processing program capable of accessing MySQL database tables was written in Python programming language (http://python.org). The data pre-processing program executed a series of SQL (Structured Query Language) queries against MySQL database tables to select students' past interactions with the computer tutor while trying to solve physics problems providing relevant learning opportunities. From 175, 267 transactions in the original log file obtained from the PSLC DataShop, the pre-processing program yielded a data set consisting of 35,549 records that were eventually analyzed in this study.

Building a statistical model predicting correct first try of students

Equation 1 shows a logistic regression model predicting problem solving performance of students, given their past interactions with Andes Physics Tutor.

$$Pr_{correct\ first\ try}(y_i = 1) = logit^{-1}(\beta_0 + \beta_1 C_{KC} + \beta_2 T_{correct} + \beta_3 T_{wrong} + \beta_4 N_{hint} + \beta_5 N_{correct} + \beta_6 N_{wrong} + \beta_7 N_{hint} C_{KC} + \beta_8 N_{hint} T_{correct} + \beta_9 N_{hint} T_{wrong} + \beta_{10} N_{hint} N_{correct} + \beta_{11} N_{hint} N_{wrong}) \quad (1)$$

Where, $logit^{-1}(x) = e^x / (1 + e^x)$,

β_i: Regression coefficients,
C_{KC}: Number of opportunities to learn KC,
$T_{correct}$: Average correct duration time,
T_{wrong}: Average error duration time,
N_{hint}: Average hint requests,
$N_{correct}$: Average correct answers,
N_{wrong}: Average wrong answers submitted while solving problems requiring an understanding of relevant KCs in the past

The outcome variable, y_i, is a binary variable indicating whether students were able to successfully accomplish a certain problem solving step in a particular Andes Physics Tutor problem at their first attempt without using hints available in the tutoring system. The statistical model shown above estimates $Pr_{correct\ first\ try}$ based on 11 explanatory variables consisting of 6 main effects and 5 interactions. The main effect explanatory variables try to capture important information that can affect the problem solving performance of students. For example, $T_{correct}$, the average correct duration time, measures the amount of time students spent before submitting their first correct answer while solving relevant problems in the past. Similarly, T_{wrong}, the average error duration time, measures the amount of time students spent before submitting their first incorrect answer or requesting hints while working on the problems requiring an understanding of relevant KC in the past. All explanatory variables were log-transformed because they were severely skewed, as shown in Figure 2, and their range was varied quite significantly. The log-transformation made explanatory variables follow an approximately normal distribution and have a comparable range (see Figure 2 and Table 2). Also, in order to help interpret the regression coefficients of the fitted model, all explanatory variables were centered by subtracting the mean of the data, which made each main effect coefficient correspond to a predictive difference with the other explanatory variables held constant at their average values (Gelman & Hill, 2007).

The development of the statistical model followed the general guidelines suggested by Gelman and Hill (2007). First, the data were fit to the logistic regression model containing only the main effect variables that might be expected to be important, for substantive reasons, in predicting the response variable. Then, for the main effect variable that has a large effect, their interactions with other main effect variables were included in the statistical model to see if they can explain the variance that was not accounted for in the main effect model. Among 6 main effect explanatory variables, N_{hint}, the number of hint requests made by students while solving relevant problems in the past, was found to explain the largest amount of variance in the data. Hence, the interaction terms involving N_{hint} were included in the final statistical model to examine whether such interaction terms can improve the predictive power of the logistic regression model.

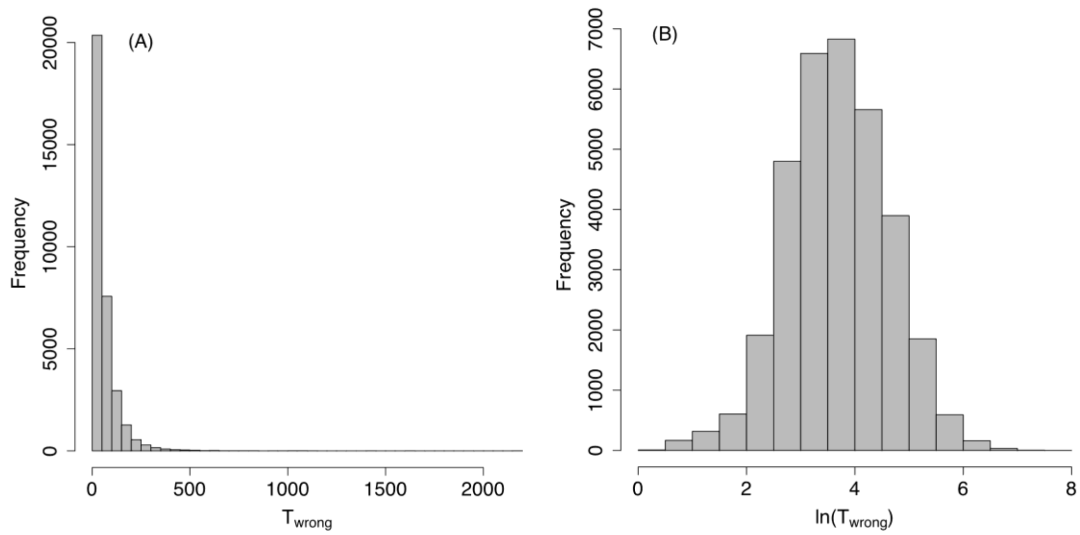

Figure 2. The distribution of original T_{wrong} and log-transformed T_{wrong}

Results

Descriptive statistics

Table 2 summarizes the descriptive statistics of the main effect variables in their linear and log-transformed forms. Students, on average, had about 10 opportunities to learn KCs for the current problem solving step in question. It is not surprising to find that students took more time before submitting their first answer when their answer was incorrect, compared to when their answer was turned out to be correct. Also, students requested about 3 hints, on average, while resolving a problem solving step requiring an understanding of relevant KC. As briefly mentioned above, the variables in their linear form are severely skewed. Therefore, Table 2 reports median and semi-interquartile range values for the original variables, instead of conventional mean and standard deviations.

Table 2. Descriptive statistics of main effect variables

Main effect	Original		Log-transformed	
variable	Median	Semi-interquartile range	Mean	*SD*
C_{KC}	10.00	12.50	2.29	1.42
$T_{correct}$	11.88	7.66	2.47	0.89
T_{wrong}	38.10	27.19	3.45	1.28
N_{hint}	3.00	1.00	1.36	0.52
$N_{correct}$	1.00	1.00	0.82	0.29
N_{wrong}	1.00	1.00	1.05	0.47

Logistic regression analysis

In order to investigate whether students' past problem solving performance and interactions with the computerized tutoring system recorded in the log files can predict $Pr_{correct\ first\ try}$, the probability for students to get each problem solving step correct at their first attempt without using hints, a logistic regression analysis was conducted. Table 3 summarizes the parameter estimates and the standard errors for the final logistic regression model developed in the study. The regression coefficients for all main effects and interactions were statistically significant at least at the 1% confidence level, indicating that the explanatory variables the data pre-processing program had produced were indeed useful in estimating the likelihood of the success of students in resolving the current problem solving step in question.

Table 3. Logistic regression model: Parameters and standard errors

Main effects	Parameter estimate	Standard error
Number of opportunities to learn KC (β_1)	-0.20^{***}	0.03
Correct duration time (β_2)	0.35^{***}	0.01
Error duration time (β_3)	0.23^{***}	0.03
Number of hint requests (β_4)	-0.75^{***}	0.03
Number of correct answers (β_5)	0.15^{**}	0.05
Number of wrong answers (β_6)	-0.11^{***}	0.03
Interactions		
Number of hint requests : Number of opportunities to learn KC (β_7)	-0.20^{***}	0.01
Number of hint requests : Correct duration time (β_8)	0.23^{***}	0.02
Number of hint requests : Error duration time (β_9)	0.21^{***}	0.02
Number of hint requests : Number of correct answers (β_{10})	-0.13^{**}	0.05
Number of hint requests : Number of wrong answers (β_{11})	0.32^{***}	0.03

$^{**}p < .01$. $^{***}p < .001$.

Evaluation of logistic regression model

In order to evaluate the discriminatory power of the developed logistic regression model, Percentage of Correct Classification (PCC) and c-statistic were computed. With the cut-off probability of 0.5, the developed logistic regression model was able to correctly identify 70.46% of the cases in the data set analyzed in this study. The model resulted in false positives (FP), classifying a student who in fact successfully resolved the problem solving step as ‘failed,’ at a rate of 13.60% (4,834 out of 35,549 cases) as shown in Table 4. The logistic regression model also resulted in false negatives (FN) 15.94% of the time by erroneously classifying 5,667 cases into the ‘successful’ group even though these students were in fact failed to resolve the problem solving step in question (see Table 4).

Table 4. Percentage of correct classification

Observed	Predicted		
	Resolved the step	Failed to resolve the step	Percentage correct
Resolved the step	9,994	4,834 (FP)	67.40
Failed to resolve the step	5,667 (FN)	15,054	72.65
Overall percentage			70.46

Data mining researchers typically perform a stratified 10-fold cross-validation analysis in order to estimate an overall error rate of a statistical learning model. In this evaluation approach, the data set is randomly divided into 10 parts in which the positive and the negative classes are represented in approximately the same proportions as in the full data set. Each part is held out in turn and the statistical learning model is built on the remaining nine-tenths; then its error rate is calculated on the holdout set. Therefore, the model building procedure is executed 10 times on different data sets and the 10 error estimates are then averaged to yield an overall error rate (Witten, Frank, & Hall, 2011). In this study, the 10-fold cross-validation analysis produced an overall prediction rate of 0.704, which is close to the PCC value obtained from the full data set, suggesting that the developed logistic regression model was fairly robust and it can have a similar predictive power on unseen data sets in the future.

Although PCC is commonly used to measure the discriminatory power of a statistical model, its error rate depends on a single cut-off probability (Kleinbaum & Klein, 2010). To address this issue, people often conduct a Receiver Operating Characteristic (ROC) curve analysis which produces a summary statistic based on a range of cut-off probabilities, not just one cut-off probability. When applied to a logistic regression model, an ROC curve is a plot of false positive vs. true positive derived from the cut-off probabilities for predicted values. Since a good logistic regression model will report small false positive and large true positive rates, the area under an ROC curve, which is often called AUC or c-statistic, will become larger as its discriminatory power grows (Kleinbaum & Klein, 2010). The AUC/c-statistic can vary from 0.5 (discriminating power not better than simple guessing) to 1.0 (perfect discriminating power) and is known to be equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (Fawcett, 2006). In this study, the AUC/c-statistic was found to be 0.78, indicating that the logistic regression model developed in this study has a fairly good predictive power for the data set used in the model building stage. In Figure 3, the black line is a ROC curve obtained from the logistic regression model of this study while the red dotted line represents a ROC curve with no discriminatory power (ACU/c-statistic = 0.5).

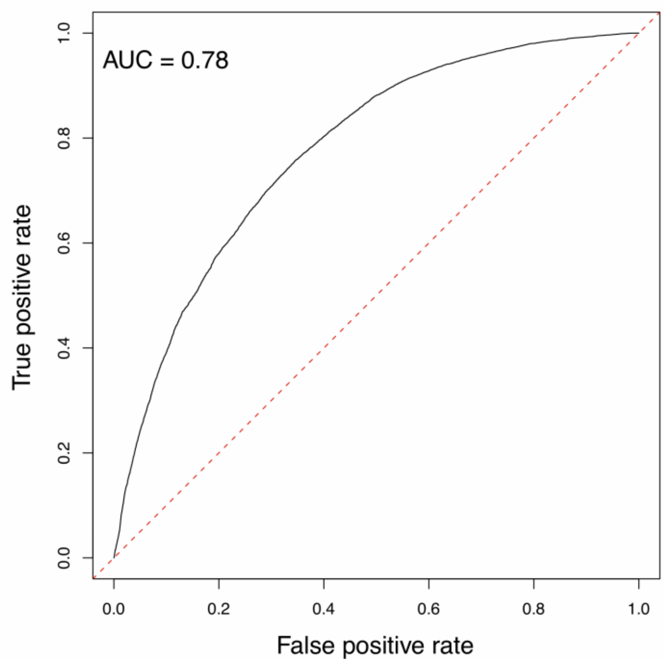

Figure 3. The ROC curve obtained from the developed logistic regression model

Since research on predicting students' problem solving performance from information captured in the log files of a computer-based learning environment is in its infancy, only a handful of studies are available that allows us to compare the predictive power of the logistic regression model reported in this study. Pardos and Heffernan (2011) incorporated problem difficulties into a standard Bayesian Knowledge Tracing (BKT) approach (Corbett & Anderson, 1995) to yield ROC values of 0.67 for ASSISTment and 0.53 for Cognitive Tutor data. Pardos, Gowda, Baker and Heffernan (2011) were able to obtain a ROC value of 0.77 when they applied a neural network that has varying numbers of hidden layers. Most recently, Pardos, Bergner, Seaton and Pritchard (2013) got ROC values of 0.65, 0.53, and 0.54 when they applied modified BKT to homework, lecture problems and exams in an edX MOOC data. Obviously, more in-depth studies are needed to systematically compare the predictive power of a logistic regression to other data mining algorithms. But, these results seem to indicate that the logistic regression model developed in this study has a comparable or better predictive power than other approaches.

Discussion

Main effects

The regression coefficients for the main effect explanatory variables suggest that the time students spend before submitting their first answer to the problems requiring an understanding of relevant KC, regardless of the correctness of their answers, is positively correlated with better problem solving performance in the future. Adding 1 to the log-transformed average correct duration time, $T_{correct}$, when the other explanatory variables are held constant at their mean values, corresponds to a positive difference in the probability of getting the current problem solving step correct about 8.8%. Similarly, a unit increase in the log-transformed average error duration time, T_{wrong}, is associated with an increase of about 5.8% in the correct first try probability. As is expected, students who submitted more correct or fewer incorrect answers, and requested fewer hints while solving relevant problems in the past showed much better performance on the current problem solving step in question. In particular, the number of hint requests, N_{hint}, showed the largest predictive power in discriminating the success of students; adding 1 to the log-transformed

N_{hint}, when the other explanatory variables are held constant at their mean values, would yield about 18.8% decrease in the probability of getting the current problem solving step correct (see Table 3).

Interestingly, the coefficient for the number of opportunities to learn required KCs, C_{KC}, was found to be negative, indicating that the more chances students had to learn relevant KCs, the more likely they are to get the current problem solving step *incorrect*; the increase of 1 in the log-transformed C_{KC} decreases the probability for students to get the current problem solving step correct by about 5% when the other explanatory variables are held constant at their mean values (see Table 3). At first glance, this result looks counter-intuitive because it would be reasonable to assume that students should be able to extract a schema while solving relevant problems in the past, which would in turn help them resolve the current problem solving step at hand. One possible explanation for this seemingly counter-intuitive result is that students with higher C_{KC} might be academically weaker students. Hence, these students did not do well even though they worked on more Andes Physics Tutor problems in the past. This interpretation is in part supported by the fact that the USNA students were allowed to solve as many Andes Physics Tutor problems as they found it necessary; the number of Andes Physics Tutor problems the USNA students solved in this semester varied greatly, from 1 to 74, indicating that Andes Physics Tutor was used as an add-on to regular learning activities, such as lectures and written problem sets. As briefly mentioned earlier, Andes Physics Tutor provides appropriate feedback and instructional supports while students are trying to solve difficult physics problems, which would make it more appealing to academically weaker students who need extra help. Further data analyses (e.g., examining academic performance of students measured independently) would be needed in order to find a more definite answer to this question.

Interaction effects

The negative coefficient for the interaction between the number of opportunities to acquire relevant KCs, C_{KC}, and the number of hint requests, N_{hint}, indicates that the importance of the number of opportunities to learn relevant KCs as an explanatory variable increases for students who used more hints while solving relevant problems in the past (and vice versa); for a unit increase in the log-transformed number of hint requests, the additional value of -0.20 is added to the coefficient for the number of learning opportunities students had, which is also negative. As a result, the probability of submitting a correct answer for the current problem solving step decreases slightly faster when students used more hints while solving relevant problems in the past.

Also, the positive coefficients for the interactions between the number of hint requests, N_{hint}, and the duration times, $T_{correct}$, T_{wrong}, suggest that the negative main effect of the number of hint requests would be mitigated if students spent more time in solving the relevant problems before submitting their first answer, regardless of their correctness. In other words, the logistic regression model predicts that the students who spent more time before submitting their first answer, among students who requested the same number of hints, are more likely to successfully resolve the current problem solving step at their first attempt without requesting hints. Previous study has found that learners often abuse instructional supports available in the compute-based learning environment without actually engaging in learning (Baker, Walonoski, Heffernan, Roll, Corbett, & Koedinger, 2008). The positive coefficients for the interactions between the number of hints and the duration times seem to support Baker et al. (2008)'s finding; using many hints without spending enough time is not positively correlated with learning probably because students did not exert enough cognitive effort to learn the presented information.

As discussed in the main effect section, the more wrong answers students submitted in the past, the less likely they are to get the current problem solving step correct. However, the positive coefficient for the interaction between the hint requests, N_{hint}, and the number of wrong answers submitted in the past, N_{wrong}, indicates that the main effect from the wrong answer submission changes depending on the hint use of students. As students use more hints, as long as they invested enough time in using them, the negative effect of submitting wrong answers seems to be decreasing. When the other explanatory variables including the number of hint requests are held constant at their mean values, the probability of correct first try decreases as the number of incorrect submissions occurred in the past increases. However, when students used more hints, the probability of correct first try *increases* because the positive coefficient of the interaction term balances out the negative main effect from the number of wrong answers (see Figure 4A).

Similarly, the negative coefficient for the interaction between hint requests, N_{hint}, and correct answer submissions, $N_{correct}$, appears to cancel out the positive main effect from the number of correct answer submissions. As shown in the solid line in Figure 4B, the main effect for the number of correct answers students submitted in the past is associated with a higher probability of getting the current problem solving step correct without using hints available in the learning environment. However, as students used more hints while solving relevant problems in the past, the probability of correct first try does not increase as much (see the dotted line in Figure 4B). These results seem to suggest the conditions under which the use of hints in Andes Physics Tutor is most beneficial to students. Students are more likely to resolve the current problem solving step successfully (1) when they use hints in conjunction with their incorrect answers; (2) when they do not use too many hints on one problem; and (3) when they invest enough time trying to understand the information presented in hints.

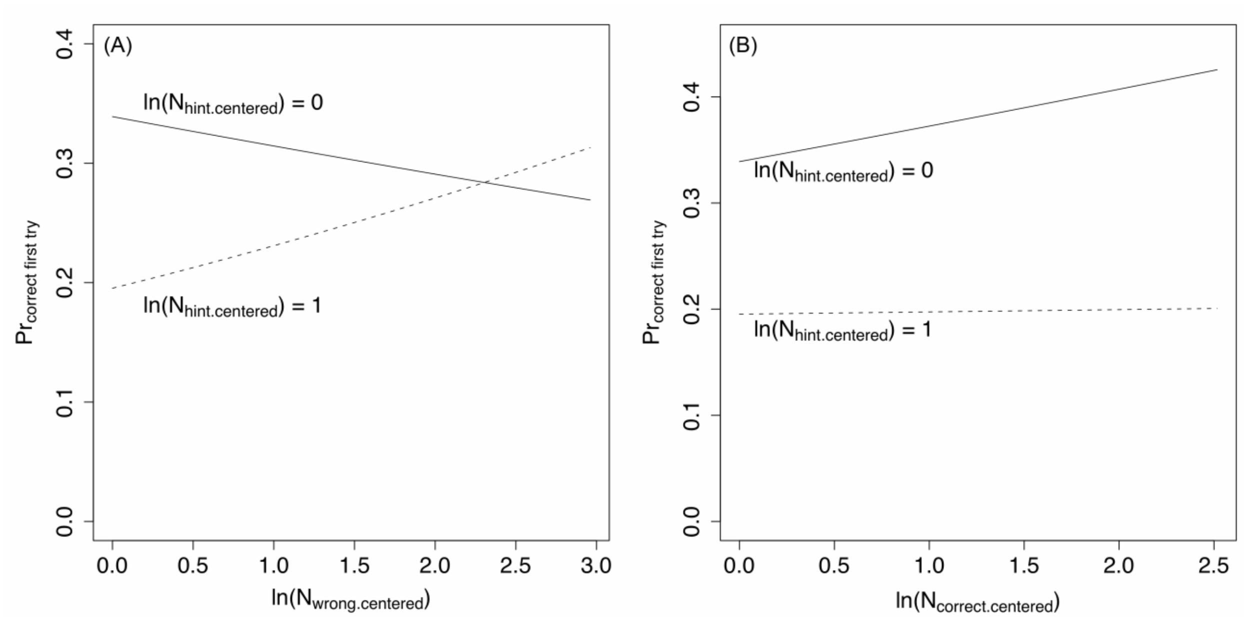

Figure 4. Interaction effects between N_{hint} and N_{wrong} (A) and N_{hint} and $N_{correct}$ (B)

Conclusion

In this study, the log files capturing how college students used a computer-based tutor while trying to solve difficult physics problems were carefully analyzed to build a statistical model that estimated the likelihood of their successful problem solving from their problem solving history observed in the past. The logistic regression model developed in this study was able to predict the problem solving performance of students with an error rate of about 30%. Various model evaluation methods, such as AUC/c-statistic and 10-fold cross-validation, suggest that the developed statistical model was fairly robust and would be able to maintain a similar predictive accuracy with unseen cases in the future.

Koedinger and Aleven (2007) point out that it is critical to balance giving and withholding information or instructional supports in a computerized learning environment in order to maximize student learning. If learning environments give students too much information prematurely, students may not be able to acquire a schema from learning tasks because they do not exert enough cognitive efforts (Kapur, 2008; Schmidt & Bjork, 1992). On the other hand, if learning environments do not provide instructional supports, academically weaker students are likely to flounder, waste their time with no success, and get frustrated with failure. In most computer-based learning environments, simple heuristics or the learner’s discretion is used to determine when to provide instructional supports. However, simple heuristics, such as giving hints or feedback after students fail to resolve a learning task N times, would be unlikely to find the right moment for providing instructional assistance that can maximize the learning outcome of students. Likewise, providing instructional supports on the learner’s demand may not lead to improved learning because previous studies found that especially novice learners do not possess enough metacognitive ability and prior knowledge required to determine the right moment to ask for help (Clark & Mayer, 2003; Lawless & Brown, 1997). The findings of this study may suggest that it is possible to develop a computerized learning environment that can provide instructional assistance based on the estimated probability of successful problem solving. A logistic regression model built from the log files created by students in the past semesters can be used to estimate a probability for students in the current semester to successfully solve physics homework problems at their first attempt. Instructional scaffolding can then be provided only when the estimated probability is less than a certain threshold value. Further research is needed to investigate how to determine an optimum probability threshold, and to examine whether this approach can in fact improve students' problem solving performance.

The statistical model reported in this study was developed from the log files created by all students enrolled in the general physics course taught at USNA. Since Andes Physics Tutor seemed to be used as an add-on to regular class activities, such as lectures and written problem sets, some students used the computer tutor more actively than others. Therefore, analyzing the subset of data including only students who used Andes Physics Tutor more actively might lead to the development of a more accurate and robust statistical model. Similarly, the current statistical model considered all previous problems that were relevant to the current problem solving step in question, which resulted in a rather long data pre-processing time. By limiting the data pre-processing to most recent, instead of all, problems relevant to the current problem solving step, it would be possible to cut down on the data pre-processing time while maintaining similar discriminatory accuracy and robustness in the final prediction model. Further research is required to determine the number of recent relevant problems that can maximize the predictive power of the statistical model while minimizing data pre-processing time.

The fact that there are many different ways to solve physics problems, especially in the college level, might have made it more difficult to build a robust statistical model that can make accurate predictions on the problem solving performance of students. Therefore, it would be interesting to build a predictive statistical model in other domains, such as elementary algebra, that are less complicated than college physics, and compare its accuracy and robustness to those obtained from more complicated knowledge domains.

Also, it would be valuable to investigate whether more advanced statistical and data mining algorithms can be used to build a better predictive model. The logistic regression used in this study belongs to a family of statistical analysis models called Generalized Linear Models (GLM) which assume a linear relationship between explanatory and response variables. Although linear models such as logistic regression have advantages over other approaches in terms of interpretation and inference, it may have weaker predictive power because the linearity assumption is almost always an approximation. There are several ways to address this issue. The simplest approach would be to extend the logistic regression model by adding extra explanatory variables by raising each explanatory variable to a certain power. Employing Generalized Additive Models (GAM) might be another approach that can relax the linearity assumption of GLM. GAM provides a general framework for extending a standard linear model by allowing non-linear functions of each of the explanatory variables while maintaining additivity (James, Witten, Hastie, & Tibshirani, 2013). Similarly, more advanced data mining algorithms that involve non-linear transformation of explanatory variables, such as Support Vector Machines (Cristianini & Shawe-Taylor, 2000), might be used to build a better predictive model.

Another line of research worth pursuing is to investigate the effect of problem difficulty. Obviously, students would take more time, make more mistakes and request more hints when they were working on more difficult problems. However, the difficulty of problems was not considered in the logistic regression model developed in this study. More research is required to find out the best way to incorporate the problem difficulty into the statistical model building process, and to examine whether it can improve the predictive power of the resulting statistical model.

References

Allen, I. E., & Seaman, J. (2010). *Learning on demand: Online education in the United States, 2009*. Retrieved August 11, 2014, from http://files.eric.ed.gov/fulltext/ED529931.pdf

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185–224.

Cen, H., Koedinger, K. R., & Junker, B. (2006). Learning factors analysis: A general method for cognitive model evaluation and improvement. In M. Ikeda, K. D. Ashley, & T.-W. Chan (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 164–175). Berlin, Germany: Springer-Verlag.

Clark, R. C., & Mayer, E. (2003). *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. San Francisco, CA: Pfeiffer.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods.* New York, NY: Cambridge University Press.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*(1), 1–38.

Hwang, W.-Y., & Wang, C.-Y. (2004). A study of learning time patterns in asynchronous learning environments. *Journal of Computer Assisted Learning*, *20*(4), 292–304.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York, NY: Springer.

Kapur, M. (2008). Productive failure. *Cognition and Instruction*, *26*(3), 379–424.

Kleinbaum, D. G., & Klein, M. (2010). *Logistic regression: A self-learning text.* New York, NY: Springer.

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, *19*(3), 239–264.

Koedinger, K. R., Baker, R. S. J. d., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A Data Repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy, R. S. J. D.Baker (Eds.), *Handbook of Educational Data Mining* (pp. 43–55). Boca Raton, FL: CRC Press.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2010). *The knowledge-learning-instruction (KLI) framework: Toward bridging the science-practice chasm to enhance robust student learning*. Retrieved August 11, 2014, from http://www.learnlab.org/documents/KLI-Framework.pdf

Lawless, K. A., & Brown, S. W. (1997). Multimedia learning environments: Issues of learner control and navigation. *Instructional Science*, *25*(2), 117–131.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, *53*(3), 950–965.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, *54*(2), 588–599.

Minstrell, J. (2001). Facets of students' thinking: Designing to cross the gap from research to standards-based practice. In K. Crowley, C. D. Schunn & T. Okada (Eds.), *Designing for Science: Implications for Professional, Instructional, and Everyday Science* (pp. 415–443). Mahwah, NJ: Lawrence Erlbaum Associates.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Palazzo, D. J., Lee, Y.-J., Warnakulassoriya, R., & Pritchard, D. E. (2010). Patterns, correlates, and reduction of homework copying. *Physical Review Special Topics – Physics Education Research*, *6*, 010104

Pardos, Z., Bergner, Y., Seaton, D., & Pritchard, D. (2013). Adapting Bayesian knowledge tracing to a massive open online course in edX. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 137–144). Memphis, TN, USA: International Educational Data Mining Society.

Pardos, Z., Gowda, S., Baker, R., Heffernan, N. (2011). The sum is greater than the parts: Ensembling models of student knowledge in educational software. *ACM SIGKDD Explorations Newsletter*, *13*(2), 37–44.

Pardos, Z., & Heffernan, N. (2011). KT-IDEM: Introducing item difficulty to the knowledge tracing model. *Lecture Notes in Computer Science*, *6787*, 243–254.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*(4), 207–217.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Robert S., Taylor, L.,…Wintersgill, M. (2005). The Andes physics tutoring system: Lesson learned. *International journal of Artificial Intelligence in Education*, *15*(3), 147–204.

van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17*(2), 147–177.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques, third edition.* San Francisco, CA: Morgan Kaufman.