Predicting Student Engagement in Intelligent Tutoring Systems Using Teacher Expert Knowledge

Nicholas M. LLOYD [a], Neil T. HEFFERNAN [a] and Carolina RUIZ [a]

[a] *Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA USA*

Abstract. Detection and prevention of off-task student behavior in an Intelligent Tutoring System (ITS) has gained a significant amount of attention in recent years. Previous work in these areas have shown some success and improvement. However, the research has largely ignored the incorporation of the expert on student behavior in the classroom: the teacher. The main goal of this project was to predict student engagement, both positive behavior and negative (gaming) behavior, within the *Assistments* system using teacher observations of student behavior in the tutoring classroom. Using a dataset incorporating attributes associated with previous findings in gaming detection research we developed two logistic regression models using stepwise regression to predict positive engagement as well as gaming behavior. Gaming detection models proved unsuccessful, however positive engagement prediction shows promising results with prediction accuracy at 71.1%. To our knowledge this is the first investigation that has shown that we can detect high levels of student engagement, thus paving the way for more accurate ways of providing positive feedback to students.

Introduction

The effectiveness of an Intelligent Tutoring System (ITS) can be undermined by students who are not engaged in the learning activity. Recent research into disengaged behavior, most notably gaming behavior where a student is exploiting the available help and feedback provided by an ITS, has shown that there is a correlation between such behavior and reduced learning [6]. Developed methods of detecting gaming have shown some success [4,5] along with studies directed towards classifying, measuring, and modeling a wider range of student engagement and disengagement as well as emotional states and attitudes with an ITS or other computer learning environment [2,11,8,9,10]. Detecting gaming behavior within the *Assistments* system, an Intelligent Tutoring System that has been developed jointly between Worcester Polytechnic Institute (WPI) and Carnegie Mellon University (CMU) [16], has proved to be less successful [19]. However, all of this research has largely ignored the expert of student behavior in the classroom: the teacher. Teachers have long been seen as the most knowledgeable of their students' behaviors in the classroom, and this has been acknowledged by some in the ITS community [9,18]. Teachers also have a direct influence on their students' engagement patterns within the classroom [17,12]. Additionally, work to determine why students game came to the conclusion that

Table 1. Student Focus Grades

1 = Very focused, model student
2 = Focused, appears to be working
3 = Unsure
4 = Unfocused, I don't think they are making an effort
5 = Very unfocused, they are messing around/not paying attention

students who are not good at math and get frustrated are typically the students who game [7]. However that study left open the question of why some students show very high persistence.

Following previous work on detecting gaming within the *Assistments* system [19], the goal of this research was to explore the use of teacher observations of student engagement within the *Assistments* system in developing predictive models of student engagement. Instead of asking the teachers to provide a simple rating of gaming or not-gaming, teachers were provided with a simple 5 value grading scheme as a code for identifying highly engaged students to moderately engaged students to highly disengaged/gaming students. This method was chosen in contrast to prior studies that focus on either gaming or not-gaming [5,4,19] in order to provide the possibility of detecting good behavior as well as bad behavior. . Positive feedback is important to student learning and detecting only the presence or absence of gaming behavior does not provide an effective means for rewarding and encouraging good engagement behavior.

Methods

Classroom Observation

Teachers being the primary source of data, it was important to be as unobtrusive as possible as well as to be as straightforward as possible in order to acquire data that accurately represents how a teacher typically monitors the behavior of their students during tutoring sessions. These issues required a different approach to the data collection process than has previously been employed [6,19].

At the beginning of every tutoring session with a new teacher, the observer was instructed to briefly explain the purpose of the study and what would be required of the teacher. The actual observations consisted of the observer shadowing the teacher for the span of the tutoring session as the teacher went about their way as they would during any tutoring session. As the teacher monitored the behavior of his/her students, they were instructed to grade the current activity of the students they were watching using a coding scheme for student engagement. Table 1 describes this coding scheme. Note that the terms "focus" and "engagement" may be used interchangeably throughout this article, however the meaning is equivalent in this domain[1].

As the teacher provided the grades for the different students, the observer would record the grade on a table associating the grade with the prerecorded student's user

[1] The term "focus" was chosen in contrast to "engagement" or "effort" in order to emphasize a more positive intention behind the teacher-given grades. Stating that a student is less focused is not as demoralizing as saying a student is putting forth less effort. This term posed no difficulty in interpretation for the teachers observed.

name as well as the minute in time of the observation. The data recording tables were constructed such that the starting hour of the period is recorded in an area at the top of the document and the column labels represent the minutes following this start period.

The observation period for this research spanned the month of March, 2007. During this period a total of 7 classes of approximately 20 students per class. Over all of these classes 3 different teachers were observed. In any given class a teacher typically yielded 2–3 observations per student over an hour long period with grade differences between observations rarely exceeding one. The final dataset yielded 265 teacher given grades for the analysis[2]. Workload for the observers was found to be approximately equivalent to that of previous studies [6,19].

Dataset Creation

The *Assistments* system maintains a record of a running dialog between the student and the tutor detailing every student action and tutor response[3]. The recorded observation times from the data collection phase of this research were used to determine what problem or problems the student was working on in the minute long time window when the observation was made. This tactic was chosen so as to characterize the relationship of engagement behavior with not only student behavior but student behavior with particular problems.

The data collected for the dataset focuses on several aspects of student activity within an ITS that have been identified as important in identifying and measuring student engagement. These aspects can be broken down into data relating to the student and data relating to the problem the student is currently working on. A student's prior knowledge of the given material is strongly correlated with their engagement patterns and, in particular, whether or not they will engage in gaming behavior [17,6,9,10,19]. However, performance alone is not an effective predictor of engagement since not all students who engage in gaming behavior are hurt by it, as demonstrated by high performing students who are among the gaming students [4]. Additionally, how quickly and in what ways a student interacts with the system are important predictors of student engagement [6,10,19]. As for problem related data, the measure of a problem's difficulty has been noted to correlate with engagement patterns [10,18] as well as the type of problem being presented, in other words multiple choice problem or short answer problem [6,19].

The dataset attributes are clustered into two distinct groups: 1) Observation period - indicating that they are directly related to the observation period, and 2) General - indicating that they are attributes that are not directly related to the observation period. Descriptions of these attributes are as follows:

Focus grade - The teacher-given grade for the student at a certain time, which will constitute the dependent variable for the analysis.

Observation period

First action - This value indicates what action the student performed after being shown the problem associated by time with the focus grade. This value consists of three possible values: 1) attempt - indicating that the student made an attempt to answer the problem 2) hint - indicating that the student requested a hint and 3) bottomHint -

[2] For more information regarding the observation process of this study see citation [13].

[3] At this time the system does not record as detailed information as mouse movement in the tutor environment.

indicating that the student requested a bottom out hint (in other words the answer to the problem). It is important to note that students do not know if the next hint they receive will be a bottom out hint, however some students who are disengaged from the learning activity will purposefully seek bottom out hints. Additionally, although this is not common, some problems only contain bottom out hints thus indicating why this is presented as a possible value for the first action variable. For the analysis these string values are encoded into the integer values 1, 2, and 3 respectively.

First action time - A millisecond value representing how long the student took to respond to the problem with the first action.

Second action - Similar to the first action except this value has the additional possibility of being empty if the first action performed was a correct answer. Empty values are encoded in the analysis as a 0 integer.

Second action time - A millisecond value representing how long the student took to respond to the results of the first action for the given problem, provided that the second action is not empty indicating that the student answered the problem correctly.

Student attempts this problem - A count of the number of attempts the student has made to answer this problem at this particular observation time.

Attempts this problem z-score - A standardized measure of the number of attempts the student has made on this problem based upon how students typically respond to this problem. The z-score[4] is calculated by taking the mean and standard deviation of the number of attempts made to answer this particular problem over all available data in the database. The mean is subtracted from the number of attempts the student has made on this problem at this time and the result is divided by the standard deviation.

Student hints this problem - A count of the number of hints requested by the student for this problem.

Hints this problem z-score - Same procedure as the "attempts this problem z-score" value except that the number of hints requested for the problem is under consideration.

Student bottom hint this problem - A "count" of the number of bottom hints requested by the student for this problem. This is a value of 0 or 1 since there is only ever one bottom out hint for any given problem.

General

Poor man's prior knowledge - A measure of the student's preceding performance. Although there is current research in the *Assistments* system that uses a student's performance as related to skills associated with different problems at varying levels of granularity [15], this does not happen live and not all problems in the *Assistments* system are associated with distinct math skills. As a result, the student's performance before each observation was estimated in a similar manner to the work by Walonoski and Heffernan [19], where the percent correct of the student's previous work is calculated.

Problem type - It is important to note that a student's interaction with the system is largely dependent upon the type of input that is required of the student to answer the problem. This input is broken into two separate categories: multiple-choice and short answer. An example of how a student will interact differently lies in the observation that a student may simply "guess and check" their way through a multiple choice problem since there are a limited number of answer options presented, whereas a short answer problem is less conducive to this behavior. In contrast to this a student presented with a short answer problem is arguably more likely to follow "help abuse" gaming patterns.

[4] Z-score is a statistical method of standardizing an observation value with respect to the properties of the population [http://en.wikipedia.org/wiki/Z-score].

Problem difficulty - This variable is a simple measure of a given problem's difficulty based upon all data available in the *Assistments* system database before the observation period. This data goes back to the year 2004. This value is a percentage of the number of times this problem was answered *incorrectly* in a problem set.

Analysis and Results

The first step in our analysis was to evaluate the distribution of focus grades in order to look for any potential bias from the teachers towards a particular focus grade. Figure 1 shows the overall distribution of grades in the dataset with the notably strong trend towards the more positive focus grades (see figure 1). The particular trend pattern displayed in this figure is equivalent for each of the three teachers with the key difference being a stronger trend towards more of the positive focus grades and less of the negative focus grades for the advanced level students [5]. This indicates that teachers tend to rate their students as being highly engaged, in this case 52.8% of the dataset, whereas negative behavior such as gaming constitutes only 6% of the data (focus grade 5). It should be noted that gaming behavior has been observed to be just as infrequent in previous studies of such activity in the *Assistments* system [19].

For the development of predictive models we chose stepwise binary logistic regression using the Likelihood Ratio Forward Selection procedure as available in the SPSS Statistical Software Suite which was used for this analysis[6]. Initial analyses of the data using linear and multinomial regression techniques provided disappointing results [13]. However, these previous analyses were focused on developing models for the entire range of focus grades rather than attempting to predict focus grades of particular importance. Following this, two binary logistic regression models were developed to predict the presence or absence of the two extremes of the focus grade range: highly engaged level 1 students and highly disengaged level 5 students. Considering the variety of categorical and

[5] These particular students were members of an advanced level mathematics curriculum. The students had to meet certain academic standards and formally apply to get into this particular class.

[6] The Likelihood Ratio forward selection method is a stepwise selection method where forward entry of variables into the model is based upon the significance of the score statistic while removal is tested using likelihood-ratio statistic and maximum partial likelihood estimates. For more information see the SPSS Logistic Regression documentation.

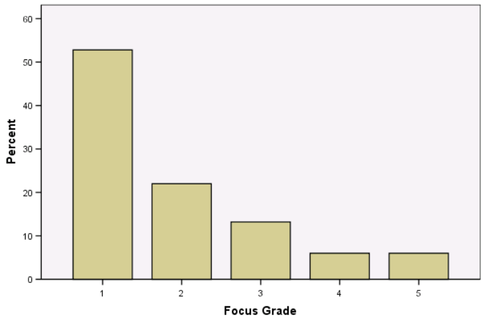

Figure 1. Distribution of Focus Grades in the Dataset

numeric data logistic regression was deemed most appropriate for this analysis. Following this two new dependent variables were added to the dataset, one representing “Great Effort” and one representing “Gamer.” Both variables are binary representations of the focus grades with “Great Effort” having a value of 1 if the associated focus grade is 1 and 0 if otherwise. The “Gamer” value is just the opposite with 1 representing a focus grade of 5 and 0 indicating all other focus grades.

The first model produced was to predict “Great Effort” for a student based upon their activity on a given problem. Table 2 shows the attributes determined to be statistically significant to the model along with their coefficients. The attributes, and their coefficients, determined to be most significant to the model were the “first action type” and “pmp” (Poor man’s prior knowledge) values. The “first action type” value was translated from its original form as being either an “attempt” or a “hint” to a value of either 1 or 0 respectively. Both have a positive correlation indicating that if the first action performed was an attempt and the student has a higher performance, then there is a greater likelihood that they are putting forth “Great Effort.”

Classification tests on the dataset, where the original dataset is classified by the model, were run to determine the accuracy of the predictions. Table 3 shows the test results for the “Great Effort” model. The overall percent correct shows that 71.1% of the time the predictions are correct. This is significantly better than simply guessing which, considering the quantity of focus grade 1, would have a 50% probability of success.

The second model produced was to predict whether or not a student was a “Gamer” based upon the teacher given grades. Table 4 shows the attribute and coefficient for the model. It is surprising that there is only one attribute and coefficient that have been selected for this model, “first action type,” however this does coincide with the “Great Effort” model. Additionally, the coefficient is almost the exact reverse of the coefficient for the “Great Effort” model for the same attribute. This makes sense considering that we are trying to predict what essentially is the polar opposite from the other model, that being gaming behavior.

Table 2. Selected Attributes for the “Great Effort” Model

Attribute	Coefficient	S.E.
Poor man’s prior knowledge	2.704	0.547
First action type (attempt)	1.932	0.429

Table 3. Great Effort Model Classification Test Results

Great Effort	Percent Correct
False	63.2%
True	78.0%
Overall	71.1%

Table 4. Selected Attribute for the “Gamer” Model

Attribute	Coefficient	S.E.
First action type (attempt)	-1.416	0.545

Table 5. Gamer Model Classification Test Results

Gamer	Percent Correct
False	100.0%
True	0.0%
Overall	94.0%

The results from the classification test on the "Gamer" model were disappointing. Table 5 shows that although the overall percent correct was a strong 94%, the low percentage of focus grade 5 rows in the dataset (6%) proved to be an insurmountable challenge for the model which was unable to accurately predict any of the focus grade 5 values correctly. The negative values for the constant and the coefficient in this model indicate that no matter what this model will predict that a student is not gaming.

Considering the significance of the poor man's prior knowledge attribute in predicting "Great Effort," it was important to re-analyze the data from the perspective of different student performance groups. The reasons for this are: 1) High performing students, though less likely to game, still do game [4] and 2) Identifying students who put forth great effort based in part on their performance does not account for students who may have low prior knowledge but are still putting forth a significant amount of effort. The dataset was subdivided into three groups based upon distinct ranges of the prior knowledge attribute. These ranges were: low performing students with prior knowledge <= 0.33, mid-range performing students with prior knowledge > 0.33 and <= 0.66, and high performing students with prior knowledge > 0.66. The distributions of these groups in the dataset are 29.2%, 41.2%, and 29.6%.

Figure 2 displays the percentage of correctly predicted Great Effort values using the Great Effort Model on each of the prior knowledge subgroups. The x-axis shows numeric labels for each subgroup with 0 representing the low performers, 1 representing the mid-range performers, and 2 representing the high performers. Though there is a notable increase in the accuracy of predicting great or not great effort in the high performance group, all groups still have a prediction accuracy level around 70%. The consistence of the prediction accuracy across these groups shows that the model does not exclusively consider all high performers to have excellent effort and at the same time low performers are not all identified as having poor effort.

Conclusion and Discussion

The results from the evaluations of the produced models indicate that the data collected on teacher grades can to a certain degree of accuracy predict whether or not a student is strongly engaged in the learning activity with the *Assistments* tutor. Although 71.1% is still not an ideal accuracy level, it provides a starting point for future work using teacher data to develop models of student engagement. In addition, since this work shows that there is potential for predicting positive engagement, there is by extension the potential for generating reward systems to go along with the gaming prevention systems that have been developed in past research [14,3,20]. However, the prediction accuracy across low, medium, and high performing students is consistent for the model suggesting that even though prior performance is a factor the model does not distinguish high performing students as being exlusively the positively engaged students. This is important considering that as much as low performing students are most likely to be frustrated and, therefore, to game [7], the performance level of a student does not preclude them from being persistent in their learning effort. What is important for a model of this nature is to identify when a student is putting forth great effort regardless of their prior performance in order to encourage the students who are doing well, as well as the students who are having the most difficulty with the subject material.

Unfortunately the prediction of gaming behavior by our model was ineffective. However, this is hardly surprising considering the low percentage of gaming instances in our dataset (6%). This difficulty has been noted in a previous study on gaming behavior in the *Assistments* system [19]. In that study the detection of non-gaming instances by their model was 98% accurate, however it is important to note that this does not indicate positive engagement behavior in the student. Evaluation of other statistical models of the data provided less effective results, however the correlation of teacher given grades with a graphical reporting tool on student engagement provided positive results [13]. Other regression models of the data splitting the focus grade groups in different ways, 1-2 vs 3-5 for instance, were ineffective.

Although this model is not accurate enough to be implemented in the *Assistments* system, since 26.67% of students who were identified as gaming were mislabeled by the model as having positive engagement, the determined accuracy is still better than guessing and the prediction accuracy is consistent across all performance levels. This shows that it is possible to predict positive engagement in an Intelligent Tutoring System, and that teacher data can be used to develop these predictive models. While models designed to detect gaming behavior can institute corrective action, either by active [14] or passive [3,20] techniques[7], more accurate models predicting positive engagement would allow for a reward system for the student potentially bringing positive reinforcement of desired learning behaviors.

[7] Active gaming prevention techniques directly alter the learning environment in order to prevent gaming while passive techniques offer unobtrusive yet highly visible feedback on student behavior.

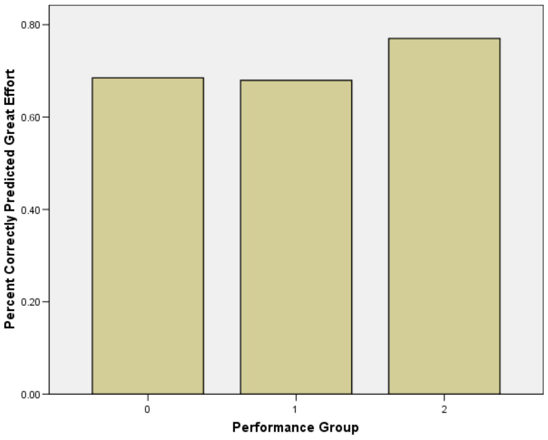

Figure 2. Great Effort Model Prediction Accuracy for Low (0), Mid-Range (1), and High (2) Performing Students

Future Work

The results in this study provide a benchmark for future research into the detection of positive engagement in an Intelligent Tutoring System as well as the incorporation of teacher expert knowledge in the development of such models. This study was instituted as an exploratory analysis in using teacher data in gaming and engagement detection, and it is clear from these initial results that there is potential for models developed from teacher given data for predicting student positive engagement. More data from a larger group of teachers could provide better models for these predictions. Additionally, a larger attribute set incorporating a wider time window to evaluate not only the problem they were working on as in this study but also the subsequent problems and possibly an average between them could improve the results of this study.

While we would like to have messages be displayed to the student along the lines of "You seem to be really focused on the assignment! Good job!" models that more accurately predict when students are positively engaged are necessary. It is important to note that misidentification of students who are engaged as being not engaged is not as problematic as rewarding students who are not engaged, or even gaming, for having positive engagement. These false positives would have the opposite of the desired effect and could convince the gaming students that their actions will go undetected.

Another question that has yet to be answered is which source of data is better at predicting student behavior patterns: teachers or outside observers. This study relied on data exclusively from the teachers, however a comparison of separate models produced by the two data sources could provide a clue as to where better data can be gathered for the development of behavior detection models in Intelligent Tutoring Systems.

Acknowledgements

This research was made possible by a large number of people listed at www.ASSISTment.org. This research was also made possible by the US Dept of Education, Institute of Education Science, "Effective Mathematics Education Research" program grant #R305K03140, the Office of Naval Research grant #N00014-03-1-0221, NSF CAREER award (REC-0448319) to Neil Heffernan, and the Spencer Foundation. All of the opinions in this article are those of the authors, and not necessarily of the funders.

References

[1] V. Aleven and K.R. Koedinger, Limitations of student control: Do students know when they need help?, In *Proceedings of the 5th International Conference on Intelligent Tutoring Systems*, Montreal, Canada, (2000), 292–303.

[2] I. Arroyo, T. Murray, and B.P. Woolf, Inferring unobservable learning variables from studentsâĂŹ help seeking behavior, In *Proceedings of the Workshop on Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes, at the 7th International Conference on Intelligent Tutoring Systems*, MaceiÃş, Alagoas, Brazil, (2004), 29–38.

[3] R.S. Baker, A.T. Corbett, K.R. Koedinger, S. Evenson, I. Roll, A.Z. Wagner, M. Naim, J. Raspat, D.J. Baker, and J.E. Beck, Adapting to When Students Game an Intelligent Tutoring System, In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, Jhongli, Taiwan, (2006), 392–401.

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

[5] R.S. Baker, A.T. Corbett, K.R. Koedinger, and I. Roll, Generalizing Detection of Gaming the System Across a Tutoring Curriculum, In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, Jhongli, Taiwan, (2006), 402–411.

[6] R.S. Baker, A.T. Corbett, K.R. Koedinger, and A.Z. Wagner, Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game the System", In *Proceedings of ACM CHI 2004: Computer-Human Interaction*, Vienna, Austria, (2004), 383–390.

[7] R.S. Baker, J.A. Walonoski, N.T. Heffernan, I. Roll, A.T. Corbett, and K.R. Koedinger (In press), Why Students Engage in "Gaming the System" Behavior in Interactive Learning Environments, *Journal of Interactive Learning Research (JILR)*, Chesapeake, VA: AACE.

[8] R.L. Bangert-Drowns and C. Pyke, A Taxonomy of Student Engagement with Educational Software: An Exploration of Literate Thinking with Electronic Text, *Journal of Educational Computing Research* **24**(3) (2001), 213–234.

[9] C.R. Beal, L. Qu, and H. Lee, Classifying Learner Engagement Through Integration of Multiple Data Sources, In *Proceedings of the 21st National Conference on Artificial Intelligence*, Menlo Park, California (2006), 2–8, AAAI Press.

[10] J.E. Beck, Engagement tracing: using response times to model student disengagement, In *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED)*, Amsterdam, The Netherlands, (2005), 88–95.

[11] C. Conati, Probabilistic assessment of user's emotions in educational games, *Journal of Applied Artificial Intelligence* **16** (2002), 555–575.

[12] R.D. Goddard, W.K. Hoy, and A.W. Hoy, Collective Teacher Efficacy: Its Meaning, Measure, and Impact on Student Achievement, *American Educational Research Journal* **37**: 479–507, 2000.

[13] N.M. Lloyd, Measuring Student Engagement in an Intelligent Tutoring System, Worcester Polytechnic Institute, Technical Report WPI-CS-TR-07-04 (2007).

[14] R.C. Murray and K. VanLehn, Effects of Dissuading Unnecessary Help Requests While Providing Proactive Help, *Artificial Intelligence in Education* (2005), 887–889.

[15] Z.A. Pardos, N.T. Heffernan, B. Anderson, and C.L. Heffernan, Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks, *Workshop in Educational Data Mining held at the 8th International Conference on Intelligent Tutoring Systems* (2006).

[16] L. Razzaq, M. Feng, G. Nuzzo-Jones, N.T. Heffernan, K. Koedinger, B. Junker, S. Ritter, A. Knight, E. Mercado, T.E. Turner, R. Upalekar, J.A. Walonoski, M.A. Macasek, C. Aniszczyk, S. Choksey, T. Livak, and K. Rasmussen, The Assistment Project: Blending Assessment and Assisting, In *Proceedings of the 12th International Conference on Artificial Intelligence in Education*, Amsterdam, The Netherlands, (2005), 555–562.

[17] J.C. Turner, C. Midgley, K.K. Meyer, M. Green, E.M. Anderman, Y. Kang, and H. Patrick, The Classroom Environment and Students' Reports of Avoidance Strategies in Mathematics: A Multimethod Study, *Journal of Educational Psychology* **94**: 88–106, 2002.

[18] A. de Vicente and H. Pain, Informing the Detection of the Students' Motivational State: An Empirical Study, In *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, Biarritz, France, (2002), 933-943.

[19] J.A. Walonoski and N.T. Heffernan, Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems, In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, Jhongli, Taiwan, (2006), 382–391.

[20] J.A. Walonoski and N.T. Heffernan, Prevention of Off-Task Gaming Behavior in Intelligent Tutoring Systems, In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, Jhongli, Taiwan, (2006), 722–724.