Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems

Ryan S.J.d. Baker
Learning Sciences Research Institute, University of Nottingham
Wollaton Road, Jubilee Campus, Nottingham, NG9 2FW, UNITED KINGDOM
ryan@educationaldatamining.org

ABSTRACT
We present a machine-learned model that can automatically detect when a student using an intelligent tutoring system is off-task, i.e., engaged in behavior which does not involve the system or a learning task. This model was developed using only log files of system usage (i.e. no screen capture or audio/video data). We show that this model can both accurately identify each student's prevalence of off-task behavior and can distinguish off-task behavior from when the student is talking to the teacher or another student about the subject matter. We use this model in combination with motivational and attitudinal instruments, developing a profile of the attitudes and motivations associated with off-task behavior, and compare this profile to the attitudes and motivations associated with other behaviors in intelligent tutoring systems. We discuss how the model of off-task behavior can be used within interactive learning environments which respond to when students are off-task.

Author Keywords
Intelligent Tutoring Systems, User Modeling, User Attitudes, Motivation, Off-Task Behavior

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous. K3.m. Computers and Education: Miscellaneous.

INTRODUCTION

In recent years, there has been considerable attention to modeling and understanding the behavior of students as they use interactive learning environments [cf. 1,3,8,9]. However, the vast majority of this past work has focused specifically on how students choose to act within the software. A student's behavior outside of a system may also affect how well the student learns from the software.

CHI 2007, April 28–May 3, 2007, San Jose, California, USA.

One such type of behavior that may affect students' learning is off-task behavior, where a student completely disengages from the learning environment and task to engage in an unrelated behavior. Examples of off-task behavior include talking to other students about unrelated subjects [7], disrupting other students [31], and surfing the web [7].

It has been hypothesized that off-task behavior is associated with poorer learning [12], but this hypothesis has only been studied to a limited degree within learning environments. In one study, Baker and his colleagues [7] reported that off-task behavior was not significantly correlated to lower learning within Cognitive Tutor software [cf. 2]. However, a later meta-analysis by the same research group [6] found a statistically significant negative correlation between off-task behavior and learning. Hence, it may be possible to make Cognitive Tutors – and other types of interactive learning environments – more educationally effective, by detecting and responding to off-task behavior.

It is worth noting that off-task behavior occurs in many types of interactive systems beyond just educational software. Many technology-supported tasks, from conducting surveillance with video [cf. 14] to driving a car, depend on a continually engaged user. Such systems might also be more effective if they could detect when their user is not paying attention to the task at hand.

Detecting whether a student is off-task, in a classroom setting, is likely to be a challenging task. In a highly instrumented setting, with microphones, gaze trackers, or fMRI machines, it might be relatively easy to determine whether a student is off-task. However, such equipment is not available to most schools; for a system to be widely useful, it must detect off-task behavior using data only from students' actions within the software. It has been found that recognizing a user's intentions solely from his or her actions within a system can be quite challenging [29]; however, off-task behavior detection need not be perfect in order to be useful. In existing learning environment-based school curricula, the responsibility lies entirely with the teacher to detect and respond to when students are off-task. Teachers cannot observe and interact with every student at the same time. By contrast, an off-task behavior detector built into the learning environment can observe every student at every moment. So long as the software's

response takes into account the possibility of errors, sensitive adaptations to each student's degree of off-task behavior become possible, with the potential of substantially improving students' learning experiences and outcomes.

In addition, recent work to detect and improve students' motivation and affect [e.g. 16,18] may benefit from information on when a student is off-task, since off-task behavior is likely to be related to motivation and affect.

In this paper, we present a machine-learned model which can determine whether a student is off-task, using only data from students' actions within the software – the model uses no audio or video data. We compare our model to a model which simply treats idle time as off-task, and show that the machine-learned model is more accurate. Then, we analyze the features that make up the model, in order to understand the model better.

Next we examine data from attitudinal and motivational surveys, in order to see what factors are associated with the choice to spend more or less time off-task. We also compare these factors to the factors associated with the choice to game the system ("attempting to succeed in an educational environment by exploiting properties of the system rather than by learning the material and trying to use that knowledge to answer correctly"). Gaming the system in Cognitive Tutors consists of behaviors such as systematic guessing and persistent overuse of hints, and has also been shown to be significantly associated with poorer learning [6,7,8]. We conclude with a discussion of potential ways that interactive systems can respond to a student going off-task, considering in particular the challenge of responding in a way that does not reduce off-task behavior at the cost of an increase in gaming the system.

DATA

Data from five studies, conducted between 2003-2005, was used in our investigation of students' off-task behavior as they used Cognitive Tutor software.

Each of the studies presented in this paper was conducted in mathematics classrooms using Cognitive Tutor software, a popular type of interactive learning environment now used by around half a million students a year in the USA. Cognitive Tutor curricula combine conceptual instruction delivered by a teacher with problem-solving where each student works one-on-one with a cognitive tutoring system which chooses exercises and feedback based on a running model of which skills the student possesses [2]. A screen shot of a Cognitive Tutor is shown in Figure 1.

Each study was conducted in the Pittsburgh suburbs, within classrooms that had used Cognitive Tutors within their regular curriculum for several months. None of the studies involved gifted or special needs students. Three of the studies involved a tutor lesson on scatterplots; the other two studies involved tutor lessons on percents and geometry.

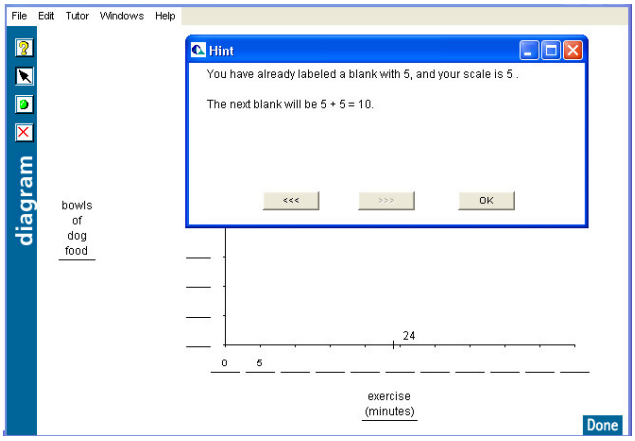

Figure 1: A screenshot from a Cognitive Tutor lesson.

The five studies shared the following general design. Each student in each of the five studies first viewed conceptual instruction on the upcoming tutor lesson, delivered via a PowerPoint presentation with voiceover and some simple animations. After viewing conceptual instruction, each student used the tutor for around 80 minutes (across 2 or 3 class periods).

Data about learning was collected using pre- and post-tests given before and after the students used the Cognitive Tutor. Two essentially isomorphic problems were constructed for the tests, for each lesson. Each problem was used as a pre-test for half of the students, and as a post-test for the other half. The problems were designed to exercise the key skills involved in the lesson (approximately six per lesson), and were graded in terms of how many of the skills a student successfully demonstrated. The test items used in each study are given in [4].

Within each of the studies, quantitative field observations [cf. 7] were used to assess each student's frequency of off-task behavior. Each student's behavior was systematically observed a number of times (around 8-10 times in total) during the course of multiple class periods, by one of three observers. In each 20 second observation, student behavior was coded as corresponding to one of a set of categories, including off-task behavior, on-task conversation, working in the tutor, and gaming the system [cf. 7]. Off-task behavior included off-task conversation (talking about anything other than the subject material), off-task solitary behavior (any behavior that did not involve the tutoring software or another individual, such as reading a magazine or surfing the web), and inactivity (such as staring into space, or the student putting his/her head down on the desk, for at least 20 seconds – brief reflective pauses by a student actively using the software were not counted as off-task). Gaming the system was not treated as a type of off-task behavior; within the observations, it was a separate category.

Across studies, most of the observations were carried out by a single observer. However, an inter-rater reliability session was carried out in 2004. In this session, two observers classified the same student at the same time. Inter-rater agreement as to whether a behavior was off-task, gaming the system, or other categories of behavior was reasonably high – Cohen's [15] $\kappa = 0.74$.

In addition, within two of the studies, motivational and attitudinal questionnaires were given to increase understanding of why students choose to game the system. In this paper, we will use these questionnaires to help us understand why students decide to engage in off-task behavior, and to compare between the motivations associated with off-task behavior and gaming the system.

A final source of data that we will use to understand off-task behavior is data from student log files as the students used the tutoring software. Across the five studies, 429 students performed between 50 and 500 actions in the tutor in each lesson, for a total of 128,887 tutor actions (due to data loss, data from 11 other students could not be used). For each student action recorded in the log files, a set of 26 features describing that student action were distilled. These features included

- Details about the action, such as whether the action was correct, the type of interface widget involved, and whether this was the student's first attempt on the problem step
- Assessment of the probability the student knew the skill involved in the action
- A hybrid feature (nonintuitively called "pknowretro") previously found useful for modeling gaming behavior [cf. 5] – pknowretro is the probability the student knew the skill if that probability changed on the current action (the first opportunity to practice the current skill on the current problem step), and -1 otherwise.
- Time taken, considered in three fashions
    - How many seconds the action took.
    - The time taken for the action, expressed in terms of the number of standard deviations this action's time was faster or slower than the mean time taken by all students on this problem step, across problems.
    - The time taken in the last 3, or 5, actions, expressed as the sum of the numbers of standard deviations each action's time was faster or slower than the mean time taken by all students on that problem step, across problems. (two variables)
- Details about relevant previous interactions, including the number of errors and help requests the student made on this problem step across problems, and how many recent actions involved this problem step.

The full list of features is given in [4].

MODELING OFF-TASK BEHAVIOR

Model Structure

Latent Response Models [22] were used as the statistical basis for all of the detectors of off-task behavior discussed in this paper. Latent Response Models have the advantage of easily and naturally integrating multiple data sources, at different grain sizes, into a single model. In addition, they were used as the basis of successful detectors of gaming behavior, within the same data [5,6].

A detector of off-task behavior, in the framework used here (shown in Figure 2), has one observable level and two hidden ("latent") levels. In a behavior detector's outermost/observable layer, the detector assesses how frequently each of n students is off-task; those assessments are labeled $OT'_0 \ldots OT'_n$. The detector's assessments for each student can then be compared to the observed proportions of time each student spent off-task, $OT_0 \ldots OT_n$.

The proportion of time each student spends off-task is assessed as follows: First, the detector makes a (binary) assessment as to whether each individual student action (denoted SA'_m) is off-task. From these assessments, $OT'_0 \ldots OT'_n$ are derived by taking the percentage of actions which are assessed to be off-task, for each student.

An action is assessed to be off-task or not, by a function on parameters composed of the features drawn from each action's characteristics. Each parameter in a candidate model is either a linear effect on one feature (a parameter value α_i multiplied by the corresponding feature value X_i – $\alpha_i X_i$), a quadratic effect on one feature (parameter value α_i multiplied by feature value X_i, squared – $\alpha_i X_i^2$), or an interaction effect on two features (parameter value α_i multiplied by feature value X_i, multiplied by feature value X_j – $\alpha_i X_i X_j$).

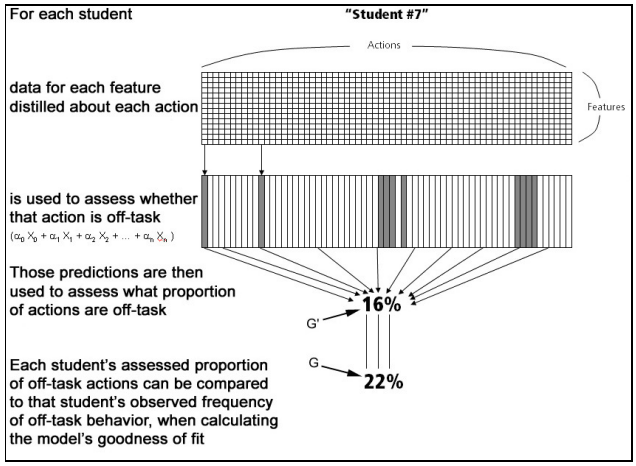

Figure 2: The architecture of the off-task behavior detector.

An assessment SA_m as to whether action m is off-task is computed as $SA_m = \alpha_0 X_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n$, where α_i is a parameter value and X_i is the data value for the corresponding feature, for this action, in the log files. Each assessment SA_m is then thresholded using a step function, such that if $SA_m \leq 0.5$, $SA'_m = 0$, otherwise $SA'_m = 1$. This gives us a set of classifications SA'_m for each action within the tutor, which are then used to create the assessments of each student's proportion of off-task behavior, $OT'_0 \ldots OT'_n$. We can then assess a model's goodness of fit by calculating the correlation between $OT'_0 \ldots OT'_n$, and the original observed data $OT_0 \ldots OT_n$.

Time-Only Modeling

Within this (or any) modeling approach, the simplest and most straightforward way to determine whether a student is off-task is probably to set a cut-off on how much time an action should take and treat all actions that take longer than that cut-off as off-task. Approaches similar to this have been used to determine whether a student is attempting to complete a problem by guessing [9], and to determine if a student is reading hints carefully [cf. 24]. Interestingly, that prior work viewed time in the opposite fashion than would be appropriate in this situation, looking for actions shorter than a time cut-off, rather than actions longer than a time cut-off.

If we use a single-parameter model which determines if an action is off-task, using only the time taken for that action, it fits well to the data, achieving a correlation of 0.47 (between the model's predictions of each student's off-task frequency and the frequency found in the original observations). According to this model's best-fitting parameter value, actions which take longer than 80 seconds are off-task. This model is also not overfit; a 10-fold student-by-student cross-validation achieves an average correlation of 0.44 across test sets.

Alternatively, it may be that we can get a better fit by taking the average time for each problem step into account (since it is conceivable that some students may legitimately need 80 seconds to input an answer on specific steps that are quite difficult or involve considerable calculation). Hence, we can set up a single-parameter model which determines whether an action is off-task, using the time taken for that action, expressed in terms of the number of standard deviations the action's time was faster or slower than the mean time taken by all students on the relevant problem step, across problems. This model also fits well to the data, achieving a correlation of 0.46. According to this model's best-fitting parameter value, actions which take more than 3.8 standard deviations longer than normal are off-task. This model is also not overfit; a 10-fold cross-validation achieves an average correlation of 0.45 across test sets.

Hence, it appears that it is possible to create a useful model of off-task behavior just by considering the time taken on each action. Considering each action's time in the context of the distribution of time students take on the relevant problem step does not appear to perform substantially better than considering time in an absolute fashion.

However, it may be that off-task behavior manifests itself in a more complex fashion than this within the tutoring environment. In the next section, we will consider whether a model trained using a fuller set of features can detect off-task behavior better than a model based just on each individual action's time.

Multiple-Feature Models

Model Selection Process

Within the model structure described above, there is a very large space of potential models that may potentially describe student behavior (if any model with 1-7 parameters is permitted, approximately 10^{13} models are possible). A combination of Fast Correlation-Based Filtering [30][1] and Forward Selection [26] was used in order to efficiently search this space of models, as follows: First, a set of single-parameter models were selected, such that:

1. Each single-parameter model was at least 60% as good as the best single-parameter model found (in terms of linear correlation to the observed data).
2. If two parameters had a closer correlation than 0.7 to each other, only the better-fitting single-parameter model was used.

Once a set of single-parameter models was obtained, each model was expanded, by repeatedly adding the potential parameter that most improved the linear correlation between the model's assessments and the original data, using Iterative Gradient Descent [11] to find the best value for each candidate parameter. Parameters were added to the model until adding a parameter worsened the model's performance in a student-by-student 10-fold cross-validation. 10-fold cross-validation is equivalent to doing a training set/test set validation ten times. Pseudocode of this algorithm can be found in [4].

Model Accuracy

The best-fitting multiple-parameter model fits well to the data, achieving a correlation of 0.62. The model is mildly overfit; a 10-fold cross-validation achieves an average correlation of 0.55 across test sets. However, while there is some decrease in performance in cross-validation, the cross-validated performance of this model is substantially better than the single-parameter time-only models (which had cross-validated correlations of 0.44 and 0.45). This

[1] In the implementation of Fast Correlation-Based Filtering used within the research presented here, linear correlation is used as the goodness-of-fit measure rather than entropy, as the overall model architecture is based on linear correlation.

indicates that the full model is likely a better explanation of the data than the time-only models.

Overall, then, the multiple-parameter model is effective at determining how much each student is off-task. In addition, this model is effective at determining how much each student is off-task, relative to other students. If the observers found that student A was off-task more often than student B, the multiple-parameter model agreed 83% of the time.

Model Details

The best-fitting multiple-parameter model is made up of six parameters. We will discuss these parameters in the order they were selected by the model; in the framework used here, each parameter after the first parameter must be understood in the context of the parameters already selected. The full model is given in Table 1.

The first parameter involves very fast actions immediately before or after very slow actions. This represents the fact that consistent very slow actions may indicate being off-task, but may also indicate careful thought or even asking the teacher for help. Careful thought or asking for help would probably not lead the student to work extremely quickly right before or after a long, thoughtful action. Hence, slow actions right before or after fast actions is more indicative of off-task behavior than slow actions alone. Taken alone, this parameter, when 10-fold cross-validated, achieves a correlation of 0.483 to the frequency of off-task behavior; hence, it already performs better than a model which labels all actions longer than a cut-off as off-task.

The second parameter indicates that if the current action is extremely slow or extremely fast, the evidence that it is an off-task action is even stronger. This feature is somewhat similar to the single-feature models considered above. The combination of the first two parameters achieves a (cross-validated) correlation of 0.522 to the frequency of off-task behavior; hence, 0.039 additional correlation is obtained by adding this parameter to the model.

The third parameter identifies specific situations where off-task behavior is more or less likely. A student is less likely to go off-task when they are inputting a string, and know the step well (inputting a string corresponds to selecting problem features, for example which variable to place on a graph). A student is more likely to go off-task when they are inputting a string, and have already made an error. Adding this parameter to the model adds 0.023 to the model's cross-validated correlation.

The fourth through six parameters together add only 0.009 to the model's cross-validated correlation. The fourth parameter indicates that repeated help-requests are not off-task behavior, regardless of how fast or slow they are. The fifth parameter indicates that two or more errors or help requests in a row are associated with off-task behaviors. Because the fourth parameter is already in the model, this parameter likely focuses on errors, suggesting that some level of carelessness may be associated with off-task behavior. (Note that two errors do not make up a pattern of systematic guessing as seen in 'gaming the system' [5]). The sixth parameter, making many errors on skills students generally know before starting the current tutor lesson, also seems to be indicate a general pattern of carelessness.

Overall, then, off-task behavior occurs in the tutor not just as slow actions, but as co-occurrence of very slow and very fast actions. In terms of student motivation, off-task behavior appears to be associated with careless actions, and possibly also with avoiding help [cf.1]. This pattern of behavior, though it has some commonalities with the knowledge-engineered time-only model of off-task behavior, represents off-task behavior in a more subtle fashion than the time-only model, and thus adds to our understanding of off-task behavior in a way that model cannot.

	param 1	param 2	value	Interpretation	Additional cross-val correlation
F1	timelast3SD	timelast5SD	-0.08	OT: Very fast actions immediately before or after very slow actions	0.483
F2	timeSD	timeSD	0.013	OT: Extremely fast actions or extremely slow actions	0.039
F3	string	pknowretro	-0.36	OT: Less likely on well-known string-input steps OT: More likely when inputting a string after error	0.023
F4	notfirstattempt	recent8help	-0.38	Not OT: Asking for a lot of help	0.004
F5	notright	pknowretro	-0.16	OT: Two errors or help-requests in a row Not OT: Errors or help requests on skills the student has already mastered	0.004
F6	pctwrong	generally-known	0.04	OT: Indicated by many errors on skills students generally know prior to starting this lesson	0.001

Table 1. The model of off-task behavior (OT). In all cases, param1 is multipled by param2, and then multipled by value.

Does this model distinguish on-task conversation from off-task behavior?

One important goal for a model of off-task behavior is that it should effectively distinguish off-task behavior from other types of behavior that occur outside of the system – for example, on-task conversation (defined as talking to the teacher or another student about the subject material or the tutoring system). A sophisticated system should not respond in the same way to a student asking a peer or the teacher for help, as it would to a student going off-task. However, there is some risk that a system may not be able to distinguish between these categories of behavior just from log files of the student's behavior within the tutor.

Fortunately, data is available to investigate whether the model of off-task behavior can distinguish these behavioral categories. Talking to the teacher or another student about the material was one of the categories of behaviors coded within the original observations, in each of the studies.

Within the model, there is some correlation between observed on-task conversation and the model's predictions of off-task behavior for each student, $r= 0.16$, $F(1,427)=11.32$, $p<0.001$. Since there is no correlation between these two categories of behavior in the observational data, $r= -0.04$, this is evidence that the model does not completely distinguish between these categories of behavior. However, the correlation between observed on-task conversation and the model of off-task behavior is much lower than the cross-validated correlation between the model of off-task behavior and the observed off-task behavior ($r=0.16$ versus $r=0.55$), $t(426)=6.85, p<0.001$, for a test of the significance of the difference between two correlation coefficients for correlated samples. Hence, the model of off-task behavior does appear to successfully distinguish between these two categories of behavior, but does not achieve complete success in doing so.

Interestingly, the model of off-task behavior that relies only upon a time cutoff (all actions longer than 80 seconds are off-task) appears to do a worse job of distinguishing between on-task conversation and off-task behavior, than the full model of off-task behavior does. The time-cutoff model correlates significantly to the frequency of on-task conversation, $r=0.22$, $F(1,427)=20.97$, $p<0.001$. This model's correlation to on-task conversation is marginally significantly higher than the full model's correlation to on-task conversation, $t(426)=1.84$, $p=0.07$. This suggests that more sophisticated models of off-task behavior not only capture those behaviors better, but are more successful at discerning the difference between off-task behavior and other behaviors which involve idle time, such as on-task conversation. This may be because the machine-learned model takes behavioral correlates (the third to sixth features in the model) into account.

Hence, it appears that the model of off-task behavior captures considerably more off-task behavior than on-task conversation, and does better at distinguishing between these behaviors than a simple time-only model of off-task behavior does. Some on-task conversation is still captured by the model, though – therefore, any system re-design which uses this model to detect and respond to off-task behavior will need to take this possibility into account.

OFF-TASK BEHAVIOR, AND MOTIVATION/ATTITUDES

Methods

Data from two self-report questionnaires was used to study the relationship between students' motivations and attitudes, and their frequency of off-task behavior.

All items on both questionnaires were drawn from existing motivational inventories or from items used across many prior studies with students from the relevant age group, and were adapted minimally (for instance, the words "the computer tutor" was regularly substituted for "in class", and some items were changed from first-person to second-person for consistency). Both questionnaires were given to students along with their unit pre-tests, before they worked through a Cognitive Tutor lesson on scatterplots (all students who received the first questionnaire, half of the students who received the second questionnaire) or percents (half of the students who received the second questionnaire). All items were given as 6-point Likert scales, except for a small number of multiple-choice and true-false items.

In order to analyze the relationship between student motivations/attitudes and off-task behavior, we correlated students' responses on the questionnaires to their frequency of off-task behavior, as assessed by the model of off-task behavior presented in this paper. It is advantageous to use the model's assessments of off-task behavior rather than the classroom observations, because the model of off-task behavior's assessments are more precise than the classroom observations. 2-3 researchers can only obtain a small number of observations of each student's behavior, and thus the estimations of each student's frequency of off-task behavior have high variance. By contrast, the model, with access to predictions about every student action, can make considerably more precise predictions.

We also compare the relationship between off-task behavior and student responses to the relationship between gaming the system and student responses, in order to better understand the relationship between these two categories of behavior. We focus this discussion on "harmful" gaming, which occurs on steps the student finds difficult. Other students game time-consuming but easy steps, in order to focus time on more challenging material [cf. 5,6] – this strategic behavior is not associated with poorer learning, and does not appear to be associated with poor motivation or negative attitudes towards the learning context [cf. 8].

Questionnaire One

Questionnaire Constructs

The first questionnaire, given in Spring 2004, is discussed in complete detail in [8]. This questionnaire consisted of items measuring:

- Whether the student had performance goals or learning goals [cf. 23]
  (Example: "We are considering adding a new feature to the computer tutors, to give you more control over the problems the tutor gives you. If you had your choice, what kind of problems would you like best?
  A) Problems that aren't too hard, so I don't get many wrong.
  B) Problems that are pretty easy, so I'll do well.
  C) Problems that I'm pretty good at, so I can show that I'm smart.
  D) Problems that I'll learn a lot from, even if I won't look so smart.") [e.g. 23]
- The student's level of anxiety about using the tutor (Example: "When you are working problems in the tutor, do you feel that other students understand the tutor better than you?") [eg. 20]
- The student's level of anxiety about using computers
  (Example: "When you use computers in general, do you feel afraid that you will do something wrong?") [eg. 20]
- How much the student liked using the tutor (Example: "How much fun were the math problems in the last computer tutor lesson you used?") [e.g. 23]
- The student's attitude towards computers (Example: "How much do you like using computers, in general?") [e.g. 19]
- If the student was lying or answering carelessly on the questionnaire – such "lie scale" items are designed such that anyone answering thoughtfully and honestly would never give one of the answers.
  (Example: "Is the following statement true about YOU? 'I never worry what other people think about me.' TRUE/FALSE") [e.g. 27]

Relations to off-task behavior

As shown in Table 2, of the quantities assessed in the first questionnaire study, only disliking computers was significantly associated with off-task behavior, $F(1,100)=5.06$, $p=0.03$, $r=0.22$. Interestingly, disliking computers is also associated with gaming the system in the harmful fashion [8], $F(1,100)=3.94$, $p=0.05$, $r=0.19$. None of the other quantities assessed in the first questionnaire study had correlations which were significantly different than chance – the closest was anxiety about using computers, $F(1,100)=1.60$, $p=0.21$, $r=0.13$.

Questionnaire Two

Questionnaire Constructs

The second questionnaire, given in Spring 2004, is discussed in complete detail in [4]. This questionnaire consisted of items measuring:

- If the student believes that computers in general, and the tutor in specific, are not very useful.
  (Example: "Most things that a computer can be used for, I can do just as well myself.") [e.g. 28]
- If the student believes that computers/the tutor don't/can't really care how much he/she learns. (Example: "I feel that the tutor, in its own unique way, is genuinely concerned about my learning.") [e.g. 10]
- If the student has a tendency towards passive-aggressiveness [25]
  (Example: "At times I tend to work slowly or do a bad job on tasks I don't want to do") [e.g. 25]
- If the student believes that computers/the tutor reduce his/her sense of being in control
  (Example: "Using the tutor gives me greater control over my work") [e.g. 17]
- If the student is not educationally self-driven
  (Example: "I study by myself without anyone forcing me to study.") [e.g. 21]
- If the student dislikes math
  (Example: "Math is boring") [e.g. 21]

Relations to off-task behavior

As shown in Table 3, two of the quantities assessed in the second questionnaire were significantly associated with off-task behavior: the student disliking math, $F(1,92)=6.97$, $p=0.01$, $r= 0.27$, and the student having a tendency towards passive-aggressive behavior, $F(1,92)=3.93$, $p=0.05$, $r= 0.20$. Another quantity was marginally significantly associated with off-task behavior: a lack of educational self-drive, $F(1,92)=2.74$, $p=0.10$, $r= 0.17$. Curiously, educational self-drive and disliking mathematics have also been found to be

	Performance Goals	Anxiety About Using Computers	Anxiety About Using the Tutor	Lying/ Answering Carelessly	Disliking Computers	Disliking the Tutor
Off-Task Behavior	0.11	0.13	0.04	-0.03	**0.22**	0.12
Gaming the System (harmful fashion)	0.00	-0.02	-0.04	0.06	**0.19**	**0.20**

Table 2. Relationships between the categories in the first questionnaire, and off-task behavior, as assessed by the model. Statistically significant relationships ($p<0.05$) are in boldface.

	Belief that Computers/ the Tutor are not useful	Belief that Computers/ the Tutor are uncaring	Tendency towards passive-aggressiveness	Belief that Computers/ the Tutor reduce control	The student is not self-driven	Disliking math
Off-Task Behavior	0.02	-0.03	**0.20**	0.00	*0.17*	**0.27**
Gaming the System (harmful fashion)	0.16	0.13	0.10	0.04	**0.25**	**0.21**

Table 3. Relationships between the categories in the second questionnaire, and off-task behavior, as assessed by the model. Statistically significant relationships ($p<0.05$) are in boldface; marginally significant relationships ($p<0.10$) are in italics.

associated with the choice to game the system in the harmful fashion [4].

Overall Pattern and Implications

Overall, off-task behavior is associated with disliking computers, disliking mathematics, passive-aggressiveness, and not being educationally self-driven. This pattern is quite similar to the pattern of attitudes in students who game the system in a fashion associated with poorer learning. Those students dislike computers, dislike the tutoring software, dislike mathematics, and are not educationally self-driven. It is somewhat curious that passive-aggressiveness is associated with off-task behavior, rather than gaming the system. Gaming the system would seem, at some level, to be related to "doing a bad job on a task I don't want to do" – however, gaming can also be seen an attempt to succeed in an undesirable task without having to put full effort into that task, rather than an attempt to intentionally perform poorly or work more slowly.

One possible explanation for the overall commonalities in the attitudes associated with off-task behavior and harmful gaming is that the same students engage in both behaviors – i.e. students who spend time off-task also game the system in the harmful fashion. However, the two behaviors are, if anything, negatively correlated with each other. Across the five studies, the frequency of harmful gaming and off-task behavior in each student's actions (each assessed by the relevant detector) are negatively correlated, $F(1,427)=8.22$, $p<0.01$, $r= -0.14$. If anything, this trend was stronger within the students for whom we have questionnaire data, $F(1,211) = 9.92$, $p<0.01$, $r = -0.21$.

The negative correlation between the two behaviors, combined with the similarity in the motivations and attitudes associated with the two behaviors, suggests that the choices to game the system or go off-task arise from relatively similar motivations but that some other factor leads students to choose between these two approaches.

One possibility is that this factor may be the degree to which the students perceive the current tutor lesson as difficult. Students game harmfully predominantly on steps they know poorly [5,6], whereas there appears to be little relationship between student knowledge and the choice to go off-task, as shown in Table 1.

Another possibility is that the students' relationship with their teacher may influence this choice. It may be that students who feel positively towards their teachers, and want their teacher to approve of them, game the system rather than engaging in more noticeable behaviors such as talking off-task or surfing the web (or asking the teacher for help, which would show lack of knowledge, and potentially cause the teacher to think less well of them). By contrast, students who feel more negatively towards their teacher may have less desire to avoid being seen off-task.

Another possibility is that students systematically differ in whether they prefer gaming the system or going off-task, for reasons that are not explicitly attitudinal or motivational. One possibility is that students learn over time that their teachers or parents respond better to one of these behaviors than the other, and adopt the behavior which they have previously found more successful, when working in the Cognitive Tutor. It is also possible that personality factors such as extraversion play a role – for example, more extroverted students may prefer to talk to their neighbors than interact with the system when they are unmotivated.

The similarity between the attitudes and characteristics associated with off-task behavior and gaming the system is striking – especially when the lack of correlation between the behaviors themselves is taken into account. In the long term, we will understand both behaviors better when we can identify what factors differentiate between the students who engage in each type of behavior.

Responding to Off-Task Behavior

Knowing which student characteristics and attitudes are associated with off-task behavior is a good start towards developing systems that can respond appropriately when a student is off-task. One important implication of our results is that off-task behavior is likely more than just evidence that a system is badly designed; instead, it is likely to be associated with deeper motivational problems. In addition, evidence that off-task behavior stems from similar motivations as gaming the system suggests that the

possibility of students switching from off-task behavior to gaming should be seriously considered in the design of system responses to off-task behavior.

In particular, redesigning systems to respond immediately and in a heavy-handed way – for example, by making a loud noise when a student is off-task – are likely to be counterproductive. Re-designing systems in this fashion may actually lead students to game the system in order to avoid the system's intervention. For example, a student might learn to type in an answer – any answer – every 20 seconds so that the system thinks he or she is actively working. In addition, heavy-handed solutions are likely to irritate a student who is off-task, and irritate students even more when the model is incorrect and the student was not off-task (which will occur some proportion of the time, since the model is not perfectly accurate).

Instead, it may be more appropriate to respond to off-task behavior with more long-term oriented, non-heavy-handed solutions. One more constructive response to off-task behavior may be to use self-monitoring, where a student is led to monitor their own on and off-task behavior– this approach has been shown to reduce off-task behavior in traditional classrooms [cf. 15] and may be feasible and effective in interactive learning environments as well. Alternatively, it may be possible to increase challenge when students go off-task, or to give rewards to students who correctly complete problems quickly without gaming the system. Rather than reducing off-task behavior by increasing gaming behavior, such an approach may even be able to remediate both off-task behavior and gaming the system at the same time, an important step towards interactive learning environments that can respond sensitively to the full spectrum of ways students choose to interact with them.

CONCLUSIONS

In this paper, we have presented a model that can automatically detect, with reasonable effectiveness, when a student is off-task in a Cognitive Tutor. This model does not rely upon sophisticated instrumentation which is unavailable in most school computer labs, such as microphones, eye-trackers, or fMRI – it relies only upon data about students' actions within the tutoring system. We have shown that this model is more accurate than a simpler approach which treats all actions longer than a certain cutoff as off-task, both at determining each student's frequency of off-task behavior, and in distinguishing off-task behavior from on-task conversation, a category of behavior which – like off-task behavior – involves idle time. The methods used to develop this model may be relevant for detecting off-task behavior in other types of interactive systems; idle time alone is generally likely to be less accurate than detecting idle time in combination with behavioral correlates. The model's accuracy is not perfect, but is likely to be sufficiently effective to drive system adaptation, so long as the system adaptation is thoughtfully designed.

We then analyzed what student attitudes, motivations, and characteristics are associated with off-task behavior, using the detector in combination with questionnaire data. We determined that off-task behavior is associated with disliking computers, disliking mathematics, passive-aggressiveness, and lack of educational self-drive.

These student attitudes and characteristics are very similar to the attitudes and characteristics found in earlier research to be associated with gaming the system – an especially surprising result in the light of the negative correlation between gaming the system and off-task behavior. One possibility is that the two behaviors are different responses to the same motivation. A student's decision of which behavior to use may interact with the student's prior learning experiences, specifics of the learning situation (such as the presence or absence of material that student finds particularly difficult), their relationship with the teacher, or personality characteristics not measured in the questionnaires.

Future work will be needed to determine why some students choose to go off-task, while others choose to game the system. Understanding the answer to this question may enable the development of systems that can respond appropriately to both of these student behaviors.

ACKNOWLEDGMENTS

I would like to thank Ido Roll, Albert Corbett, Ken Koedinger, and Angela Wagner for the significant role they played in the original design, administration, and analysis of the studies re-analyzed here, as well as their helpful comments and suggestions on this paper. I would also like to thank Jason Walonoski and the anonymous reviewers for their very helpful comments and suggestions.

REFERENCES

1. Aleven, V., McLaren,B.M., Roll, I., and Koedinger, K.R. Toward tutoring help seeking: Applying cognitive modeling to meta-cognitive skills. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems (ITS 2004),* 227-239.
2. Anderson, J.R., Corbett, A.T., Koedinger, K.R., Pelletier, R. Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences 4,* 2 (1995), 167-207.
3. Amershi, S., and Conati, C. Automatic Recognition of Learner Groups in Exploratory Learning Environments. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems (ITS 2006),* 463-472.
4. Baker, R.S. (2005) Designing Intelligent Tutors That Adapt to When Students Game the System. Doctoral Dissertation. *Carnegie Mellon University Technical Report CMU-HCII-05-104.*

5. Baker, R.S., Corbett, A.T., and Koedinger, K.R. Detecting Student Misuse of Intelligent Tutoring Systems. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems (ITS 2004),* 531-540.
6. Baker, R.S.J.d., Corbett, A.T., Roll, I., Wagner, A.Z., and Koedinger, K.R. The Relationship Between Gaming the System and Learning in Cognitive Tutor Classrooms. Manuscript Under Review.
7. Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z. Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game the System". *Proceedings of ACM CHI 2004: Computer-Human Interaction*, 383-390.
8. Baker, R.S., Roll, I., Corbett, A.T., Koedinger, K.R. Do Performance Goals Lead Students to Game the System? *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED 2005),* 57-64.
9. Beck, J.E. Engagement tracing: using response times to model student disengagement. *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED 2005),* 88-95.
10. Bickmore, T.W., Picard, R.W. Towards Caring Machines. *CHI Extended Abstracts* (2004)*,* 1489-1492.
11. Boyd, S., and Vandenberghe, L. *Convex Optimization.* Cambridge University Press, Cambridge, UK, 2004.
12. Carroll, J. A Model For School Learning. *Teachers College Record 64* (1963), 723-733.
13. Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*, 1 (1960), 37-46.
14. Collins, R.T., Lipton, A.J., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N., Hasegawa, O., Burt, P., Wixson, L. *Carnegie Mellon University Technical Report CMU-RI-TR-00-12: A System for Video Surveillance and Monitoring* (2000).
15. Dalton, T., Martella, R.C., and Marchand-Martella, N.E. The effects of a self-management program in reducing off-task behavior. *Journal of Behavioral Education 9,* 3-4 (1999), 157-176.
16. de Vicente, A., and Pain H. Informing the detection of the students' motivational state: an empirical study. *Proceedings of the 6th International Conference on Intelligent Tutoring Systems (ITS 2002),* 933-943.
17. Dillon, T.W., Garner, M., Kuilboer, J., Quinn, J.D. Accounting Student Acceptance of Tax Preparation Software. *Journal of Accounting and Computers 13* (1998), 17-29.
18. D'Mello, S.K., Craig, S.D., Gholson, B., Franklin, S., Picard, R.W., and Graesser, A.C. Integrating Affect Sensors in an Intelligent Tutoring System. *Affective Interactions: The Computer in the Affective Loop Workshop at 2005 International Conference on Intelligent User Interfaces,*7-13.
19. Frantom, C.G., Green, K.E., Hoffman, E.R. Measure Development: The Children's Attitudes Towards Technology Scale (CATS). *Journal of Educational Computing Research 26,* 3 (2002), 249-263.
20. Harnisch, D.L., Hill, K.T., Fyans, L.J. Development of a Shorter, More Reliable, and More Valid Measure of Test Motivation. Paper presented at the 1980 annual meeting of the National Council on Measurement in Education. ERIC Document #ED193273.
21. Knezek, G., Christensen, R. *Computer Attitudes Questionnaire* (1995). Denton, TX: Texas Center for Educational Technology.
22. Maris, E. Psychometric Latent Response Models. *Psychometrika 60,* 4 (1995), 523-547.
23. Mueller, C.M., and Dweck, C.S. Praise for Intelligence Can Undermine Children's Motivation and Performance. *Journal of Personality and Social Psychology 75,* 1 (1998), 33-52.
24. Murray, R.C., and vanLehn, K. Effects of Dissuading Unnecessary Help Requests While Providing Proactive Help. *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED 2005),* 887-889.
25. Parker, G., Hadzi-Pavlovic, D. A Question of Style: Refining the Dimensions of Personality Disorder Style. *Journal of Personality Disorders, 15,* 4 (2001), 300-318.
26. Ramsey, F.L., Schafer, D.W. *The Statistical Sleuth: A Course in Methods of Data Analysis.* Duxbury Press, Belmont, CA, USA, 1997.
27. Sarason, S.B. *Anxiety in Elementary School Children: A Report of Research.* Greenwood Press, Westport, CT, USA, 1978.
28. Selwyn, N. Students' Attitudes Towards Computers: Validation of a Computer Attitude Scale for 16-19 Education. *Computers & Education, 28* (1997), 35-41.
29. Suchman, L. *Plans and Situated Actions: The Problem of Human-Machine Communication.* Cambridge University Press, Cambridge, UK, 1987.
30. Yu, L., and Liu, H. Feature selection for high-dimensional data: a fast correlation-based filter solution. *Proceedings of the International Conference on Machine Learning (2003),* 856-863.
31. Ziemek, T.R. Two-D or not Two-D. Gender Implications of Visual Cognition in Electronic Games. *Proceedings of the 2006 Symposium on Interactive 3D Graphics and Games,* 183-190.