# Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School

Maria Ofelia Z. San Pedro[1], Ryan S.J.d. Baker[1], Alex J. Bowers[1], Neil T. Heffernan[2]
[1]Teachers College Columbia University, 525 W 120th St. New York, NY 10027
[2]Worcester Polytechnic Institute, 100 Institute Rd. Worcester, MA 01609

mzs2106@tc.columbia.edu, baker2@exchange.tc.columbia.edu, bowers@tc.edu, nth@wpi.edu

## ABSTRACT

Research shows that middle school is an important juncture for a student where he or she starts to be conscious about academic achievement and thinks about college attendance. It is already known that access to financial resources, family background, career aspirations and academic ability are indicative of a student's choice to attend college; though these variables are interesting, they do not necessarily give sufficient actionable information to instructors or guidance counselors to intervene for individual students. However, increasing numbers of students are using educational software at this phase of their education, and detectors of specific aspects of student learning and engagement have been developed for these types of learning environments. If these types of models can be used to predict college attendance, it may provide more actionable information than the previous generation of predictive models. In this paper, we predict college attendance from these types of detectors, in the context of 3,747 students using the ASSISTment system in New England, producing detection that is both successful and potentially more actionable than previous approaches; we can distinguish between a student who will attend college and a student who will not attend college 68.6% of the time.

## Keywords

College Enrollment, Affect Detection, Knowledge Modeling, Educational Data Mining

## 1. INTRODUCTION

The processes leading a student to choose to attend college starts early, and decisions can begin to solidify as early as middle school (ages 12-14). Especially in the United States, successful learning experiences which develop key skills build positive self-beliefs, interests, goals and actions, making students likely to actively seek and plan higher educational goals and career aspirations [31] As students go through middle school, they increasingly find themselves engaged or disengaged from school and learning. This process is driven in part by changes in students' self-perceptions, whether they see themselves as smart and capable of taking the courses in high school. This leads to students making decisions about how academic achievement, certain careers, and college majors fit into their self-perception [14].

It is during middle school that students either start to value academic achievement or begin to get off track and start to become frustrated and disengaged in school [8]. Research findings suggest that middle school students often think about going to college but fail to get support in planning how to achieve this [14].The transition to middle school in the United States has been associated with a decline in academic achievement, performance motivation, and self-perception [33]. For students who fail to develop a plan (or obtain support) for a college future, this has a dramatic effect on what eventually happens to the students.

Disengagement not only leads to negative attitudes about higher education, but also to poorer learning [17, 27, 29]. This leads to an unsuccessful learning experience when they reach high school, ultimately leading to dropping out, or disinterest in pursuing post-secondary education [7]. Multiple studies have shown that it is possible to predict which students will eventually drop out of high school, as early as late elementary or middle school [7, 9, 10, 35], with evidence that some particularly predictive factors include problem behaviors [7, 35] and shifts in academic achievement over time [9, 10]. Indicators of fairly extreme forms of disengaged behavior (low attendance, misconduct) and academic failure in sixth grade have been shown to be strong predictors of students falling off the path towards graduation and therefore eventual college attendance, within longitudinal analysis [7].

By contrast, students who have already made college plans when they are in middle school tend to be more likely to attend college in spite of challenges [13]. They tend to plan appropriate courses to take when they enter high school, and get involved in relevant extracurricular activities that contribute to college admission [14]. In effect, they become interested in achieving a good academic record. Thus, examining the factors that influence students' engagement and disengagement during middle school is crucial so as to understand better the factors that lead students to fail to attend college, and the possible paths to re-engaging students.

However, one of the major limitations to this past research is that it identifies changes in student engagement only through fairly strong indicators of disengagement, such as failing grades [9, 10, 11], problem behaviors such as violence in school [47] and non-attendance [35]. By the time these indicators are commonplace, it may be quite late to make an intervention. If it were possible to identify useful and actionable antecedents to these changes, it might be possible to intervene more effectively.

One potential source of data on early change in engagement is the log files from educational software. In recent years, the use of educational software at the middle school level has expanded considerably, with systems such as ASSISTments [40] being used by rapidly increasing numbers of students. At the same time, educational data mining (EDM) techniques have been applied to logs from educational software to model a range of affect and engagement constructs, including gaming the system [4], off-task behavior [1], carelessness [45], boredom [22, 37, 44], frustration [22, 37, 44], and engaged concentration [22, 37, 44],. Automated detectors of disengaged behaviors can predict differences in learning, both in the relatively short term [cf. 17] and over the course of a year [26, 37]. Within this paper, we extend this work to study how learning and engagement in middle school – as assessed by this type of automated detector – can be used to predict college enrollment. To our knowledge, this is the first study that aims at predicting college enrollment from affect and engagement inferred from the logs of educational software used years earlier. We conduct this research in a data set of 3,747 students who used the ASSISTment system [40], between 2004 and 2007. We discuss which aspects of learning and engagement predict college enrollment, and conclude with a discussion of potential implications for the design and interventions of interactive educational systems for sustained attendance and engagement in school.

# 2. METHODOLOGY

## 2.1 The ASSISTment System

Within this paper, we investigate this issue within the context of ASSISTments. The ASSISTment system, shown in Figure 1, [40] is a free web-based tutoring system for middle school mathematics that *assesses* a student's knowledge while *assisting* them in learning, providing teachers with detailed reports on the skills each student knows. The ASSISTment system, shown in Figure 1, provides feedback on incorrect answers. Within the system, each mathematics problem maps to one or more cognitive skills. When students working on an ASSISTment answer correctly, they proceed to the next problem. If they answer incorrectly, they are provided with scaffolding questions which break the problem down into its component steps. The last step of scaffolding returns the student to the original question (as in Figure 2). Once the correct answer to the original question is provided, the student is prompted to go to the next question.

## 2.2 Data

### *2.2.1 ASSISTments Data*

Action log files from the ASSISTment system were obtained for a population of 3,747 students that came from middle schools in New England, who used the system at various times starting from school years 2004-2005 to 2006-2007 (with a few students continuing tutor usage until 2007-2008 and 2008-2009). These students were drawn from three districts who used the ASSISTment system systematically during the year. One district was urban with large proportions of students requiring free or reduced-price lunches due to poverty, relatively low scores on state standardized examinations, and large proportions of students learning English as a second language. The other two districts were suburban, serving generally middle-class populations. Overall, the students made 2,107,108 actions within the software (where an action consisted of making an answer or requesting help), within 494,150 problems, with an average of 132 problems per student. Knowledge, affect, and behavior models were applied to this dataset, creating features that could be used for our final prediction model of college enrollment.

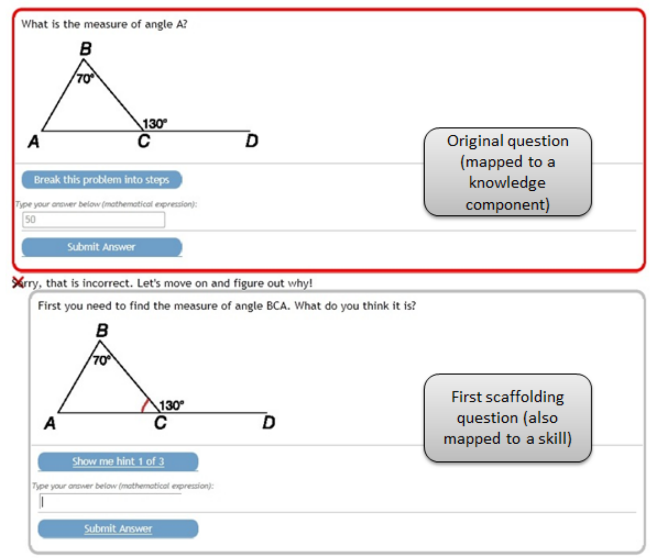


**Figure 1. Example of an ASSISTments Problem. If a student gets it incorrect, scaffolding problems are there to aid the student in eventually getting the correct answer.**

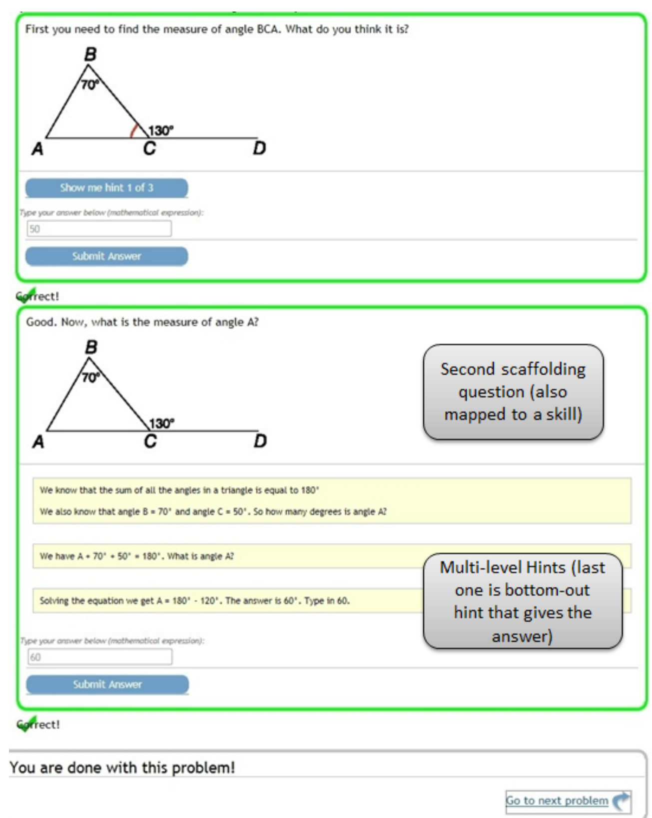


**Figure 2. Example of Scaffolding in an ASSISTments Problem.**

### 2.2.2 College Enrollment Data

For college enrollment information, enrollment records of these 3,747 students were requested from the National Student Clearinghouse (http://www.studentclearinghouse.org). For the purposes of the analyses in this paper, we identified solely whether each student was enrolled in a college or not, and used this as our labels in training our model. Additional information (such as whether the student graduated from college) is generally available from the Clearinghouse, but will not be available for these students for a few more years.

## 2.3 Creation of Model Features

In order to predict and analyze college enrollment, we distilled a range of features from the log files of ASSISTments, including student knowledge estimates, student affect (boredom, engaged concentration, confusion), student disengaged behaviors (off-task, gaming the system, carelessness), and other information of student usage (the proportion of correct actions and the number of first attempts on problems made by the student, a proxy for overall usage). These features were either directly distilled from the logs or obtained from automated detectors applied to the data set.

### 2.3.1 Student Knowledge Features

Estimates of student knowledge were computed using Bayesian Knowledge Tracing (BKT) [20], a model used in several ITSs to estimate a student's latent knowledge based on his/her observable performance. This model can predict how difficult the current problem will be for the current student, based on the skills involved in that problem. As such, this model can implicitly capture the tradeoff between difficulty and skill for the current context. This model can inform us whether student skill is higher than current difficulty (resulting in a high probability of correctness), when current difficulty is higher than student skill (resulting in a low probability of correctness), and when difficulty and skill are in balance (medium probabilities of correctness). To assess student skill, BKT infers student knowledge by continually updating the estimated probability a student knows a skill every time the student gives a first response to a new problem. It uses four parameters, each estimated separately per skill – $L_O$, the initial probability the student knows the skill; T, the probability of learning the skill at each opportunity to use that a skill; G, the probability that the student will give a correct answer despite not knowing the skill; and S, the probability that the student will give an incorrect answer despite knowing the skill. In this model, the four parameters for each skill are held constant across contexts and students (variants of BKT relax these assumptions). BKT uses Bayesian algorithms after each student's first response to a problem in order to re-calculate the probability that the student knew the skill before the response. Then the algorithm accounts for the possibility that the student learned the skill during the problem in order to compute the probability the student will know the skill after the problem [20]. With the data from the 2004-2005 to 2006-2007 logs, BKT model parameters were fit by employing brute-force grid search [cf. 3].

### 2.3.2 Affect and Behavior Features

To obtain affect and behavior assessments, we leverage existing detectors we developed of student affect and engaged/disengaged behavior within the ASSISTment system [37], to help us understand student affect and behavior across contexts. Detectors of three affective states are utilized: engaged concentration, boredom, and confusion. Detectors of three disengaged behaviors are utilized: off-task, gaming, and slip or carelessness. These detectors of affect and behavior are identical to the detectors used in [37]. They were developed in a two-stage process: first, student affect labels were acquired from the field observations (reported in [37]), and then those labels were synchronized with the log files generated by ASSISTments at the same time (forming our first dataset). This process resulted in automated detectors that can be applied to log files at scale, specifically the data set used in this project (the 2004-2005 to 2006-2007 data set). To enhance scalability, only log data was used as the basis of the detectors, instead of using physical sensors. The research presented in this paper could not have been conducted if physical sensors were used. The detectors were constructed using only log data from student actions within the software occurring at the same time as or before the observations, making our detectors usable for automated interventions, as well as the discovery with models analyses presented in this paper.

The affect detectors' predictive performance were evaluated using A' [28] and Cohen's Kappa [18]. An A' value (which is approximately the same as the area under the ROC curve [28]) of 0.5 for a model indicates chance-level performance for correctly determining the presence or absence of an affective state in a clip, and 1.0 performing perfectly. Cohen's Kappa assesses the degree to which the model is better than chance at identifying the affective state in a clip. A Kappa of 0 indicates chance-level performance, while a Kappa of 1 indicates perfect performance. A Kappa of 0.45 is equivalent to a detector that is 45% better than chance at identifying affect.

As discussed in [37], all of the affect and behavior detectors performed better than chance. Detector goodness was somewhat lower than had been previously seen for Cognitive Tutor Algebra [cf. 6], but better than had been seen in other published models inferring student affect in an intelligent tutoring system solely from log files (where average Kappa ranged from below zero to 0.19 when fully stringent validation was used) [19, 22, 44]. The best detector of engaged concentration involved the K* algorithm, achieving an A' of 0.678 and a Kappa of 0.358. The best boredom detector was found using the JRip algorithm, achieving an A' of 0.632 and a Kappa of 0.229. The best confusion detector used the J48 algorithm, having an A' of 0.736, a Kappa of 0.274. The best detector of off-task behavior was found using the REP-Tree algorithm, with an A' value of 0.819, a Kappa of 0.506. The best gaming detector involved the K* algorithm, having an A' value of 0.802, a Kappa of 0.370. These levels of detector goodness indicate models that are clearly informative, though there is still considerable room for improvement. The detectors emerging from the data mining process had some systematic error in prediction due to the use of re-sampling in the training sets (models were validated on the original, non-resampled data), where the average confidence of the resultant models was systematically higher or lower than the proportion of the affective states in the original data set. This type of bias does not affect correlation to other variables since relative order of predictions is unaffected, but it can reduce model interpretability. To increase model interpretability, model confidences were rescaled to have the same mean as the original distribution, using linear interpolation. Rescaling the confidences this way does not impact model goodness, as it does not change the relative ordering of model assessments.

#### 2.3.2.1 *Application of Affect and Behavior Models to Broader Data Set*

Once the detectors of student affect and behavior were developed, they were applied to the data set used in this paper. As mentioned, this data set was comprised of 2,107,108 actions in 494,150 problems completed by 3,747 students in three school districts. The result was a sequence of predictions of student affect and behavior across the history of each student's use of the ASSISTment system.

#### 2.3.2.2 *Carelessness Detection using Logs*

Different from the process above, the incidence of carelessness within the Cognitive Tutor was traced with a model designed to assess "slips" [2]. Slips in that paper are operationalized in a fashion essentially identical to prior theory of how to identify careless errors [16]. The model used in [2], termed the Contextual Slip model, contextually estimates the probability that a specific student action indicates a slip/carelessness, whenever the student reaches a problem step requiring a specific skill, but answers incorrectly. The probability of carelessness/slip is assessed contextually, and is different depending on the context of the student error. The probability estimate varies based on several features of the student action and the situation in which it occurs, including the speed of the action, and the student's history of help-seeking from the tutor. As such, the estimate of probability of carelessness/slip is different for each student action.

The Contextual Slip model is created using the BKT approach previously discussed. Note that in the BKT model – used in creating the Contextual Slip model – the four parameters for each skill are invariant across the entire context of using the tutor, and invariant across students. We use BKT as a baseline model to create first-step estimations of the probability that each action is a contextual slip. These estimations are not the final Contextual Slip model, but are used to produce it. Specifically, we use BKT to estimate whether the student knew the skill at each step. In turn, we use these estimates, in combination with Bayesian equations, to label incorrect actions with the probability that the actions were slips, based on the student performance on successive opportunities to apply the rule. More specifically, given the probability that the student knows the skill at a specific time, Bayesian equations and the static BKT parameters are utilized to compute labels for the Slip probabilities for each student action (A) at time N, using future information (two actions afterwards – N+1, N+2). In this approach, we infer the probability that a student's incorrectness at time N was due to not knowing the skill, or whether it is due to a slip. The probability that the student knew the skill at time N can be calculated, given information about the actions at time N+1 and N+2 ($A_{N+1,N+2}$), and the other parameters of the Bayesian Knowledge Tracing model:

$$P(A_N \text{ is a Slip} \mid A_N \text{ is incorrect}) = P(L_n \mid A_{N+1,N+2}). \quad (1)$$

This gives us a first estimate that a specific incorrect answer is a slip. However, this estimate uses data on the future, making it impossible to use to assess slip in real-time. In addition, there is considerable noise in these estimates, with estimates trending to extreme values that over-estimate slip in key situations due to limitations in the original BKT model [2]. But these estimated probabilities of slip can be used to produce a less noisy model that can be used in real time, by using them as training labels (e.g. inputs) to machine-learning. Specifically, a linear regression model is created that predicts slip/carelessness contextually. The result is a model that can now predict at each practice opportunity whether an action is a slip, using only data about the action itself, without any future information. This model has been shown to predict post-test scores, even after student knowledge is controlled for [3].

Once these labels are obtained from BKT, the labels are smoothed by training models with each student action originally labeled with the probability estimate of slip occurrence, using information on that student action generated on our tutor logs. For each action, a set of numeric or binary features from the ASSISTments logs were distilled, based on earlier work by [5].

As in previous work to model slipping, the features extracted from each student action within the tutor were used to predict the probability that the action represents a slip/carelessness. The prediction took the form of a linear regression model, fit using M5-prime feature selection in the RapidMiner data mining package [32]. This resulted in numerical predictions of the probability that a student action was a careless error, each time a student made a first attempt on a new problem step. Linear regression was chosen as an appropriate modeling framework when both predictor variables and the predicted variable are numeric. In addition, linear regression functions well with noisy educational data, creating relatively low risk of finding an "over-fit" model that does not function well on new data.

Six-fold student-level cross-validation [24] was conducted to evaluate the carelessness detector's goodness. Cross-validating at this level allows us to assess whether the model will remain effective for new students drawn from the same overall population of students studied. Carelessness models were trained separately per school year of the ASSISTments data set. They were assessed in terms of cross-validated correlation. The carelessness models trained within the ASSISTments data achieved a cross-validation correlation of r = 0.458 on the average.

## 2.4 Logistic Regression

Models were built to predict whether a student attended college. Aggregate predictor variables were created by taking the average of the predictor feature values for each student, resulting in one record per student (in other words, taking the average boredom per student, average confusion per student, etc.).

A multiple-predictor logistic regression model was fitted to predict whether a student will enroll in college from a combination of features of his or her student affect, engagement, knowledge and other information on student usage (the proportion of correct actions, and the number of first attempts on problems made by the student, a proxy for overall usage) of a tutoring system during middle school. We used logistic regression analysis since we have a dichotomous outcome – whether or not the student would be enrolled in college – resulting in a non-linear relationship between our predictors and outcome variable. Choosing logistic regression allows for relatively good interpretability of the resultant models, while matching the statistical approach used in much of the other work on predicting college attendance [12, 23, 36, 46]. In essence, the logistic model predicts the logit (natural logarithm of an odds ratio [cf. 39]) of an outcome variable from a predictor or set of predictors. The odds ratio in logistic regression is the odds of an event occurring given a particular predictor, divided by the odds of an event occurring given the absence of that particular predictor. An odds ratio over 1.0 signifies that the independent variable increases the odds of the dependent variable occurring; correspondingly, an odds ratio under 1.0 signifies that the independent variable decreases the odds of the dependent variable occurring.

Features were selected using a simple backwards elimination feature selection, based on each parameter's statistical significance. All predictor variables were standardized (using z-scores), in order to increase interpretability of the resulting odds ratios (note that this does not impact model goodness or predictive power in any fashion). The odds ratio indicates the odds that a class variable increases per one unit change of a predictor (per one SD change for standardized predictors). Standardizing the predictors enables us to show a clear indication of each predictor's contribution to the class variable (college enrollment).

## 3. RESULTS

Before developing the model, we looked at our original, non-standardized features and how their values compare between those who were labeled to have attended college and those who have not (Table 1).

**Table 1. Features for Students who Attended College (1, n = 2166) and did not Attend college (0, n = 1581)**

| | College | Mean | Std. Dev. | Std. Error Mean | t-value |
|---|---|---|---|---|---|
| Slip/ Carelessness | 0 | 0.132 | 0.066 | 0.002 | -13.361 |
| | 1 | 0.165 | 0.077 | 0.002 | (p<0.01) |
| Student Knowledge | 0 | 0.292 | 0.151 | 0.004 | -15.481 |
| | 1 | 0.378 | 0.180 | 0.004 | (p<0.01) |
| Correctness | 0 | 0.382 | 0.161 | 0.004 | -17.793 |
| | 1 | 0.483 | 0.182 | 0.004 | (p<0.01) |
| Boredom | 0 | 0.287 | 0.045 | 0.001 | 5.974 |
| | 1 | 0.278 | 0.047 | 0.001 | (p<0.01) |
| Engaged Concentration | 0 | 0.483 | 0.041 | 0.001 | -11.979 |
| | 1 | 0.500 | 0.044 | 0.001 | (p<0.01) |
| Confusion | 0 | 0.130 | 0.054 | 0.001 | 5.686 |
| | 1 | 0.120 | 0.052 | 0.001 | (p<0.01) |
| Off-Task | 0 | 0.304 | 0.119 | 0.003 | 1.184 |
| | 1 | 0.300 | 0.116 | 0.002 | p=0.237 |
| Gaming | 0 | 0.041 | 0.062 | 0.002 | 8.862 |
| | 1 | 0.026 | 0.044 | 0.001 | (p<0.01) |
| Number of First Actions | 0 | 114.500 | 91.771 | 2.308 | -8.673 |
| | 1 | 144.560 | 113.357 | 2.436 | (p<0.01) |

From Table 1, initial observations show that average knowledge estimate, average correct, number of first actions, average slips/carelessness, and average engaged concentration had higher mean values for students who attended college. Average boredom, average confusion, average off-task and average gaming had higher mean values for those who did not attend college. Conducting an independent samples t-test (equal variances assumed) indicates that, with the exception of off-task, the difference of means of each feature between the two groups are statistically significant.

These observations align with the individual effects of each feature on the prediction of college enrollment. For example, there is a strong positive relationship between college enrollment and average correct answers (CollegeEnrollment = 0.612 Correctness + 0.346, $\chi^2(df = 1, N = 3747) = 304.141$, $p < 0.001$, Odds Ratio (Correctness) = 1.844), indicating that success in ASSISTments lead to higher probability of attending college. The same strong positive relationship is seen between college enrollment and student knowledge estimate as the student learns with ASSISTments (CollegeEnrollment = 0.543 Student Knowledge + 0.345, $\chi^2(df = 1, N = 3747) = 236.683$, $p < 0.001$, Odds Ratio (StudentKnowledge) = 1.722). Engaged Concentration is also shown to positively predict college attendance (CollegeEnrollment = 0.403 Engaged Concentration + 0.325, $\chi^2(df = 1, N = 3747) = 140.557$, $p < 0.001$, Odds Ratio (Engaged Concentration) = 1.497), a finding that supports studies relating this affective state to effective learning [21, 42]. And the more a student uses ASSISTments, the more likely that student will attend college (CollegeEnrollment = 0.321 Number of First Actions + 0.327, $\chi^2(df = 1, N = 3747) = 79.159$, $p < 0.001$, Odds Ratio(Number of First Actions) = 1.378). One non-intuitive relationship is between carelessness and college enrollment. Taken by itself, the more a student becomes careless or commits more slips, the more likely the student is to attend college (CollegeEnrollment = 0.477 Slip/Carelessness + 0.338, $\chi^2(df = 1, N = 3747) = 185.208$, $p < 0.001$, Odds Ratio(Slip/Carelessness) = 1.612), evidence in keeping with past results that careless errors are seen in more successful students [16].

Conversely, the more a student is bored, the less likely that student is to attend college (CollegeEnrollment = -0.197 Boredom + 0.318, $\chi^2(df = 1, N = 3747) = 35.387$, $p < 0.001$, Odds Ratio(Boredom) = 0.821) a result in keeping with past evidence that boredom is associated with poorer learning [38], as well as high school dropout [25, 34, 43]. Confusion also is shown to be negatively associated with eventual college enrollment (CollegeEnrollment = -0.188 Confusion + 0.317, $\chi^2(df = 1, N = 3747) = 32.051$, $p < 0.001$, Odds Ratio(Confusion) = 0.829). Gaming the system is also negatively correlated with eventual college enrollment (CollegeEnrollment = -0.313 Gaming + 0.314, $\chi^2(df = 1, N = 3747) = 78.821$, $p < 0.001$, Odds Ratio (Gaming) = 0.731), perhaps unsurprising given its relationship with poorer learning [17].

A model for college enrollment including all data features was developed using Logistic Regression, and cross-validated at the student level (5-fold). The full data set model (Table 2) which included all features achieved a cross-validated A' of 0.686 and cross-validated Kappa value of 0.239. This model was statistically significantly better than a null (intercept-only) model, $\chi^2(df = 9, N = 3747) = 390.146$, $p < 0.001$. Statistical significance was computed for a non-cross-validated model, as is standard practice.

**Table 2. Full Data Set Model of College Enrollment**

| Features | Coefficient | Chi-Square | p-value | Odds Ratio |
|---|---|---|---|---|
| Student Knowledge | 1.078 | 16.193 | <0.001 | 2.937 |
| Slip/ Carelessness | -1.100 | 25.873 | <0.001 | 0.333 |
| Correctness | 0.758 | 33.943 | <0.001 | 2.133 |
| Boredom | 0.069 | 0.308 | 0.579 | 1.071 |
| Engaged Concentration | -0.175 | 2.207 | 0.137 | 0.839 |
| Confusion | 0.201 | 20.261 | <0.001 | 1.223 |
| Off-Task | -0.036 | 0.188 | 0.665 | 0.965 |
| Gaming | -0.047 | 0.720 | 0.396 | 0.954 |
| Number of First Actions | 0.269 | 27.094 | <0.001 | 1.308 |
| Constant | 0.354 | 99.735 | <0.001 | 1.421 |

This model can be refined by removing all features that are not statistically significant, using a backwards elimination procedure. Our final model (Table 3) achieves a cross-validated A' of 0.686 and a cross-validated Kappa value of 0.247, almost identical to the initial model with a full data set. The reduced model is both more parsimonious and more interpretable, so it is preferred. (It is not more generalizable within the initial data set, but its parsimony increases the probability that it will be generalizable to entirely new data sets). This model is also statistically significantly better than the null model, $\chi^2(df = 6, N = 3747) = 386.502$, $p < 0.001$. Our final model also achieved a fit of $R^2$ (Cox & Snell) = 0.098, $R^2$ (Nagelkerke) = 0.132, indicating that our predictors explaining 9.8% to 13.2% of the variance of those who attended college. Note that for our models, our $R^2$ values serve as measures of effect sizes; when converted to correlations, they represent moderate effect sizes in the 0.31-0.36 range.

**Table 3. Final Model of College Enrollment**

| Features | Coefficient | Chi-Square | p-value | Odds Ratio |
|---|---|---|---|---|
| Student Knowledge | 1.119 | 17.696 | <0.001 | 3.062 |
| Correctness | 0.698 | 47.352 | <0.001 | 2.010 |
| Number of First Actions | 0.261 | 28.740 | <0.001 | 1.298 |
| Slip/ Carelessness | -1.145 | 28.712 | <0.001 | 0.318 |
| Confusion | 0.217 | 24.803 | <0.001 | 1.242 |
| Boredom | 0.169 | 12.249 | <0.001 | 1.184 |
| Constant | 0.351 | 100.011 | <0.001 | 1.420 |

As can be seen in Table 3, the first three predictors (student knowledge, correctness and number of first actions) maintained the same directionality as in Table 1, but slip/carelessness, confusion and boredom flipped direction when incorporated into the final multiple logistic regression model, though each remained significant. For example, in this model, the likelihood of college enrollment increases with boredom, once the other variables are taken into account (e.g. once we control for student knowledge, software use, and so on, and other forms of disengagement). This may be because once we remove unsuccessful bored students, all that may remain are students who become bored because the material is too easy [cf. 37]. Similarly, once we control for other variables in the model, confusion is positively associated with college attendance. Again, once we remove students who are both confused and unsuccessful, all that is likely to remain are students who addressed their confusion productively [cf. 30]. For carelessness, once we control for other variables in the model, it is negatively associated with college attendance. Once we remove careless but successful students, all that is likely to remain are students who haven't overcome their carelessness [cf. 16].

# 4. DISCUSSION AND CONCLUSION

Many factors influence a student's decision to enroll in college. A lot of them external or social factors: financial reasons, parental support and school support. Another major factor, however, is one's ability and engagement, which develop over early years, and begin to manifest strongly during the middle school years. In this paper, we apply fine-grained models of student knowledge, student affect (boredom, engaged concentration, confusion) and behavior (off-task, gaming, slip/carelessness) to data from 3,747 students using educational software over the course of a year (or more) of middle school to understand how the development of student learning and engagement during this phase of learning, can predict college enrollment. A logistic regression model is developed, and we find that a combination of features of student engagement and student success in ASSISTments can distinguish a student who will enroll in college 68.6% of the time. In particular, boredom, confusion, and slip/carelessness are significant predictors of college enrollment both by themselves and contribute to the overall model of college enrollment.

The relationships seen between boredom and college enrollment, and gaming the system and college enrollment indicate that relatively weak indicators of disengagement are associated with lower probability of college enrollment. Success within middle school mathematics (indicated by correct answers and high probability of knowledge in ASSISTments) is positively associated with college enrollment , a finding that aligns with studies that conceptualize high performance during schooling as a sign of college readiness [41] and models that suggest that developing aptitude predicts college attendance [15, 23].

Findings in our data and final model support existing theories about indicators of college enrollment (academic achievement, grades). More importantly, it further sheds light on behavioral factors the student experiences in classrooms (which are more frequently and in many ways more actionable than the behaviors which result in disciplinary referrals). As the results here suggest, affect and disengagement are associated with college enrollment, suggesting that in-the-moment interventions provided by software (or suggested by software to teachers) may have unexpectedly large effects, if they address negative affect and disengagement. Confused students can be properly guided and encouraged to resolve their confusion. Bored students can be provided with greater novelty to reduce boredom or support in emotional self-regulation. Students who game the system can be given alternate opportunities to learn material bypassed through gaming, as in past successful interventions. Further work can be explored in the interactions of these various factors which influence our predictions.

Future endeavors in evaluating college attendance through data mining of interaction logs (pre-college) can further benefit from including additional possible interaction features in our model. Other machine learning algorithms or modeling can also be employed in our data in further understanding our research problem. It is possible that other classifiers, such as decision trees or support vector machines, may have performed better in predicting college enrollment. However, interpretability of the models may be reduced for these algorithms. Together with findings in this paper, further design considerations for educational software can be investigated that can influence not only effective learning during secondary education, but contribute as well to college interest and readiness.

## 5. ACKNOWLEDGMENTS

This research was supported by grants NSF #DRL-1031398, NSF #SBE-0836012, and grant #OPP1048577 from the Bill & Melinda Gates Foundation. We also thank Zak Rogoff, Adam Nakama, Aatish Salvi, and Sue Donas for their assistance in conducting the field observations, and Adam Goldstein for help in data processing.

## 6. REFERENCES

[1] Baker, R.S.J.d. 2007. Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems. In *Proceedings of ACM CHI 2007: Computer-Human Interaction*, 1059-1068.

[2] Baker, R.S.J.d., Corbett, A.T., and Aleven, V. 2008. More Accurate Student Modeling through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (eds Aimeur E. & Woolf B.), Springer Verlag, Berlin, 406-415.

[3] Baker R.S.J.d., Corbett A.T., Gowda S.M., Wagner A.Z., MacLaren B.M., Kauffman L.R., Mitchell A.P., and Giguere S. 2010. Contextual Slip and Prediction of Student Performance after Use of an Intelligent Tutor. In *Proc. UMAP 2010*, 52-63.

[4] Baker, R.S., Corbett, A.T., and Koedinger, K.R. 2004. Detecting Student Misuse of Intelligent Tutoring Systems. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, 531-540.

[5] Baker, R.S.J.d., Goldstein, A.B., and Heffernan, N.T. 2011. Detecting Learning Moment-by-Moment. *International Journal of Artificial Intelligence in Education*, 21, 5-25.

[6] Baker, R.S.J.d., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J., and Rossi, L. 2012. Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. In *Proc. EDM 2012*, 126-133.

[7] Balfanz, R., Herzog, L., and Mac Iver, D. 2007. Preventing Student Disengagement and Keeping Students on the Graduation Path in the Urban Middle Grade Schools: Early Identification and Effective Interventions. *Educational Psychologist*, 42(4), 223–235.

[8] Balfanz, R. 2009. *Putting middle grades students on the graduation path: A policy and practice brief*. Baltimore, MD: Everyone Graduates Center & Talent Development Middle Grades Program, Johns Hopkins University.

[9] Bowers, A.J. 2010. Analyzing the Longitudinal K-12 Grading Histories of Entire Cohorts of Students: Grades, Data Driven Decision Making, Dropping Out and Hierarchical Cluster Analysis. *Practical Assessment, Research & Evaluation (PARE)*, 15(7), 1-18.

[10] Bowers, A.J. 2010. Grades and Graduation: A Longitudinal Risk Perspective to Identify Student Dropouts. *The Journal of Educational Research*, 103(3), 191-207.

[11] Bowers, A.J., Sprott, R., and Taff, S.A. 2013. Do we Know Who Will Drop Out? A Review of the Predictors of Dropping out of High School: Precision, Sensitivity and Specificity. *The High School Journal*. 96(2), 77-100.

[12] Cabrera, A. F. 1994. Logistic regression analysis in higher education: An applied perspective. *Higher Education: Handbook of Theory and Research*, Vol. 10, 225-256.

[13] Cabrera, A. F., La Nasa, S. M. and Burkum, K, R. 2001. *Pathways to a Four-Year Degree: The Higher Education Story of One Generation*. Center for the Study of Higher Education. Penn State University.

[14] Camblin, S. 2003. The middle grades: Putting all students on track for college. Honolulu, HI: *Pacific Resources for Education and Learning*.

[15] Christensen, J. Melder, and B. Weisbrod. 1975. Factors affecting college attendance. *Journal of Human Resources*, 10 (1975), pp. 174–188

[16] Clements, K. 1982. Careless errors made by sixth-grade children on written mathematical tasks. *Journal for Research in Mathematics Education*, 13, 136–144.

[17] Cocea, M., Hershkovitz, A., and Baker, R.S.J.d. 2009. The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate? In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 507-514

[18] Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1, 37-46.

[19] Conati, C., and Maclaren, H. 2009. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction* 19, 3, 267-303.

[20] Corbett, A.T., and Anderson, J.R. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction* 4, 4, 253-278.

[21] Craig, S.D., Graesser, A.C., Sullins, J., and Gholson, B. 2004. Affect and learning: an exploratory look into the role

of affect in learning with autotutor. *Journal of Educational Media*, 29, 241-250.

[22] D'Mello, S.K., Craig, S.D., Witherspoon, A. W., McDaniel, B. T., and Graesser, A. C. 2008. Automatic Detection of Learner's Affect from Conversational Cues. *User Modeling and User- Adapted Interaction*, 18 (1-2), pp. 45-80.

[23] Eccles, J. S., Vida, M. N., and Barber, B. 2004. The relation of early adolescents' college plans and both academic ability and task-value beliefs to subsequent college enrollment. *Journal of Early Adolescence*, 24, 63-77.

[24] Efron, B. and Gong, G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, 37, 36–48.

[25] Farrell, Edwin. 1988. Giving Voice to High School Students: Pressure and Boredom, Ya Know What I'm Sayin'? *American Educational Research Journal* 4, 489-502.

[26] Feng, M., Heffernan, N.T., and Koedinger, K.R. 2009. Addressing the assessment challenge in an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research (UMUAI). 19*(3), pp. 243-266.

[27] Fredricks, J. A., Blumenfeld, P. C., and Paris, A. H. 2004. School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74, 59–109.

[28] Hanley, J., and McNeil, B. 1982. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* 143, 29-36.

[29] Lau, S. and Liem, A.D. 2007. The role of self-efficacy, task value, and achievement goals in predicting learning strategies, task disengagement, peer relationship, and achievement outcome. *Contemporary Educational Psychology*, 3, 1-26.

[30] Lee, D.M., Rodrigo, M.M., Baker, R.S.J.d., Sugay, J., and Coronel, A. 2011. Exploring the Relationship Between Novice Programmer Confusion and Achievement. In *Proceedings of the 4th bi-annual International Conference on Affective Computing and Intelligent Interaction, 2011*.

[31] Lent, R. W., Brown, S. D., and Hackett, G. 1994. Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior*, 45, 79-122.

[32] Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M. and Euler, T. 2006. YALE: rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (eds Ungar L., Craven M., Gunopulos D. & Eliassi-Rad T.), ACM, New York, 935-940.

[33] National Middle School Association research summary #12: Academic achievement. 2002. Retrieved January 14, 2003, from www.nmsa.org/research/ressum12.htm

[34] National Research Council & Institute of Medicine. 2004. *Engaging schools: Fostering high school students' motivation to learn*. Washington, DC: National Academy Press.

[35] Neild, R. C.2009. Falling Off Track during the Transition to High School: What We Know and What Can Be Done. *The Future of Children* 19(1), 53-76. Princeton University.

[36] Nunez, A.M. and Bowers, A.J. 2011. Exploring What Leads High School Students to Enroll in Hispanic-Serving Institutions: A multilevel analysis. *American Educational Research Journal*, 48(6), 1286-1313.

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

[38] Pekrun, R., Goetz, T., Daniels, L.M., Stupnisky, R.H. and Perry, R.P. 2010. Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology* 102, 531-549.

[39] Peng, C.-Y. J., Lee, K. L., and Ingersoll, G. M. 2002. An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, 96(1), 3-14.

[40] Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N.T., Koedinger, K. R., Junker, B., Ritter, S., Knight, A., Aniszczyk, C., Choksey, S., Livak, T., Mercado, E., Turner, T.E., Upalekar. R, Walonoski, J.A., Macasek. M.A. and Rasmussen, K.P. 2005. The Assistment project: Blending assessment and assisting. In *Proc. AIED 2005*, 555-562.

[41] Roderick, M., Nagaoka, J., and Coca, V. 2009. College readiness for all: The challenge for urban high schools. *The Future of Children* 19(1): 185-210.

[42] Rodrigo, M.M.T., Baker, R.S., Jadud, M.C., Amarra, A.C.M., Dy, T., Espejo-Lahoz, M.B.V., Lim, S.A.L., Pascua, S.A.M.S., Sugay, J.O., and Tabanao, E.S. 2009. Affective and Behavioral Predictors of Novice Programmer Achievement. In *Proc. ACM-SIGCSE 2009*, 156-160.

[43] Rumberger, R. W. 1987. High school dropouts: A review of issues and evidence. *Review of Educational Research*, 57, 101-121

[44] Sabourin, J., Mott, B., and Lester, J. 2011. Modeling Learner Affect with Theoretically Grounded Dynamic Bayesian Networks. In *Proc. ACII 2011*, 286-295.

[45] San Pedro, M.O.C., Baker, R., and Rodrigo, M.M. 2011. Detecting Carelessness through Contextual Estimation of Slip Probabilities among Students Using an Intelligent Tutor for Mathematics. In *Proceedings of 15th International Conference on Artificial Intelligence in Education*, 304-311

[46] Stephan, J. and Rosenbaum, J. 2012. Can High Schools Reduce College Enrollment Gaps with a New Counseling Model? *Educational Evaluation and Policy Analysis*, first published on October 16, 2012. DOI: 10.3102/0162373712462624.

[47] Tobin, T. J. and Sugai, G. M. 1999. Using sixth-grade school records to predict school violence, chronic discipline problems, and high school outcomes. *Journal of Emotional and Behavioral Disorders*, 7, 40-53.