Measuring Student Engagement in an Intelligent Tutoring System

by Nicholas M. Lloyd

A Thesis

Submitted to the Faculty

of

Worcester Polytechnic Institute

in partial fulfillment of the requirements for the

Degree of Master of Science

in

Computer Science

May 2007

Approved:

____
Professor Neil Heffernan
Thesis Advisor

____
Professor Carolina Ruiz
Thesis Reader

____
Professor Michael Gennert
Department Head

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. Our research re-evaluates the subjects of off-task behavior detection and prevention by developing metrics for student engagement in an ITS using teacher observations of student behavior in the classroom. We present an exploratory analysis of such metrics and the data gathered from the teachers. For off-task prevention we developed a visual reporting tool that displays a representation of a student's activity in an ITS as they progress and gives a valuable immediate report for the instructor.

Contents

List of Figures

1 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 [5]. Developed methods of detecting gaming have shown some success [2, 4] along with studies directed towards classifying and measuring a wider range of student engagement and disengagement with an ITS [6, 7]. Methods of preventing gaming have proven to be less successful either from students adapting to changed in a system resulting from prevention mechanisms [16] or by simply ignoring more passive prevention mechanisms [3]. Passive prevention mechanisms have also encountered difficulty in identifying what specifically is effective/not-effective about such techniques in the prevention of gaming behavior [3, 20]. 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 [6, 10]. Additionally, teachers have a direct influence on their students' engagement patterns within the classroom [19, 12].

The first objective of this research is to explore the feasibility of using teacher reports on student activity in the tutor classroom to develop models for measuring student engagement within the *Assistment* system, an Intelligent Tutoring System that has been developed jointly between Worcester Polytechnic Institute (WPI) and Carnegie Mellon University (CMU) [18]. A previous study on gaming behavior within the *Assistment* system provided some rules for detecting certain gaming behavior patterns, however actual machine learned models for detecting gaming proved to be unreliable [20]. Teacher evaluations of student behavior may provide a more successful model for detecting gaming behavior in the *Assistment* system. Additionally, by measuring engagement as a whole instead of identifying the presence or absence of gaming, the possibility of implementing a reward system to encourage good engagement behavior presents itself.

The second objective was to design and implement a live visual reporting tool for exclusive use by the teachers to assist in the evaluation of the engagement of their students during live tutoring sessions. The influence teachers have over there students [19, 12] coupled with the challenges faced by previous attempts to prevent gaming behavior directly through an ITS [16, 3] are strong indicators that providing a more effective report directly to the teacher for their intervention may be what is required. A current graphical tool that reports on student activity in the *Assistment* system provides the teachers with much valued information [20], however this tool is known to confuse and distress some students who misinterpret the very detailed information the tool displays.

2 Background

The roots of this project reside in the following areas of research: gaming detection and prevention, student engagement, assessment, and motivation. First and foremost our research is focused in the area of gaming detection and prevention within the larger subject of Intelligent Tutoring Systems. Additionally aspects of student engagement research from within the classroom has been incorporated to provide this project with a firmer grounding in educational research. Finally, student assessment and motivation in learning environments was studied for the development portions of this project.

2.1 Gaming Detection

Gaming behavior is defined by Ryan Baker and his colleagues [5] as behavior aimed at obtaining correct answers and advancing within the tutoring curriculum by systematically taking advantage of regularities in the softwares feedback and help. It was shown by their research that off-task behavior, most notably the behavior they term gaming the system, is associated with significantly reduced learning [5]. As a result, gaming detection and prevention research has been a strong focus in the field of ITS research in recent years. Models that have been developed by Baker et al. [2] have proved to be effective at identifying students who are gaming the system, however some difficulty has been encountered when detecting specifically those students who are hurt by gaming as opposed to students who learn regardless of gaming activity.

Joseph Beck expanded upon the this area of study by looking at modeling the wider area of student engagement and disengagement within an Intelligent Tutoring System as opposed to the lowest level of student disengagement which is gaming behavior [7]. His work in particular looks at the correlation between problem difficulty, student ability, and student response times in determining the probability that a student will get the problem right. This approach proved effective in detecting changes in student engagement within the scope of the subject tutoring system (Project LISTEN reading tutor) as well as characterizing important factors for detecting negative learning behavior from within an ITS [7].

Beal et al. [6] were the first researchers to integrate data collected from teachers in their study of learner engagement as related to ITS interaction. Their work focused primarily on classifying students into distinct groups based upon engagement level and associated behavior using multiple data sources, including student self reports on motivation. While this research proved successful in accomplishing its own goals, it is important to note two distinct facts about their process and results:

1. *Response time* to questions was first determined to be a factor in distinguishing the highly motivated learners from those who are either having trouble or are abusing the help system.

2. That this is the first study in gaming detection to use data from the *teachers*.

2.2 Gaming Prevention

While models have been developed and deployed for the effective detection of gaming activities and patterns, attempts to actively prevent and deter gaming activity in an ITS have met with mixed results. Murray and VanLehn [16] implemented a simple timed delay strategy in their Calculus Tutor that would encourage students to not use help request features of the tutor until a certain amount of time has passed. Their results showed that students were successfully dissuaded from requesting help in their experimental group and that in both experimental and control groups higher post-test scores were associated with less help requests. However, while this method was effective at preventing some help abuse they did observe new gaming patterns to circumvent the help request delays [16]. While these results are promising, they also show that students can adapt to observed changes in an ITS and thus discover new ways to exploit those changes and hinder their own learning.

It became clear to some in the gaming detection and prevention community that while active and direct prevention mechanisms as mentioned above meet with some success, there is still a problem prevalent when students determined to game can adapt to these changes, and the determined gamers are the primary targets of this research. Baker et al. [3] took this into consideration by designing a tutoring companion for their ITS which they named Scooter the Tutor. This animated gaming detection agent was present on the students screen during a tutoring session, changing its displayed emotion based upon detected patterns in the students interaction with the tutoring environment. Additionally the tutor would, at certain times when gaming activity was detected, provide supplementary exercises to the student. The results of their research showed that when using Scooter with the ITS there is a sizeable, though only marginally statistically significant, reduction in the frequency of observed gaming. Additionally they noted that while these results indicate success in deterring some students from gaming, those who had been determined to regularly game continued to game [3].

From the results of the above cases, two important issues present themselves. The first issue is that students who are most likely to engage in gaming behavior will pursue that path regardless of system-based prevention mechanisms. The second, and possibly less apparent, issue is that both of these techniques affect the tutoring environment in some way and, as a result, affect not only the students who game but also the students who are focused on their work. While these approaches show some success, it is clear that an alternative approach is necessary to readjust the attitudes of the students who game without hindering the work of the focused students or, even worse, falsely accusing those positively engaged students of engaging in gaming activity.

2.3 Student Engagement

Gaming is one state of disengagement with an ITS environment; however it is not the only form of disengagement. The original work of Baker et al. identified four different categories of student disengagement with their system, including the gaming the system category [5]. While they focused their work towards the gaming category, they do indicate that their results show that in general off task behavior seems to correlate with reduced learning. Research in the field of education and educational psychology has acknowledged the correlation between performance and engagement from within the classroom [14, 19]. It is possible that by addressing the larger problem of student disengagement, a solution could be reached for curbing gaming behavior as well. Additionally, expanding the study to encompass all of student engagement/disengagement could potentially allow for the detection of good behavior as well as bad behavior, thus permitting some level of positive encouragement and rewards for the focused students.

Student engagement in the classroom has been heavily investigated and recent work has been done to evaluate engagement behavior within ITS. Previous research has shown that student engagement patterns remain consistent across age groups and grade levels [14]. More recent research has shown that teaching strategies and the social environment within the classroom both have a strong affect on the avoidance strategies students use in Mathematics classrooms [19]. For instance, a teacher who works with their students in a team building environment, encouraging and guiding them towards the correct answer is more likely to motivate the students in future academic endeavors. In contrast to that teaching method, a teacher who asks students to answer questions in class and simply says whether they are right or wrong before moving on to another student with the same question is more likely to result in a student avoiding learning opportunities for fear of embarrassment before their peers [19]. Other research shows evidence to suggest that incorporating computer use in a curriculum has a positive effect on a students motivation to learn [8]. Regardless of the presence of computers within a classroom curriculum, it is clear from these data that student engagement patterns are strongly connected to the classroom environment and teacher interaction. Considering tutoring systems are an addition to the classroom learning environment, it is clear that engagement patterns that appear in the ITS environment likely originate in the classroom.

The importance of evaluating engagement within an ITS environment has been acknowledged by recent research. Beck has developed an effective model for student engagement within a reading tutor using a specially modified form of the standard formula from the Item Response Theory of psychology [7]. Beal et al. evaluated student engagement within an ITS using multiple data sources to construct five different models of student engagement behavior represented as finite-state machines [6].

Similar work has been done to make tutors that are more emotionally responsive to a students engagement by incorporating digital cameras to monitor facial features and chair sensors to detect the students posture [13, 15]. While initial work shows promise, the incorporation of sensor and camera technology could prove to be overly expensive for a school system.

From research in student engagement it is important to identify two points. First, student engagement patterns are directly influenced by teacher interaction. Second, gaming behavior is one end of the larger spectrum of student engagement behavior.

2.4 Assessment and Motivation

Having characterized the importance of evaluating student engagement in order to reduce disengagement behavior it is necessary to find ways to influence a students motivation to use an ITS effectively. The effectiveness of teachers in the classroom is strongly connected with improved student achievement [12]. Additionally, a review and integration of education and psychology research literature has shown that careful monitoring and feedback given to a student is essential in encouraging self-regulated learning, which is crucial to student achievement [9]. Students themselves have difficulty in evaluating their own capabilities, especially when they need help [1].

Research by Vicente and Pain [10] have shown that the motivational state of a student can be analyzed based upon the student's recorded activity in an ITS. In their study they used prerecorded student activity in an ITS, including mouse movements, along with student reports on motivation during the tutoring sessions as input. Their study subjects were individuals with prior teaching experience who would watch these recordings and give reasoned reports on the perceived motivation of the student at given times. They found that certain patterns presented themselves and were readily identified by their study subjects despite the initially perceived difficulty of the task [10]. They also note that since the students themselves were not under observation, only the information available to the tutoring system itself was under scrutiny.

The Assistment ITS already has an extensive reporting system providing many important details on student performance within the system [11]. Jason Walonoski has done significant research on gaming behavior within the system and has developed active and passive methods of gaming prevention [20]. The passive gaming prevention approach Walonoski developed displays a graph representing a students activity in their current curriculum within the system, offering a level of graphical reporting for the teachers to see over their students shoulders. This graph also permits students to see their own activity within the system, potentially discouraging certain known gaming patterns that are clearly seen on the graph [20].

2.5 Similarities in Prior Research

From the previous research in the related areas of student engagement, gaming behavior, and motivation the following common points are present:

1. Student response time is strongly related to engagement level and motivational state.
2. Low motivation and engagement are strongly correlated, and both tend to result in gaming behavior.
3. Teacher interaction is a strong influence on student engagement.
4. Active and passive gaming prevention mechanisms are effective at reducing overall gaming in a student population, but do not appear to influence the students who frequently engage in gaming behavior.

Additionally, it is important to note three aspects of previous gaming detection research as these aspects are the foundation of the work presented in this document. First, such research has primarily focused on detecting the presence or absence of gaming rather than measuring a student's engagement. Gaming behavior is arguably the lowest level of student disengagement with an ITS. Second, gaming prevention mechanisms that interact with the student either directly or indirectly through the ITS environment have not been shown to prevent the target audience, that being the students who are most likely and most determined to engage in gaming behavior. Finally, except in one case there is no interaction with those who have the most influence with their students as well as the most knowledge of their students' learning patterns and behaviors, the teachers.

3 Objectives

The primary goals of this research were to take a new approach to the subject of gaming detection and prevention research. Teachers have long been identified as the experts in the classroom environment, and it has been demonstrated in previous research that teachers have a direct influence on the engagement patterns of their students [19, 12]. Bearing this in mind, the absence of teacher presence in previous research into gaming detection and prevention is surprising. From this standpoint there were two main objectives for this research. The first was to determine if teacher observation data could be used to construct a reliable metric for measuring student engagement in the *Assistment* tutoring system. The second goal was to develop tools for reporting engagement-related information on student activity to the teachers during *Assistment* tutoring sessions.

3.1 Measuring Engagement

Gaming detection models previously developed have relied upon the observations of the researchers [2, 5]. While this has shown some success in this research, it has not been successful universally [20]. Considering teachers are commonly acknowledged to be experts on their students' behaviors in the classroom, the question arises as to whether or not that knowledge can be captured and emulated within an ITS to better detect gaming. Therefore, the first objective of this project is to determine if teacher reports of student engagement in the classroom can be used to generate models for measuring student engagement within the *Assistment* tutoring system.

It is important to note that previous research has primarily focused on detecting only the most detrimental form of student disengagement with an ITS, that of gaming, in the sense of their being two states for any student: gaming and not gaming Previous work on gaming detection within the *Assistment* system has also focused directly on the presence and absence of gaming behavior independent of other engagement behaviors [20]. More recent research began to look beyond that and explore student engagement as a whole as opposed to one aspect of student engagement [6, 7]. Following the success of these recent findings, this project objective was further specified towards determining if metrics from teacher observations could be made to measure the range of student engagement.

One of the early challenges which will be discussed later resided in how the teachers in the classroom could be elicited for engagement reports on their students with out disrupting their regular activities during a tutoring session. Teachers are frequently active during a tutoring session; assisting students when necessary and monitoring the activity of their students continually. Different teachers have different methods for how they act during *Assistment* sessions, however it has been observed that even teachers who prefer to leave their students be will still be quietly observing the classroom. We have yet to observe a tutoring session where the teacher is not actively watching their students for a good majority of the tutoring period.

3.2 Reporting Engagement

One of the key strengths of the *Assistment* tutoring system resides in its reporting system [11]. The *Assistment* sysem has an extensive live reporting system that can provide teachers with detailed information on their students' current and previous activities even during a tutoring session. During tutoring sessions, it has been observed that some teachers do, in fact, check some of these reports during a tutoring session to evaluate the progress of their students. Considering the effectiveness of this reporting system and prior research that has shown that teachers have the most influence on engagement patterns [19, 12], the second goal of this research was to develop reporting tools for the teachers to assist in the evaluation of student engagement during tutoring sessions.

The results of previous research in gaming prevention methods have shown that although active and passive approaches towards gaming prevention show some effect, their is still much to be desired [16, 3, 20]. Some students can be deterred from gaming behavior, but not all, and not the most determined and stubborn students, those students who likely need the most help [5, 1]. As such this research aimed to produce tools that represent engagement patterns and report student activity directly to the teachers in a manner that can be rapidly interpreted for immediate action. Additionally, the passive prevention tool developed by prior work [20] was improved.

4 Measuring Engagement

The first major focus of this research was to perform a preliminary study to determine whether or not teacher evaluations of students could be used to develop a reliable metric for measuring student engagement within the *Assistment* ITS. While current research has shown some success in determining with varying levels of reliability whether or not a student is or is not gaming the system [2, 5, 20], and other research has shown that it is possible to measure engagement [7], as of yet no other research group has used teacher input alone to develop metrics or models.

Since this part of our work was more experimental in nature it is broken down into two sub sections. The first sub section will detail the data collection and analysis procedures used in this study while the second section will cover the results and their implications.

4.1 Methods

The methods we employed for constructing a metric using teacher given data closely follow what previous researchers have done [5], howevery they differ in several key aspects. The process used in this study consists of 3 steps:

1. Classroom observation during active tutoring sessions, collecting data from the teachers as they monitored their students.
2. Dataset creation using observation times to associate ITS recorded activity with the teacher given data for the metric generation step.
3. Metric generation using piece-wise regression techniques in order to develop an effective equation for measuring student engagement as a teacher would.

4.1.1 Classroom Observation

The first challenge of this research was to collect data from the teachers without disrupting their normal classroom activity. 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 accurate 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 [5, 20].

Previous research on gaming behavior within the *Assistment* system relied on classroom observations at schools where the computer lab setup was conducive to effective observations, and the same strategy applies to this research. Tutoring sessions at the Sullivan

Middle School (SMS) in Worcester, Massachusetts were observed for several different teachers, multiple sessions per teacher. SMS was chosen for two reasons. First, the school holds regular tutoring sessions with the *Assistment* System providing more opportunities for observation as well as teachers who are more familiar with the system and how their students interact with the system. Secondly, the classroom structure was more conducive to teacher observation of their students. The computer classroom at SMS is an open room with the computers positioned along three of the four walls with an administrator's desk in front of the last wall, oriented so the administrator can observe the students while working on the administrative computer. This gives any teacher in the classroom a panorama of their students for easy monitoring.

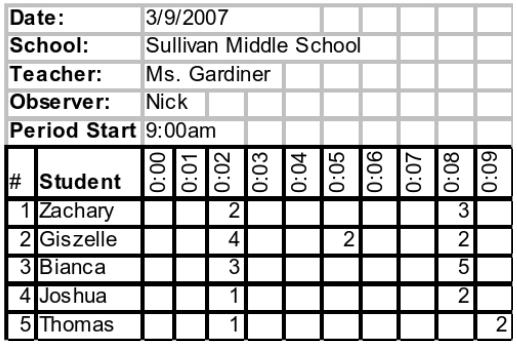
At the beginning of every tutoring session with a new teacher, the observer was instructed to briefly explain the purpose of the study and what would be required of the teacher. The actual observations consisted of the observer shadowing the teacher for the span of the tutoring session as the teacher went about their way as they would during any tutoring session. As the teacher monitored the behavior of their students, they were instructed to grade the current activity of the students they were watching using the coding scheme depicted in figure 1.

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

Figure 1: Student Focus Grades

It is important to note that the terms "focus" and "engagement" in this research refer to the same thing. While "student engagement" is the more formal term for what is under study, that being the student's engagement in the learning activity, it was determined to be more appropriate to use the term "student focus" when explaining the purpose of the research to the teachers. While "engagement", "focus", and a third term "effort" all represent the same thing, the latter two are much less formal and tended to coincide more with the terminology the teachers were familiar with, based upon initial observation sessions. However, it is important to note that while "focus" and "effort" are similar, "focus" tends to be a friendlier term to use. For example, saying a student is putting in "less effort" implies that they are not trying hard, but a student who is having difficulty with a problem and does not know when to ask for help [1] might be defensive about being labeled as putting in "less effort." This could potentially be demoralizing to students. As opposed to this, the word "focus" is arguably less threatening. The reporting tools that are discussed later are represented as tools to assist in the measure of "student focus" within the *Assistment* system.

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 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 (see figure 2 below).



Date:	3/9/2007
School:	Sullivan Middle School
Teacher:	Ms. Gardiner
Observer:	Nick
Period Start	9:00am

#	**Student**	0:00	0:01	0:02	0:03	0:04	0:05	0:06	0:07	0:08	0:09
1	Zachary			2						3	
2	Giszelle			4			2			2	
3	Bianca			3						5	
4	Joshua			1						2	
5	Thomas			1							2

Figure 2: Data Collection Table Mock-Up

A hypothetical data collection table is depicted in figure 2. This figure shows data collected on 5 different students in a tutoring session held at 9:00AM on March 9th, 2007 in a class run by Ms. Gardiner. The table indicates that grades were given for all students at 9:02AM and again at 9:08AM save for one student, Thomas, who was given a grade of 2 for his activity at 9:09AM. An individual observation was made for the second student, Giszelle, at 9:05AM. This table reflects certain observation patterns that were common for the teachers that were observed. The teachers in question had the tendency to survey the classroom at fairly regular intervals throughout the class period. In the intervening time the teachers would either be helping various students, monitoring particularly troublesome students, or monitoring the progress of the class through the *Assistment* system's extensive reporting system [11]. As such the real data collection tables tend to consist of large clusters of grades given at a common time with various individual or smaller group grade recordings in between.

The observation period for this research spanned the month of March, 2007. During this period a total of 7 classes were observed with 3 teachers providing the data. There were approximately 20 students per class yielding 265 teacher given grades in total for the analysis. All of this information is detailed in figure 3.

Time Period	March 2007
School	Sullivan Middle School
Teachers	3
Classes	7
Avg. Students Per Class	20
Total Teacher Observations	265

Figure 3: Summary of Observation Period

4.1.2 Dataset Creation

Before any analysis can take place it is important to evaluate what data is important for the analysis and to construct one or many datasets to be incorporated into the analysis. It is important to note at this point that piecewise regression was the selected analysis methodology for two reasons: 1) the technique evaluates each independent variable (student actions) in relation to the prediction of the dependent variable (engagement rating) 2) the result is an equation that can be implemented into an ITS. As a result, one large dataset was constructed incorporating as much relevant data as possible based upon the findings of previous research.

As has been discussed in the Classroom Observation section above, the data collected during observation periods consists of a teacher given “focus grade” in conjunction with the minute in time of the observation. The time value was recorded in order to associate the student’s grade with the time logged data stored within the *Assistment* system’s database. The *Assistment* system has an extensive logging system to coincide with its large and detailed reporting system [11], as such the first task of dataset creation was to evaluate what data was necessary.

The *Assistment* system automatically logs information on student activity, with the exception of mouse movement and mouse click data, within the tutoring runtime to a table called the “Actions” table. This table constitutes what is essentially an active dialog between the student on the client-side and the system runtime on the server-side, with the server sending a message relating to the user a change in the runtime state (such as correct/incorrect answer messages, new problem content, or hint messages) and the student sending a response (such as an attempt at an answer or a hint request). Along with this data is information on what problem has been displayed at any given instant and, most importantly, the time in milliseconds between a server sent message and a client sent response.

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. It has been identified by previous researchers that 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 [19, 5, 6, 7, 20]. However, performance alone is not an effective predictor of engagement since not all students who engage in gaming behavior are hurt by it [2]. Additionally, how quickly and in what ways a student interacts with the system are also acknowledged by prior studies as being important predictors of student engagement [5, 7, 20]. As for problem related data, the measure of a problem’s difficulty has been noted to correlate with engagement patterns [7, 10] as well as the type of problem being presented, in other words multiple choice problem or short answer problem [5, 20].

Based upon previous research, a dataset was constructed and its elements are listed and described in detail below in figure 4. 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. It is important to note that this data focuses on the activity of each student based upon whatever problem they were working on at the time of the observation. This tactic was chosen so as to characterize the relationship of engagement behavior with not only student behavior but student behavior with particular problems.

Focus grade - The teacher-given grade for the student at a certain time, what 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 connected 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 - 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 student's typically respond to this problem. The z-score[1] 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

[1] 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].

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 at this particular observation time.

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 essentially is a value of 0 or 1 since there is only ever one bottom out hint for any given problem.

Bottom hint this problem z-score - The same as the above z-score calculations except for the fact that this calculation is subject to the unique property of the bottom out hint in that only one is present for any given problem.

Percent of available hints requested - The number of hints requested divided by the number of hints connected to the problem the student was working on.

Percent of remaining Multiple Choice options - Exclusively for Multiple Choice problems: The number of attempts made to answer the question divided by the number of options available in the Multiple Choice problem. This is a value to represent the "guess-and-check" pattern that students' are known to follow for Multiple Choice problems.

General

Poor man's prior knowledge - A measure of the student's preceding performance. Although there is current research in the *Assistment* system that uses a student's performance as related to skills associated with different problems at varying levels of granularity [17], this does not happen live and not all problems in the *Assistment* system are associated with distinct math skills. As such a measure of the student's performance before each observation has been estimated in a similar manner to the work by Jason Walonoski [20], 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, where as a short answer problem is less conducive to this behavior.

Problem difficulty - This variable is a simple measure of a given problem's difficulty based upon all data available in the *Assistment* 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.*

Figure 4: Dataset Attributes

In addition to these values, 6 separate values were generated and added to the dataset. These values represent variations on a simple equation set for generating partial credit scores. A partial credit score can be described simply as a more detailed evaluation of what the student has accomplished during the course of their activity within the extent of a problem. At present measurements of student performance in this and other studies [20] have relied upon the notably simplistic measurement of performance that simply looks at what percentage of problems have been answered correctly. In this sense a value of 1 is given for every correct problem, and 0 for every incorrect problem. This leaves the evaluation limited to the first action a student takes upon a problem. The partial credit equation sets each hold two equations, one for Short Answer questions and the other for Multiple Choice questions. Figure 5 below displays these equations. In both equations the score is a value between 0 and 1, 0 clearly indicating the lowest possible score and 1 indicating a correctly answered problem. Both equations make an attempt to characterize the importance of how many attempts a student has made to answer a problem in addition to the number of hints the student has requested. In the Short Answer problem equation the emphasis is towards the percentage of available hints requested, depending upon the value of the constant A, whereas the Multiple Choice problem equation represents the importance of both how many hints were requested as well as how many attempts were made to answer the problem in relation to the total available answer choices. The breakdown of the equation sets used to generate the 6 additional values follows: Partial Credit 1-3 have A = 0.1, 0.15, 0.2 respectively with B = 1, Partial Credit equation sets 4-6 follow the same pattern for A as 1-3 with B = 2.

$$1 - \text{ \% available hints requested} - \text{ incorrect attempts} \times A \text{ where A is a constant}$$

(a) Short Answer Problem Equation

$$(1 - \text{ \% available hints requested})\left(1 - \frac{\text{incorrect attempts}}{\text{total choices}}\right)^B \text{ where B is a constant}$$

(b) Multiple Choice Problem Equation

Figure 5: Partial Credit Equations

Prior to the assembly of the dataset detailed above, the initial design consisted of z-scores for the first and second action times instead of counts and z-scores of the counts of each action type. However, calculating z-scores based upon very specific sequences of actions for specific problems proved to be more problematic than was initially anticipated. Initial dataset construction yielded very few z-scores based upon action sequence. As a result, a revised dataset, detailed above in figure 4, was constructed resulting in a wider range of factors for the final analysis.

4.2 Results and Discussion

The data was evaluated in several stages. First, the data was analyzed to identify any potential bias in the focus grades given by the teachers. After this, the distribution of the focus grades in relation to different data elements was evaluated graphically to search for particular trend patterns. Finally the data was analyzed using SPSS statistical software in order to generate different regression models and to determine the reliability of those models at predicting the focus grades. Regression was selected as the approach to take in developing models since the aim was to produce equations that could provide as a result a value between 1 and 5 to correspond with the focus grades.

4.2.1 Focus Grade Distributions by Teacher

From the start of this study it was understood that the possibility for teacher bias towards grading in a certain manner would be a potential risk to the data. While during the observation period the observers noted that it was indeed true that the observed teachers could effectively identify, and even predict, certain behavior patterns in their students, something that a researcher would not be able to do, it is highly likely that the teachers would not be able to distance themselves emotionally from their students to provide an objective evaluation of student behavior. Despite this risk, the study was put forth to present an alternative to prior methods in gaming detection research.

Figure 6 shows the unfortunate imbalance in the data used in this evaluation. Although three different teachers were incorporated into this study, one of these teachers did not have enough time to provide much data whereas the rest of the observations were dominated by the other two teachers. It must be noted that Teacher 2 (all names having been omitted from this report for the sake of privacy) held the most class sessions observed of all teachers and additionally had a wider variety of students, in other words a substantial mix of high, medium, and low performing students. Teacher 1 also had primarily medium to low performing students, whereas Teacher 3's classes were predominantly advanced level students. It is important to bear in mind at this point that while there is a known correlation between performance and potential for gaming behavior [5], there are students who perform well but still engage in gaming behavior [2].

Although there is potential for bias between teachers based upon the disproportionate amount of data collected from the different teachers, this can easily be discounted by the distribution of the actual focus grades given by each teacher to their students. Figure 7 shows the percentage of each focus grade given by each teacher to their students. Teacher 1 did not end up providing any 5's to their students, however the proportions of the other grades coincides with the proportions given by the other two teachers with a correspondingly high percentage of 1s given to students. If we equate the focus grades to the behavior categories detailed in prior studies [5, 20] then we end up with grade 5 representing the

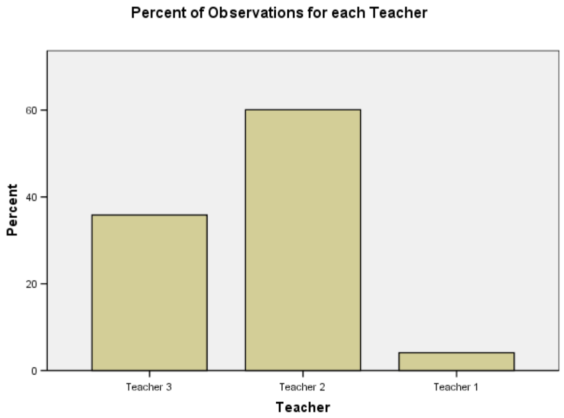
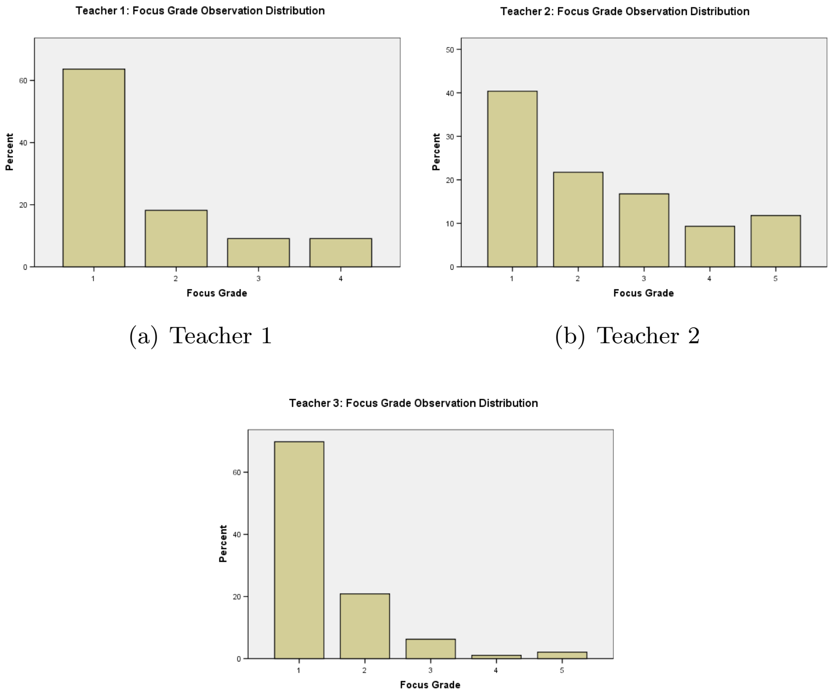


Figure 6: Percentage of Observations Made for Each Participating Teacher

gaming students with the rest of the grades representing non-gaming behavior. Following from this the low percentage of 5s distributed by each teacher is representative of the low frequency of gaming in a student population as observed by those prior studies. The significantly high proportion of 1s given by each teacher, with the second highest proportion of grades given being grade 2, indicates that teachers primarily observe positive engagement behavior in their students. The significant difference in the percent of level 4 and 5 students seen when comparing Teacher's 2 and 3 distributions of grades with a correspondingly higher proportion of 1s given by Teacher 3 correlates with the fact that the majority of Teacher 3's classes are advanced level students who, according to that teacher, tend to be very focused in their learning activities.

It is important to note that while the bias for teachers towards distributing level 1 grades is not necessarily unexpected, indeed one would presume that for the most part students would tend to be engaged with the learning activity with the occasional drift into off-task behavior with a low quantity of gaming behavior as has been observed before [20], it does present a problem with one of the desired outcomes of a positive result from this experiment. This outcome is the development of a reward system for highly focused students. Unfortunately such a system would be impractical and even distracting for the students considering such a high percentage would be marked for reward by the system. This would work well if the focus grade percentages represented a normal curve with low percentages of observations in the level 1 and 5 categories, but this is not the case. As such, although such a study is not thorough enough to provide a concrete conclusion, it can be surmised from *primae facie* evaluation of the data displayed in figure 7 that development of a reward system for student engagement is not feasible with the available knowledge.



(a) Teacher 1

(b) Teacher 2

(c) Teacher 3

Figure 7: Distribution of Focus Grades Given By Each Observed Teacher

4.2.2 Focus Grade Distributions by Dataset Attribute

Previous research has shown that several different attributes of a student's activity within an ITS are strongly correlated with detecting gaming. These attributes include student performance, problem difficulty, student time to response, and what kinds of responses the student makes to a given problem [5, 16, 7, 6, 20]. Each of these values is represented in the dataset for this study in order to play a part in this analysis.

In order to better understand the data present in the dataset and to evaluate what type of curve should be used to fit the data to in the actual regression analysis each of the variables with an identified significance in gaming detection, based upon prior research, was plotted in relation to the associated focus grades. Two dimensional and three dimensional scatter plots were employed in this phase of the analysis in order to represent the association of individual and multiple attributes with the focus grades.

The first and possibly most widely recognized attribute is the student's ability. In this study this value is represented by the percentage of problems the student has answered correctly since they have been using the *Assistment* system (see figure 4). Figure 8 below shows the scatter plot of focus grades in the dataset plotted against the associated student prior performance values. The scatter plot shows what could potentially be a linear curve to the data following the denser clusters of data points for each focus grade. Additionally, the predominance of low performance grades with the level 5 grade correlates with the findings of prior studies. However, the spread of all of these data at each level beyond 5, and the uniformity of those values at each focus grade level, pose a potential problem for the development of reliable metrics.

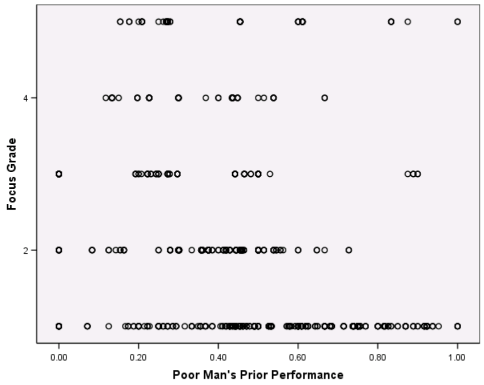


Figure 8: Scatter Plot of Focus Grades Against Student Performance

Figure 9 depicts a box plot of the same data as depicted in figure 8 which better illustrates potential association as well as the problematic distribution of the data.

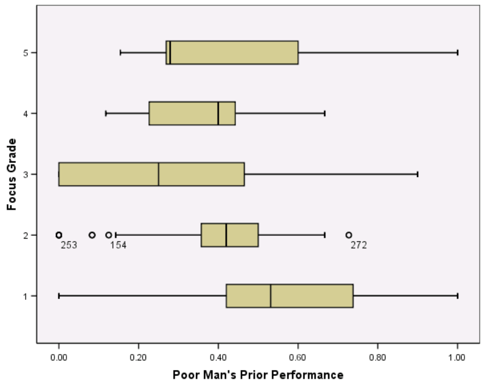


Figure 9: Box Plot of Focus Grades Against Student Performance

Problem difficulty, the next attribute considered important by prior research, shows a much more uniform distribution across each of the focus grades in the scatter plot displayed in figure 10. The clustering of the data groups in each focus grade level align too much with each other suggesting that in the case of the *Assistment* system this measure of problem difficulty (see figure 4) is not an adequate attribute for the purposes of this study.

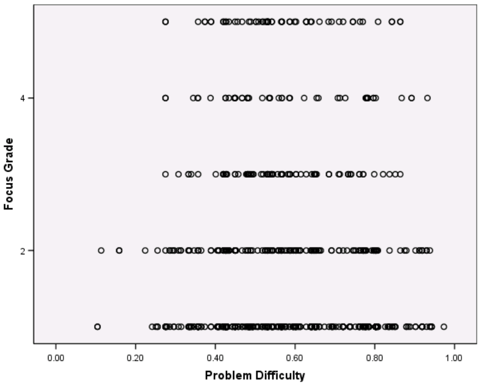


Figure 10: Scatter Plot of Focus Grades Against Problem Difficulty

A student's response time to a given problem is represented in two different but related attributes, the time taken before the student's first response and the time between the student's first response and the student's second response, in the event that a second response exists. Figure 11 depicts a three dimensional scatter plot of focus grades against the first and second action response times respectively. This plot shows a particularly appropriate clustering of values for the level 5 focus grade around the low end of both action times, indicating that students who are not focused are probably not taking the time to read the problem before either making an attempt or requesting help. It is important to note that the types of actions associated with these time values and the focus grades did not provide any further insight into this relationship between gaming and fast action times. Despite this observed correlation between level 5 focus grades and prior gaming studies involving action time as an attribute, the same problem of uniformity across all values presents itself again in this data. The wide spread of other focus grade data points presents a significant challenge to the development of reliable metrics.

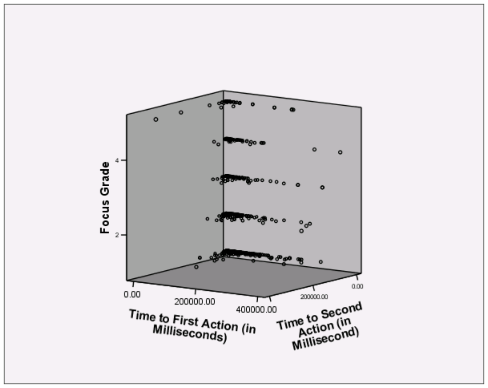


Figure 11: Scatter Plot of Focus Grades Against Student's First Action Time and Second Action Time

The student actions most associated with gaming behavior in the *Assistment* system based on prior research are hint requests and bottom out hint requests [20]. As such the last plot of focus grades against other attributes is a three dimensional scatter plot showing the relationship between the focus grades and Z-Scores for the number of hints and bottom out hints requested for the given problem (see figure 4). Figure 12 shows this plot and as with the other attributes examined in this section the data for the level 5 focus grade corresponds appropriately with what was expected of gaming students from the results of prior studies. In this case the close clustering of the level 5 focus grade data elements towards larger Z-Score values of both attributes indicates that level 5 focus grades are more directly associated with a higher than average number of hint and bottom out hint requests for the extent of a problem. This correlated with the known pattern of help abuse in gaming behavior. However, despite all of this the data shows the same troublesome uniformity that is present in the other attributes.

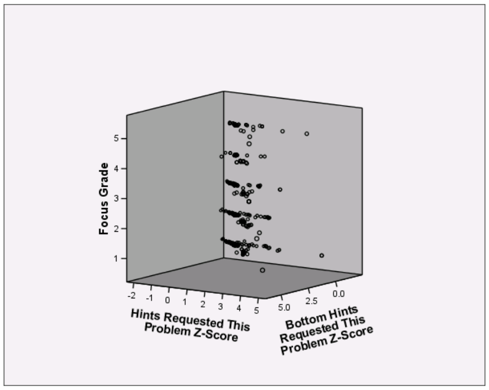


Figure 12: Scatter Plot of Focus Grades Against Z-Score of the Number of Hints Requested and Z-Score of the Number of Bottom Out Hints Requested this Problem

The results of this visual analysis of the data show that for each of the key attributes there is a strong correlation between the level 5 focus grade, which is the gaming behavior associated grade, with previous findings in gaming behavior. This indicates that teacher observations of gaming behavior correspond well with researcher observations in previous studies. Though a thorough comparison of both observation approaches is beyond the scope of this research, this result does show that teacher observations do not difer greatly from previous findings. However, the wider distribution of data across the more positive focus grades, in particular grade levels 1-3, presents a problem for the development of an actual metric from this data. This distribution covers the distributions of the level 5 data, and coupled with the higher percentage of level 1-3 grades in the dataset, reveals the possibility that the high variability may make it very difficult to reliably predict focus grades.

These data plots could suggest that measuring student engagement is a more challenging endeavor than detecting gaming behavior. This is as of yet an unexplored concept in gaming behavior research and would be worthwhile to explore further, though it is beyond the scope of this research to answer that question.

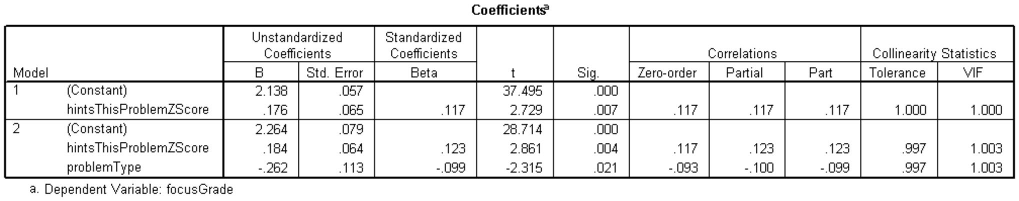
4.2.3 Regression Models

The last phase of the analysis involved the development and evaluation of equation based models of the data using statistical regression techniques within the SPSS statistical software suite.[2] Regression was selected as the technique for developing the models for two reasons: 1) the equation provided will provide as a result a value between 1 and 5 and 2) the equation can be easily implemented in the *Assistment* system. Two methods of regression analysis were performed on the data: linear and multinomial logistic. Based on the observations in section 4.2.2, despite the variability in the data there is strong visual evidence of a linear trend in each graph plotting focus grades against other attributes in the dataset. Both linear and logistic regression techniques attempt to fit data to a linear model. Linear regression treats the dependent value, the focus grades, as a continuous range from 1 to 5 where as logistic regression treates each focus grade as an independent categorical value. In order to determine what attributes are best at predicting the data, a stepwise approach was followed in both cases where each attribute in the dataset was added to an evolving model if the attribute provided a statistically significant contribution to predicting the dependent variable, in this case the focus grades.

Prior to the regression analyses the non-numeric categorical data within the dataset was encoded into numeric form. For the problem types: Multiple Choice was replaced with a 0 while Short Answer was replaced with a 1. For the first and second action types: an attempt was replaced with a 1, a hint replaced with a 2, a bottom out hint replaced with a 3, and for the second actions that do not exist (when a problem was answered correctly on the first attempt) a value of 0.

Stepwise linear regression of the dataset attributes yielded two models. The first model has one attribute which is the Z-Score for the number of hints requested while the second model has the same attribute as the first model with the addition of the problem type element. Figure 13 shows the coefficients table from the SPSS results. The t column values indicate the significance of the different attributes, as well as the constant, in the prediction of the dependent variable as well as the direction of the correlation. In this case the expectation that Short Answer problems are less likely to be present in gaming instances is presented since there is a negative correlation between the problem type value and the focus grade value. Note that the encoding scheme for problem type represents Short Answer with a value of 1 and a numerically high focus grade represents greater likelihood of gaming. Additionally, there is a positive correlation between the Z-Score of the number of hints requested and the focus grade, which is also expected based upon prior research since a higher Z-Score indicates more hints requested and thus a greater likelihood of gaming.

[2] http://www.spss.com/, software licensed through WPI.



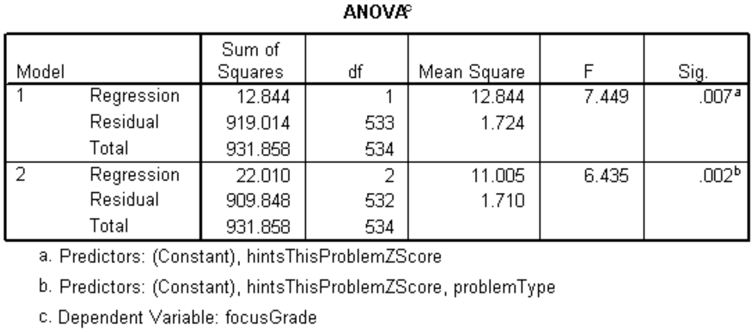
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	2.138	.057		37.495	.000					
	hintsThisProblemZScore	.176	.065	.117	2.729	.007	.117	.117	.117	1.000	1.000
2	(Constant)	2.264	.079		28.714	.000					
	hintsThisProblemZScore	.184	.064	.123	2.861	.004	.117	.123	.123	.997	1.003
	problemType	-.262	.113	-.099	-2.315	.021	-.093	-.100	-.099	.997	1.003

a. Dependent Variable: focusGrade

Figure 13: Linear Regression Models and Coefficient Values

Despite the positive indications of the linear model coefficients and the expected correlations, the model itself is disappointing. Results from Analysis of Variance (ANOVA) evaluation of the models, displayed in figure 14, show substantially low quantities for the sum of squares of the regression model predicted values in comparison to the sum of squares of the residual values[3]. This difference indicates that the model accounts for only a small quantity of the variation present in the data. As such a linear regression model can not accurately predict the focus grades in this dataset.



ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12.844	1	12.844	7.449	.007[a]
	Residual	919.014	533	1.724		
	Total	931.858	534			
2	Regression	22.010	2	11.005	6.435	.002[b]
	Residual	909.848	532	1.710		
	Total	931.858	534			

a. Predictors: (Constant), hintsThisProblemZScore

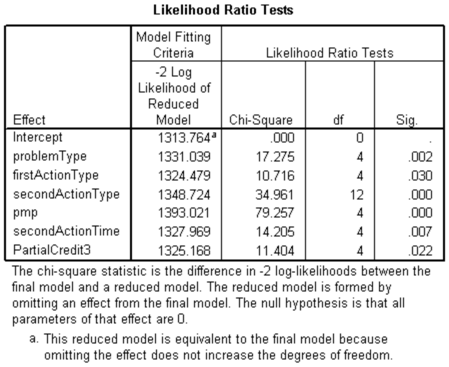
b. Predictors: (Constant), hintsThisProblemZScore, problemType

c. Dependent Variable: focusGrade

Figure 14: ANOVA Results for the Linear Regression Models

Stepwise multinomial logistic regression was attempted after the linear regression analysis and provided more promising results. First and foremost the model resulting from the stepwise process included a wider range of attributes. Figure 15 shows the list of attributes that were determined to be most statistically significant to the model during the stepwise procedure. The inclusion of problem type corresponds well with the second linear model from the previous analysis, however it is important to note that the Z-Score for the number of hints requested which was present in both of the linear models is absent. Additionally this model includes factors for the actions first performed in response to the problem as well as some action timing data. The high Chi-Square value for the "pmp" attribute, "pmp" representing "Poor Man's Prior Performance," suggests that that parameter is of particular importance to the model. Additionally the third partial credit equation appears to be of some importance indicating that some measure of the activity of the student for the extent of the problem was significant.

[3] Residual values in regression are the difference between the actual data points and the regression line.



Likelihood Ratio Tests

	Model Fitting Criteria	Likelihood Ratio Tests		
Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	1313.764[a]	.000	0	.
problemType	1331.039	17.275	4	.002
firstActionType	1324.479	10.716	4	.030
secondActionType	1348.724	34.961	12	.000
pmp	1393.021	79.257	4	.000
secondActionTime	1327.969	14.205	4	.007
PartialCredit3	1325.168	11.404	4	.022

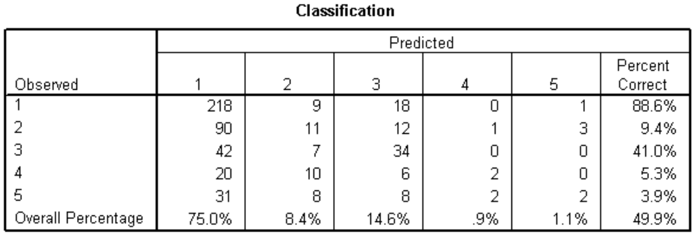
The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

[a] This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Figure 15: Logistic Regression Selected Attributes

In multinomial logistic regression a probabilistic equation is produced for each of the categorical values being predicted. Each of these equations has a different set of coefficient values for each parameter. The complete table of all parameters for the logistic model produced in this analysis is presented in Appendix A.

After the logistic regression model was produced an evaluation of the model was performed by attempting to classify the dataset data with the model. Figure 16 displays the results of this test. Looking at the percentages shown on the rightmost column we see that the model is capable of classifying level 1 focus grade actions 88% of the time with all but the level 3 focus grade being classified accurately less than 10% of the time. This is not unexpected considering the observations made of the distribution of the data across focus grade values in section 4.2.2. However, these accuracy rates indicate that this model is not capable of reliably interpreting the data from the dataset to predict teacher given focus grades.



Classification

	Predicted					
Observed	1	2	3	4	5	Percent Correct
1	218	9	18	0	1	88.6%
2	90	11	12	1	3	9.4%
3	42	7	34	0	0	41.0%
4	20	10	6	2	0	5.3%
5	31	8	8	2	2	3.9%
Overall Percentage	75.0%	8.4%	14.6%	.9%	1.1%	49.9%

Figure 16: Classification Accuracy Table for the Logistic Regression Model

These results are discouraging considering the accuracy of the teachers in predicting very specific student behaviors as noted by the classroom observers. However, these results are not surprising considering the significant variability in the data as described in section 4.2.2. The classification accuracy results from the multinomial logistic regression model slightly correspond to the classification accuracy of the machine learned models produced by a prior study of gaming behavior in the *Assistment* system [20]. This previous study produced models that could detect non-gaming behavior with a high degree of accuracy, however the prediction of gaming behavior was unacceptably low as well. This could indicate that the data present in the dataset is not appropriate for the detection of gaming behavior within the *Assistment* system.

5 Reporting Engagement

The second goal of this work was to develop tools for teachers to assist in the evaluation of their students' engagement in the learning activity during an active tutoring session. Previous research in gaming prevention methods have yielded inconsistent results which begs the question of what actually is effective at preventing negative engagement behavior [3, 16, 20]. Jason Walonoski developed a passive gaming indicator meant to deter students from gaming, however instead it proved to be an effective reporting tool for teachers [20]. Since teacher involvement is strongly related to student engagement [19, 12], its clear that the best approach to take to prevent off-task behavior is to better inform the teacher and as such have the prevention, or rather intervention, come directly from the teacher.

This section is broken down into two distinct subsections. The first subsection details the design, implementation, and initial teacher impressions of a new visual reporting tool called the Summary Chart. The second subsection details the results of a cluster analysis of the data from the dataset discussed in section 4.1.2 that is equivalent to the data presented by the Summary Chart and how those clusters correspond with known gaming behavior patterns.

5.1 Summary Chart

In a previous study of gaming behavior within the *Assistment* system, a passive gaming prevention mechanism known as the "Progress Bar" was developed to help deter students from gaming activity [20]. Additionally it was thought by the researchers that it could be an effective tool for the teachers evaluating their students, and this did in fact turn out to be the case. In fact, at present of all the teachers that have seen the Progress Bar, only one teacher has expressed the desire to have the component deactivated for their classroom sessions. It is clear that this tool, although primarily designed for the student, has become more effective at reporting behavior to the teacher. As for the student, however, the individuals involved in this research who observed classrooms during the data collection phase (see Section 4) noted that some students were often confused and even distressed at what they believed the Progress Bar was showing. Certainly passive prevention mechanisms are meant to in some sense scare the student away from gaming behavior through the impression that they are being monitored, however for a student who is concerned about doing well who perhaps may not be a strong performer confusion and distress are the wrong emotions we want to be evoking. In response to this it has been determined that a live graphical reporting tool that is available exclusively to the teachers is what is necessary.

The we have developed is what we call a "Summary Chart" simply because it constructs a visual "summary" of a student's activity within the *Assistment* system. The chart displays student activity over time, with the X axis depicting time and the Y axis as a scale for the number of actions the student has taken. The chart contents represent every main problem as a single stacked bar. A main problem is defined as being a single distinct problem composed of a question text, input type, correct answer or correct answer set, one to many hints, and 0 or many scaffolding problems. Another term that will be used from this point on is a problem extent, which can be defined as the sequence of events taken by a student from the moment they are first shown the main problem to the time they complete the main problem. Following these definitions, the Summary Chart shows bars representing summaries of each problem extent of a main problem the student has completed over the course of a specified day's activity. Figure 17 below shows a Summary Chart as it would appear in the reporting tool to a teacher.

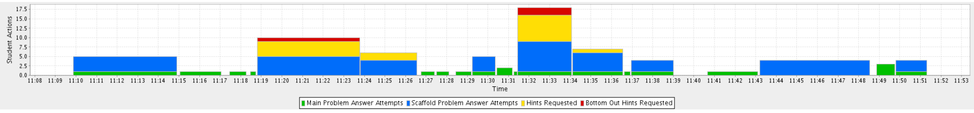
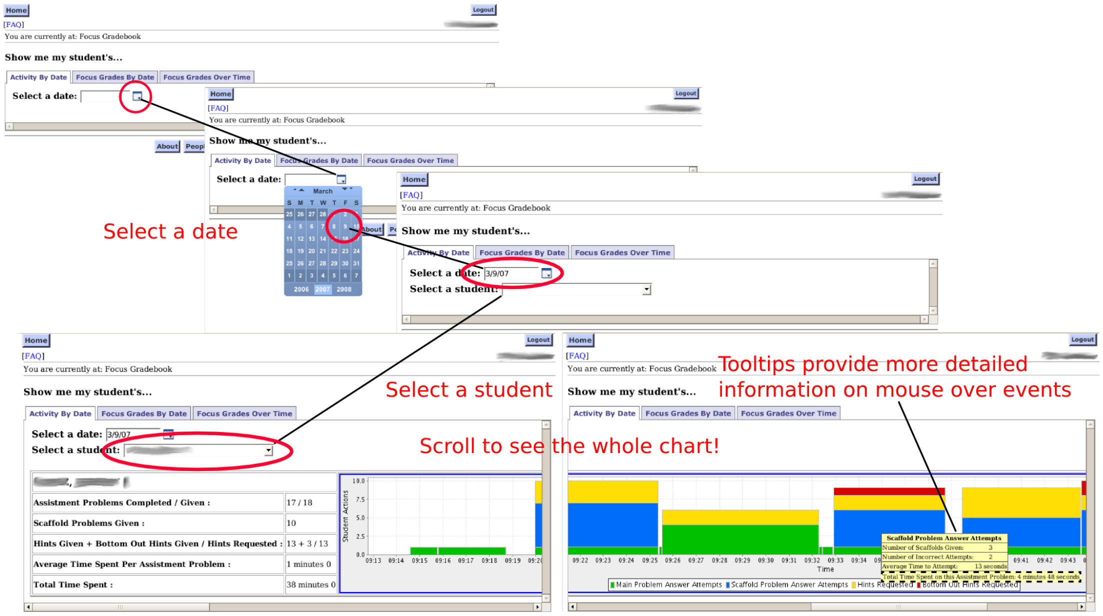


Figure 17: Sample Summary Chart generated from real student data

Going into further detail (see figure 17 for reference), the width of each bar represents the total amount of time spent on the problem from the time it was first shown to the student to the time the student either correctly answered the problem or correctly answered the last associated scaffold problem. Thus a student who spends a lot of time on a main problem will have a very wide bar for that particular problem. The height of the bar represents the number of actions the student has performed during the extent of their work on the problem. Thus a student who answers a problem correctly on the first attempt will have a very short bar for that problem while a student who requested two hints and made two attempts, the last being a correct answer, will have a bar 4 times the size of the correctly answered question's bar. Each bar is vertically subdivided into as many as 4 color coded sections. From the bottom up these sections represent the following: 1) Green for the number of attempts made to answer the main problem, 2) Blue for the number of attempts made to answer any connected scaffold problems, 3) Yellow for the number of regular hints requested, and 4) Red for the number of bottom out hints requested.

The web-based reporting tool itself was designed with simplicity in mind. Figure 18 depicts a sequence of screenshots of the web-based tool showing what a teacher does in order to display a Summary Chart for one of their students (names have been blurred out to maintain the privacy of the users). The teacher may select either the current date or any previous date when any of their students have been using the system to produce a selection list of their students who have been using the system that day. Selecting any one of these students will produce an up-to-date Summary Chart with some additional textual summary information to the left of the chart itself. Additionally, moving the mouse over any of the vertical subsections of any bar will display more specific information regarding the data that that particular subsection displays.

So how effective is this tool in application? Does the chart correspond well with teacher observations of student activity, as described in section 4? The chart and its web reporting



Select a date: 3/9/07

Select a student: [redacted]

[redacted]	
Assistment Problems Completed / Given :	17 / 18
Scaffold Problems Given :	10
Hints Given + Bottom Out Hints Given / Hints Requested :	13 + 3 / 13
Average Time Spent Per Assistment Problem :	1 minutes 0
Total Time Spent :	38 minutes 0

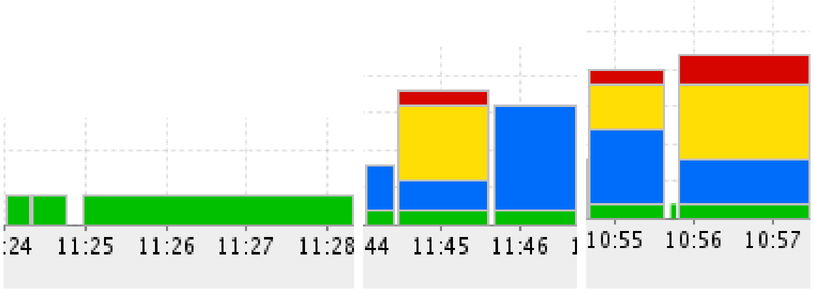
Scaffold Problem Answer Attempts	
Number of Scaffolds Given:	3
Number of Incorrect Attempts:	2
Average Time to Attempt:	13 seconds

Total Time Spent on this Assistment Problem: 4 minutes 48 seconds

Figure 18: Screenshots of a typical users interaction with the Summary Chart Reporting Tool

tool was developed before the classroom observation stage of the work detailed in section 4. As a result, most of the teachers involved were shown the reporting tool during live tutoring sessions. Although this does not constitute a formal evaluation, and such an evaluation would require additional work beyond the scope of this research, these initial impressions from the teachers were very positive and encouraging for future work. Additionally, it has been noted by the observers of these classrooms that some of the teachers tended to monitor the progress of their students from a central computer using the *Assistment* reporting system, and these same teachers noted that this tool would be useful to them as a result. Additionally, one teacher noted certain visible patterns in the chart that coincided with recent observations they had made of their owne students.

Figure 19 shows three segments of Summary Charts, produced from three separate students who were monitored during the classroom observation periods (see section 4). Subfigure 19(a) shows a chart segment from a period when the teacher on-duty observed the student to be engaged in the learning activity. This is clearly seen by the low green bars. The variation in width could indicate that the problem that took longer was more difficult; however, based solely upon the larger bar in this graph it is clear by the absense of any requests for hints that the student took the time to find a correct solution to the problem. Subfigure 19(b) tells a different story. In this segment the bars represented show a high level of activity for each of the three problems, with the presence of a bottom-out hint in one of the bars. While students having more trouble might end up with bars similar in composition, the width of these bars is markedly small for so much activity. As such the engagement of the student is arguably not what it should be. Subfigure 19(c) depicts ac-



(a) Focus Grade: 1 (b) Focus Grade: 3 (c) Focus Grade: 5

Figure 19: Summary Chart segments corresponding to focus grades (see figure 1) given by teachers during classroom observation periods

tivity that was awarded the lowest of the possible focus grades, indicating that the student was completely off-task and, potentially, gaming the system. The presence of red alone is a negative engagement indicator, however it must be noted that the thicker red section of the bar to the far right indicates that the student went to bottom-out hints multiple times in a single problem. This, coupled with the dominant proportion of yellow in the same bar, indicates that the student was probably abusing the help system. Additionally, the high and noteably thinner bars indicate that the student was not spending the time necessary.

For comparative analysis, Summary Charts for three different students are shown in figures 20, 21, and 22. These charts were generated using data from a completed hour-long class period. Additionally, these students were rated by their teacher using the focus grading scheme (see figure 1) based upon their typical classroom behavior. It is clear that some of the same patterns that were previously discussed are also present in some of these figures and tend to coincide with the teacher given ratings of the students to whom these charts refer.

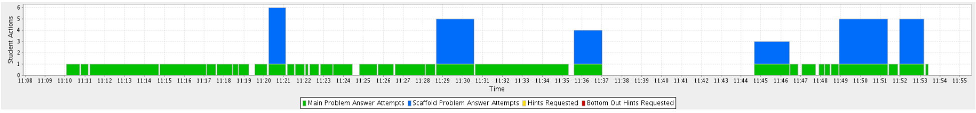


Figure 20: Full Summary Chart for a student who typically displays level 1 behavior

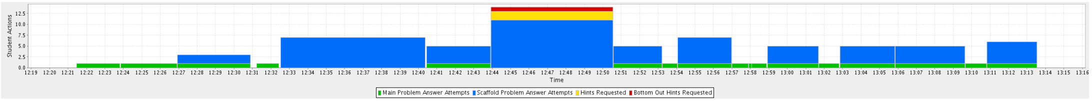


Figure 21: Full Summary Chart for a student who typically displays level 3 behavior

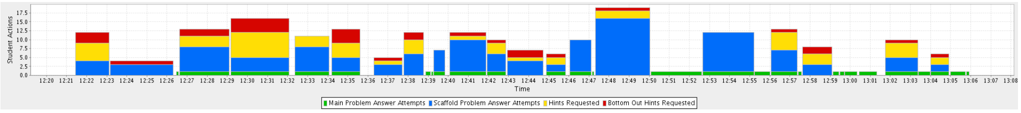
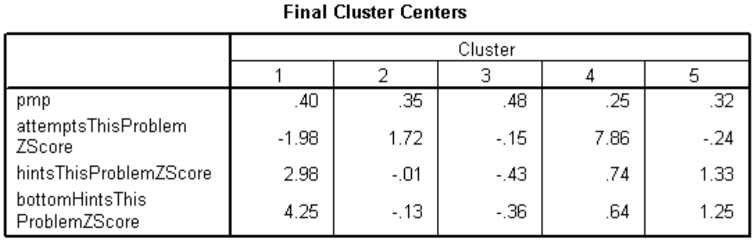


Figure 22: Full Summary Chart for a student who typically displays level 4 behavior

5.2 Cluster Analysis of Summary Chart Related Data

The strong correlation between particular patterns in the Summary Charts and the teacher given focus grades of previous sections suggested that further study of attributes that correspond to the data presented in the Summary Chart was necessary. Such an analysis was performed using the SPSS software suite k-means cluster analysis procedure using the Z-Score values for the number of attempts made, hints requested, and bottom out hints requested for each row in the dataset described in section 4.1.2. The analysis procedure attempted to classify the data provided into 5 different clusters.

The results of the cluster analysis provide some insight into the behaviors of students in the *Assistment* system that corresponds with previously identified student behavior patterns. Figure 23 below shows the characteristics of each cluster center as represented by the values of each of the provided attributes for those centers. Clusters 1 and 4 particularly stand out with noticeably high values for hint requests of both types of hints and attempts made respectively. These two clusters appear to represent help abuse behavior, for cluster 1, and guess-and-check behavior, for cluster 4. Cluster 2 seems to represent a lesser form of cluster 4 while cluster 5 represents a lesser form of cluster 1. Cluster 3 would appear to be the grouping of the most determined students trying to answer the corrections with the absolute minimum amount of attempts made and hints requested.

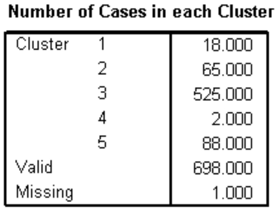


Final Cluster Centers

	Cluster				
	1	2	3	4	5
pmp	.40	.35	.48	.25	.32
attemptsThisProblem ZScore	-1.98	1.72	-.15	7.86	-.24
hintsThisProblemZScore	2.98	-.01	-.43	.74	1.33
bottomHintsThis ProblemZScore	4.25	-.13	-.36	.64	1.25

Figure 23: Characteristics of the Cluster Centers

Evaluation of the quantity of rows placed in each cluster, shown in figure 24, reveals the same pattern as seen in the proportional distributions of teacher given focus grades in section 4.2.1 with the most focused students maintaining the majority, as indicated by the quantity of cluster 3. This and the low quantity of cases in the gaming behavior clusters (1 and 4) suggest that this clustering corresponds appropriately with the proportions of gaming and non-gaming instances as found in previous studies [5, 20]. It is important to note that the actual focus grades when associated with the clusters do not correspond well with the observed properties of each cluster.



Number of Cases in each Cluster

Cluster	1	18.000
	2	65.000
	3	525.000
	4	2.000
	5	88.000
Valid		698.000
Missing		1.000

Figure 24: Distribution of Dataset Rows For Each Cluster

The Summary Chart is most effective at displaying the behaviors described by clusters 1 and 4, with cluster 1 behavior being represented by bars with a red section and a noteably high quantity of yellow and cluster 4 behavior being represented by excessively high bars with a high quantity of green and blue with some yellow and red. This analysis provides some further insight into the capabilities of the Summary Chart at displaying particular student behavior patterns.

6 Conclusions

The off-task behavior known as "gaming the system" [5] has received a lot of attention in recent years due to its connection with reduced learning in Intelligent Tutoring Systems. The main goal of this research was to approach the study of gaming behavior from a previously unexplored direction. The new approach was to bring the teacher into the situation as the expert witness of student behavior in the classroom as well as the method of prevention of gaming behavior. This was accomplished in two parts: 1) the exploratory analysis of teacher reports on student behavior being used to create metrics for measuring student engagement and 2) the development of visual reporting tools to provide live feedback on student behavior to the teachers during active tutoring sessions.

The exploratory analysis of teacher expert data showed a significant bias towards grading students more favorably in their engagement with the system. Despite the appropriateness of the clustering of low engagement grades plotted against attributes known to be associated with gaming behavior, linear and logistic regression models proved to be incapable of overcoming the strikingly high level of variance in the rest of the data. The classification accuracy of non-gaming behavior correlated with machine learned models in a preceding study of gaming within the *Assistment* system which also suffered from an inability to accurately detect gaming behavior.

The other aspect of this research, the development of a graphical reporting tool for the teacher, proved to be substantially more successful. This tool, called the Summary Chart, presents the teacher with a visual summary of a student's progress on a given day through the *Assistment* system with particular emphasis on the time spent on each problem and the quantity and proportions of the different actions performed by the student for the extent of their interaction with the problem. Informal evaluation using input from the teachers observed in earlier aspects of this research yielded encouraging results and comparison of student Summary Charts with teacher given engagement ratings from earlier in this study showed a strong correlation between high and low engagement ratings and certain highly visible patterns in the visualization. Cluster analysis of the data in the dataset used in the first part of this research correlates with known gaming behaviors that are displayed particularly clearly in the Summary Chart.

7 Future Directions

The results of this research provide a solid starting point from which new research into gaming detection and prevention methods can take place. Despite the poor performance of the models for gaming detection presented in this research, the conclusion can not yet be made as to whether or not teacher expert data is a viable, or even better, alternative to gaming detection models developed from researcher observations. Additionally, the question arises from the results as to whether or not detecting gaming is more or less difficult than measuring student engagement in an ITS.

The inaccuracies of the logistic regression model and the machine learned models of the preceding study of gaming behavior in the *Assistment* system [20] suggest that the data elements used in both studies are inadequate for the development of reliable gaming or engagement models for the *Assistment* system. Possible missing attributes include: using student performance and problem difficulty ratings based upon math skill models being studied within the *Assistment* system [17], mouse movements and mouse clicks in the interface not currently recorded by the *Assistment* system, and possibly specific attributes for the visual patterns presented by the Summary Chart.

The Summary Chart shows promise as an effective visual reporting tool for teachers, however a thorough evaluation of its usability and effect have yet to be performed. In addition to this it is possible that the Summary Chart could be more effective at deterring student gaming behavior than the Progress Bar currently used by the system as a passive gaming prevention mechanism [20]. Even though the Summary Chart was designed as an engagement reporting tool specifically for teacher use, it is possible that it could be a less confusing and, therefore, more effective passive deterrent of gaming activity.

These unanswered questions provide a multitude of directions and possible contributions that can be made to the study of gaming behavior in Intelligent Tutoring Systems.

Acknowledgements

I would like to thank my advisor, Neil Heffernan, for his support and guidance. I would also like to thank Professor Carolina Ruiz for her advice for the first part of this research; Ms. Ohanesian, Ms. Gardiner, Ms. Fulk, and Mrs. Adamowicz for their invaluable assistance; and Bob Jennings for his comments and suggestions.

References

[1] V. Aleven and K. R. Koedinger. Limitations of student control: Do students know when they need help? In C. F. G. Gauthier and K. VanLehn, editors, *Proceedings of the 5th International Conference on Intelligent Tutoring Systems*, pages 292–303. Springer Verlag, 2000.

[2] 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*, pages 531–540, 2004.

[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*, pages 392–401, 2006.

[4] 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*, pages 402–411, 2006.

[5] 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*, pages 383–390, 2004.

[6] 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*, pages 2–8, Menlo Park, California, 2006. AAAI Press.

[7] 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)*, pages 88–95, Amsterdam, 2005.

[8] H. J. Becker. Pedagogical Motivations for Student Computer Use that Lead to Student Engagement. *Educational Technology*, 40:5–17, 2000.

[9] D. L. Butler and P. H. Winne. Feedback and Self-Regulated Learning: A Theoretical Synthesis. *Review of Educational Research*, 65:245–281, 1995.

[10] A. de Vicente and H. Pain. Informing the Detection of the Students' Motivational State: An Empirical Study. *Sixth International Conference on Intelligent Tutoring Systems*, pages 933–943, 2002.

[11] M. Feng and N. T. Heffernan. Informing Teachers Live about Student Learning: Reporting in the Assistment System. *The 12th International Conference on Artificial Intelligence in Education Workshop on Usage Analysis in Learning Systems*, 2005.

[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] A. Kapoor, S. Mota, and R. W. Picard. Towards a Learning Companion that Recognizes Affect. In *Proceedings from Emotional and Intelligent II: The Tangled Knot of Social Cognition, AAAI Fall Symposium*, 2001.

[14] H. M. Marks. Student Engagement in Instructional Activity: Patterns in Elementary, Middle, and High School Years. *American Educational Research Journal*, 37:153–184, 2000.

[15] S. Mota and R. W. Picard. Automated Posture Analysis for Detecting Learner's Interest Level. *IEEE Workshop CVPRHCI in conjunction with CVPR '03*, 2003.

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

[17] 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.

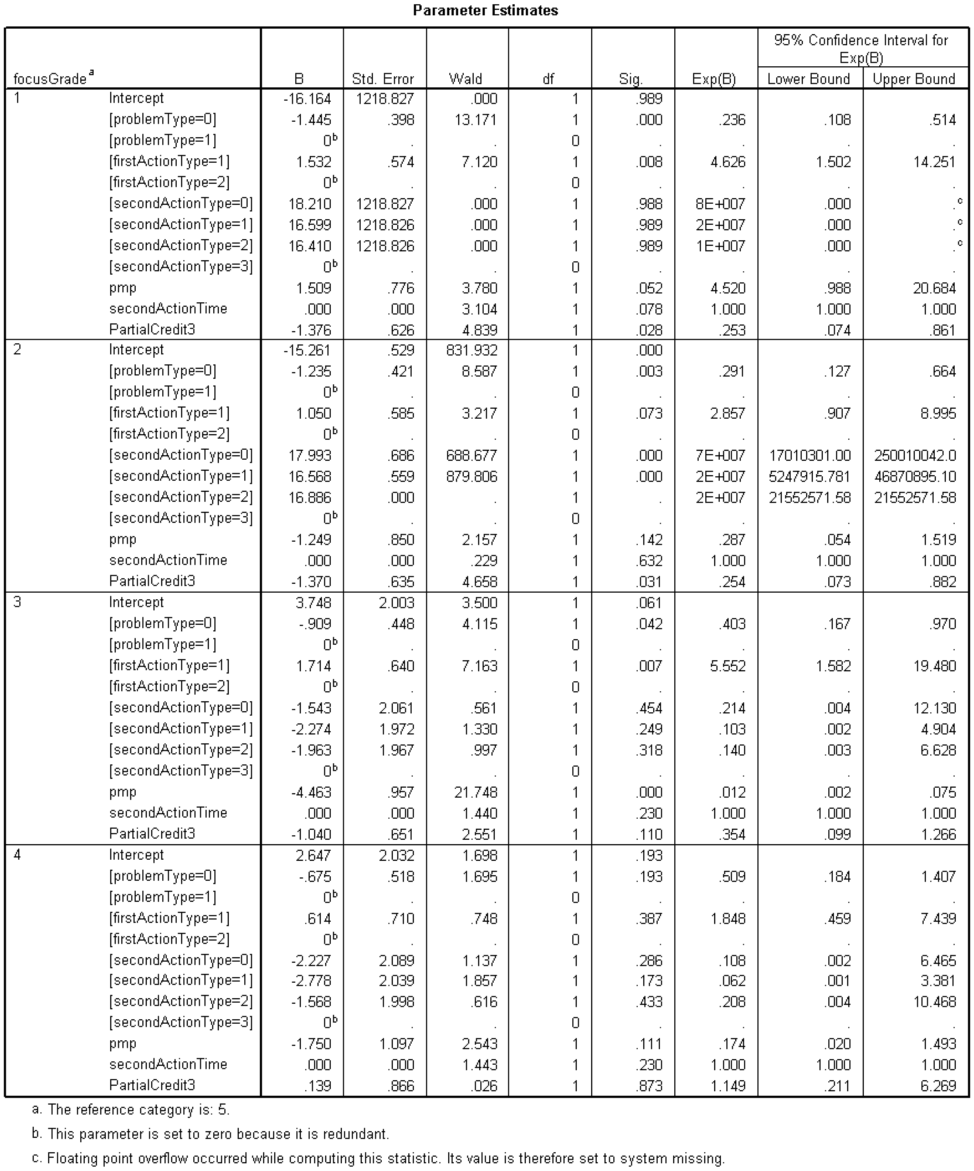
[18] L. Razzaq, M. Feng, G. Nuzzo-Jones, N. T. Heffernan, K. Koedinger, B. Junker, S. Ritter, A. Knight, E. Mercado, T. E. Turner, R. Upalekar, J. A. Walonoski, M. A. Macasek, C. Aniszczyk, S. Choksey, T. Livak, and K. Rasmussen. The Assistment Project: Blending Assessment and Assisting. In B. B. . J.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

[19] 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.

[20] J. A. Walonoski. Visual Feedback for Gaming Prevention in Intelligent Tutoring Systems. Master's thesis, Worcester Polytechnic Institute, 2005.

A - Logistic Regression Model Parameters

The following table displays the full output of all parameters in each of the 5 logistic regression model equations produced by the logistic regression analysis detailed in section 4.2.3. In multinomial logistic regression, the output is a set of equations that produce a probability value based upon certain inputs for predicting a categorical value. An equation for each of the values to be predicted is produced and, after each equation is provided with inputs, the result with the highest returned probability value is determined to be the prediction based upon the input data. Each of the parameter values below has been determined by the analysis to be statistically significant for each equation in the regression with the value for "Intercept" representing the constant in the equation.



Parameter Estimates

								95% Confidence Interval for Exp(B)	
focusGrade[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	Lower Bound	Upper Bound
1	Intercept	-16.164	1218.827	.000	1	.989			
	[problemType=0]	-1.445	.398	13.171	1	.000	.236	.108	.514
	[problemType=1]	0[b]	.	.	0	.	.	.	.
	[firstActionType=1]	1.532	.574	7.120	1	.008	4.626	1.502	14.251
	[firstActionType=2]	0[b]	.	.	0	.	.	.	.
	[secondActionType=0]	18.210	1218.827	.000	1	.988	8E+007	.000	.[c]
	[secondActionType=1]	16.599	1218.826	.000	1	.989	2E+007	.000	.[c]
	[secondActionType=2]	16.410	1218.826	.000	1	.989	1E+007	.000	.[c]
	[secondActionType=3]	0[b]	.	.	0	.	.	.	.
	pmp	1.509	.776	3.780	1	.052	4.520	.988	20.684
	secondActionTime	.000	.000	3.104	1	.078	1.000	1.000	1.000
	PartialCredit3	-1.376	.626	4.839	1	.028	.253	.074	.861
2	Intercept	-15.261	.529	831.932	1	.000			
	[problemType=0]	-1.235	.421	8.587	1	.003	.291	.127	.664
	[problemType=1]	0[b]	.	.	0	.	.	.	.
	[firstActionType=1]	1.050	.585	3.217	1	.073	2.857	.907	8.995
	[firstActionType=2]	0[b]	.	.	0	.	.	.	.
	[secondActionType=0]	17.993	.686	688.677	1	.000	7E+007	17010301.00	250010042.0
	[secondActionType=1]	16.568	.559	879.806	1	.000	2E+007	5247915.781	46870895.10
	[secondActionType=2]	16.886	.000	.	1	.	2E+007	21552571.58	21552571.58
	[secondActionType=3]	0[b]	.	.	0	.	.	.	.
	pmp	-1.249	.850	2.157	1	.142	.287	.054	1.519
	secondActionTime	.000	.000	.229	1	.632	1.000	1.000	1.000
	PartialCredit3	-1.370	.635	4.658	1	.031	.254	.073	.882
3	Intercept	3.748	2.003	3.500	1	.061			
	[problemType=0]	-.909	.448	4.115	1	.042	.403	.167	.970
	[problemType=1]	0[b]	.	.	0	.	.	.	.
	[firstActionType=1]	1.714	.640	7.163	1	.007	5.552	1.582	19.480
	[firstActionType=2]	0[b]	.	.	0	.	.	.	.
	[secondActionType=0]	-1.543	2.061	.561	1	.454	.214	.004	12.130
	[secondActionType=1]	-2.274	1.972	1.330	1	.249	.103	.002	4.904
	[secondActionType=2]	-1.963	1.967	.997	1	.318	.140	.003	6.628
	[secondActionType=3]	0[b]	.	.	0	.	.	.	.
	pmp	-4.463	.957	21.748	1	.000	.012	.002	.075
	secondActionTime	.000	.000	1.440	1	.230	1.000	1.000	1.000
	PartialCredit3	-1.040	.651	2.551	1	.110	.354	.099	1.266
4	Intercept	2.647	2.032	1.698	1	.193			
	[problemType=0]	-.675	.518	1.695	1	.193	.509	.184	1.407
	[problemType=1]	0[b]	.	.	0	.	.	.	.
	[firstActionType=1]	.614	.710	.748	1	.387	1.848	.459	7.439
	[firstActionType=2]	0[b]	.	.	0	.	.	.	.
	[secondActionType=0]	-2.227	2.089	1.137	1	.286	.108	.002	6.465
	[secondActionType=1]	-2.778	2.039	1.857	1	.173	.062	.001	3.381
	[secondActionType=2]	-1.568	1.998	.616	1	.433	.208	.004	10.468
	[secondActionType=3]	0[b]	.	.	0	.	.	.	.
	pmp	-1.750	1.097	2.543	1	.111	.174	.020	1.493
	secondActionTime	.000	.000	1.443	1	.230	1.000	1.000	1.000
	PartialCredit3	.139	.866	.026	1	.873	1.149	.211	6.269

a. The reference category is: 5.

b. This parameter is set to zero because it is redundant.

c. Floating point overflow occurred while computing this statistic. Its value is therefore set to system missing.

Figure 25: Logistic Regression Model Parameters