Human Classification of Low-Fidelity Replays of Student Actions

Ryan S.J.d. Baker
Learning Sciences Research Institute
University of Nottingham
rsbaker@cmu.edu

Albert T. Corbett, Angela Z.Wagner
Human-Computer Interaction Institute
Carnegie Mellon University
corbett@cmu.edu, awagner@cmu.edu

Abstract. Human observations and classifications have shown to provide substantial leverage for developing models of students' motivation, attitudes, and strategic choices, as a student interacts with an intelligent tutoring system. However, human observation and classification is highly time-consuming, which has limited its use. We present a technique for conducting human classification on "low-fidelity" text-based replays of student behavior derived from logs of tutor usage. We show that low-fidelity classification is much faster than live classification, and that low-fidelity classification is approximately as accurate as live classification for detecting a behavior known as gaming the system, using a machine-learning detector of this behavior as the gold standard.

Keywords: Observation, Classification, Gaming the System, Machine Learning, Data Mining

INTRODUCTION

In recent years, there has been increasing interest in using human observations and classifications to improve both our understanding of how students interact with intelligent tutoring systems, and to develop systems which can adapt to differences in student motivation, attitudes, and strategic choices. Human judgment can provide substantial leverage for studying how students interact with intelligent tutoring systems: human beings regularly make judgments about other humans' affective state, motivation, attitudes, and strategies in order to participate effectively in daily life. However, although humans are very good at making this sort of judgments about others, they are not as good at describing how they make these judgments, likely due to the large role that unconscious processing plays (Adolphs, 2006). Because humans are good at making judgments about affect, motivation, and strategy but not as good at explaining their decision-making process, one strategy which has recently become popular is to combine quantitative data from human observation and classification, with machine learning or data mining techniques.

In such an approach, a set of categories of interest are first developed. These may be specific categories of behavior (working with a tutor, talking off task to a neighbor, gaming the system – cf. Baker et al., 2004), categories of affect (effort, confidence, satisfaction – cf. de Vicente and Pain, 2002; boredom, frustration, and flow – cf. Craig, Graesser, Sullins, and Gholson, 2004), or even whether the user is receptive to a specific interaction (interruptible versus non-interruptible – cf. Fogarty et al., 2005). Next, a human observer observes a set of students over a period of time – either one at a time, or switching between students according to a pre-determined schedule (cf. Baker et al, 2004). In each observation, a student is classified into one of the pre-determined categories. This process of observation and classification may be done live, or it may be done later, using video recordings or screen capture.

Once classifications have been obtained, data from logs (or videotapes, or other records) of the student or user's behavior is distilled into a set of features that can be used within a machine learning algorithm. Finally, a machine learning algorithm is used to develop a detector of the categories of interest, predicting the data from the human observations using some combination of the distilled features. This approach has successfully been used in each of the cases mentioned above: to predict student affect (de Vicente and Pain, 2002), whether a student is gaming the system (Baker et al., 2004), and whether a user is interruptible (Fogarty et al., 2005).

Hence, human observation and classification has proved its value for the development of detectors which can effectively detect a wide variety of student characteristics. However, human observation and classification is highly time-consuming, a disadvantage which has limited its use. As shown in Table 1, while the official observation periods of human classifications are usually around 20 or 30 seconds, the actual time required to obtain each observation is considerably more. de Vicente and Pain (2002) report making 85 classifications in around 6 hours, an average of one classification every 4.2 minutes. Baker et al. (2004) report making 563 classifications in around 7.5 hours of class time. However, since the observations in Baker et al. were made in

Table 1. Estimations of the actual time per classification, in three prior studies using high-fidelity observations

Study	Total Time Spent Classifying (approx)	Session Logistical Time (approx)	Number of Classifications	Theoretical Classification Time	Actual Time per Classification
de Vicente and Pain 2002	6 hours	minimal	85	N/A	4.2 minutes
Baker et al. 2004	7.5 hours	6 hours	563	20 seconds	1.3 minutes
Craig et al. 2004	20 hours	8.5 hours	?	30 seconds	>5 minutes

classrooms distant from the researchers' offices, driving time and setup time should also be counted as part of the overall time cost of conducting these observations. If we assume around a half hour driving time in each direction, and 15 minutes setup time before each session, that works out to around 6 hours logistical time. Hence, while each observation took 20 seconds to conduct, the actual time cost per classification is 1.3 minutes. Craig et al. (2004) do not report exactly how many classifications were made, but since there was some setup time for each of the 34 study participants, and the observation technique used was to conduct one 30 second observation every 5 minutes, the time cost of this method was at minimum 5 minutes per classification. Hence, human observation and classification of student behavior/attitudes/affect within tutoring systems appears to have a general time cost of around 1-10 minutes per classification, depending on specifics of the categories being classified and the classification method chosen. This breakdown is shown in Table 1.

A number of factors account for the difference in time taken per classification, in the studies discussed. One factor is the study setting. Conducting studies with multiple simultaneous participants (for example, in a classroom) offers time savings over studying the same behavior in the lab. However, lab studies allow for more precision and instrumentation in measurement. In addition, classroom studies generally involve more startup costs (school recruitment, approval, and scheduling), though this can be avoided by overlaying a classification study on top of an already occurring study (as in the Baker et al study). It is also worth noting that some behaviors (such as talking off task) many not occur in the lab, and must be studied in a classroom context.

Overall, though, it appears that the benefits of human observation and classification are currently offset by the large time cost. This time cost currently hinders the large-scale utilization of these methods. While small-scale human observation studies can be useful for improving our understanding of student interactions with software (cf. Craig et al, 2004), they are less immediately applicable to the problem of developing detectors of student behaviors and attitudes that can be deployed on a large scale and across different tutor lessons or even different tutoring systems. Recent work suggests that conducting observations across multiple tutor lessons may be sufficient to train a behavior detector that could transfer between tutor lessons (Baker et al, to appear). However, even in that example, collecting enough classifications to develop and verify a generalizable detector of gaming behavior required over 60 hours of observation and logistical time.

In this paper, we will discuss and validate an alternative observational technique: observations of low-fidelity replays of student behavior. We investigate this technique within the context of studying whether students are "gaming the system", a behavior for which we already have a considerable amount of observational data and a validated machine-learned detector (Baker et al, to appear). We find that the individual classifications obtained through low-fidelity replays have lower inter-rater reliability than individual classifications obtained through live observations; however, the aggregate predictions made about each student's frequency of gaming agree equally well with the machine-learned model as the live classifications (used to train the model) do.

FIDELITY

Each of the three previous studies discussed here involved "high-fidelity" observations, meaning that the observer had access to a very broad range of data to use when making inferences. For instance, in both the Craig et al. and Baker et al. studies, the observer was physically co-located with the observed participant. Hence, the observer could watch the observed student's actions in the user interface in considerable detail: not just what the participant entered as answers, but partial responses, backspacing, pauses in the middle of responding, switching windows, and mouse movements. Additionally, the observer could view the observed participant's posture, facial expression, and gestures, and could hear whatever comments or exclamations the observed participant made. In the de Vicente and Pain study, observations were not co-located but were based on exact replays of the student's screen. The observer could therefore watch the observed student's actions in the user interface in the same detail as in the co-located studies, but did not have access to potential data from the observed student's behavior outside of the user interface (e.g. posture, facial expressions, gestures, comments, and exclamations).

There is a fairly wide space for what data can be available to the human observer, beyond what was utilized in these three studies. A simple model of this space is shown in Figure 1.

Between live co-located observation and exact screen replays (in terms of fidelity), there are exact video replays. Exact video replays show the same overall data as live co-located observation, but only from a single

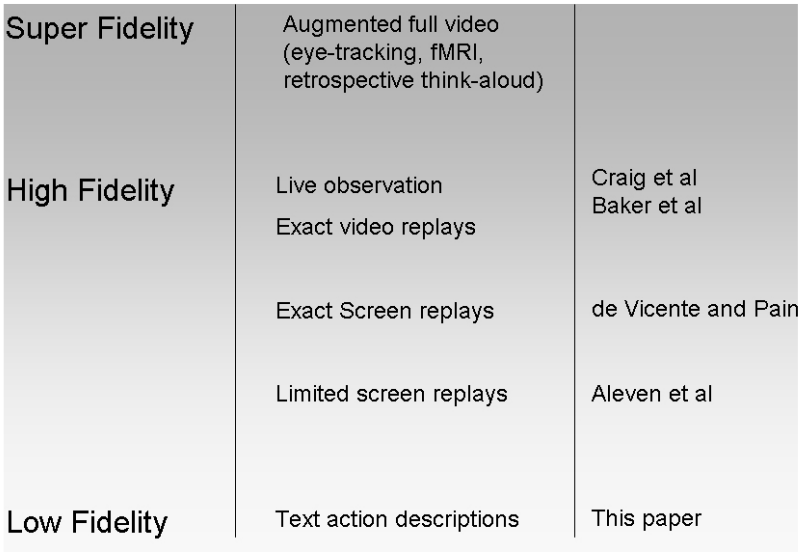

Fig. 1. Differing degrees of fidelity in the data a human observer has at hand when making a classification.

camera angle. Whereas a live observer can adjust his/her angle to see the student's facial expression or the screen (when obscured), a single camera can sometimes be obscured. Additionally, not all utterances may be picked up by a single camera (as compared to a live observer with good hearing). On the other hand, exact video replays enable repeated re-coding and can be directly analyzed with image-processing software. Exact video replays are common in research into interactive systems (cf. Fogarty et al., 2005).

A lower-fidelity option is limited, or low-fidelity, screen replays (cf. Aleven et al, 2004). Limited screen replays attempt to show the best possible approximation of the student's behavior, given log files not explicitly designed for that purpose. For instance, many if not most intelligent tutoring systems now log each student action at the level of type of action, interface widget/problem step, and input entered. From these logs, it is possible to create an approximation of the student's screen, containing the same windows (and information, prior answers, etc., within them) as the student's screen had; it is not possible, however, to correctly represent window position or when one window occludes another. Hence the context of the student's behavior can be shown, but may include some information that was not salient to the student at the time of their action(s). Additionally, the student's actions which were directly processed by the software (such as answers or help requests) can be shown in the replay, but mouse movements and partial responses are not included. Limited screen replays provide less rich information than an exact screen replay; however, since they can be generated automatically from existing log files, they can be conducted on the pre-existing tutor data which now exists in large quantity in several intelligent tutor research groups.

An even lower-fidelity option is text replays. Text replays are generated from the same data as limited screen replays, but make little attempt to show context or represent actions as they occurred. A sequence of actions of a pre-selected duration are shown in a textual format that shows each action's time (relative to the first action in the clip), what type of action it was, the interface widget selected, the input entered, and how the system assessed the action (for instance, correct, incorrect, a help request, or an incorrect action indicating a known misconception). Other information in the logs (such as the probability the student knew the skill) can also be included. An example of a sequence of text action descriptions is shown in Figure 2. Text replays omit context such as students' previous answers, and omit visual information such as mouse movements or partial responses. It is also necessary to use an annotated printout of the user interface to interpret what widgets such as "x-axis-glb-ns" mean. Hence, text replays have a very low bandwidth of information, compared to live co-located observations or full screen replays. However, they are likely to be very quick to classify (though the cost in terms of accuracy may be high), and like limited screen replays can be generated automatically from existing log files.

Though live co-located observations is the highest-fidelity technique which is currently commonly used, they are not the highest fidelity technique possible. For example, video replays could be augmented with other information such as eye-tracking data (Jacob, 1995), Functional Magnetic Resonance Imaging (fMRI) data, or retrospective think-aloud data (Russo, Johnson, and Stephens, 1989).

IS HIGH FIDELITY NECESSARY?

High fidelity observations have proven their utility, but have also been shown to be very time-consuming. In this section, we will compare high-fidelity observations to a very low-fidelity technique, text replays, in order to clarify the relative advantages and disadvantages of using high-fidelity and low-fidelity techniques. In specific,

Fig. 2. A text replay of a student's actions

we will study a behavior using low-fidelity techniques that has already been studied using high-fidelity techniques, gaming the system, in order to compare the data which results from each technique.

Previous Studies, using High-Fidelity Observations

From 2003 to 2005, we collected data on student behavior in a set of 16 classrooms in 2 schools in the Pittsburgh suburbs. The goal of these studies was to investigate the prevalence of a set of student behaviors in intelligent tutoring systems, and the learning gains associated with those behaviors. Within these studies, we used live, co-located observations to classify a student as engaging in one of the following behaviors:

1. on-task -- working on the tutor
2. on-task conversation -- talking to the teacher or another student about the subject material
3. off-task conversation – talking about anything other than the subject material
4. 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)
5. inactivity -- for instance, the student staring into space or putting his/her head down on the desk for the entire observation period
6. gaming the system – inputting answers quickly and systematically, and/or quickly and repeatedly asking for help until the tutor gives the student the correct answer

In each of these studies, each student's behavior was observed and classified several times during the course of each class period, by one of three observers. Most of the observations involved a single observer and a single student; however during an inter-rater reliability session in 2004, two observers classified the same student at the same time. The observational method used was based on prior techniques used to code on-task and off-task classroom behavior (cf. Karweit & Slavin, 1982; Lloyd & Loper, 1986); in order to avoid bias towards more interesting or dramatic events, the coder observed the set of students in a specific order determined before the class began, as in Lloyd and Loper (1986). Any behavior by a student other than the student currently being observed was not coded. In each study, between 500 and 1000 observations were taken, with around 6-10 observations taken for each student in each study, with some variation due to different class sizes and students arriving to class early or leaving late.

Each observation lasted for 20 seconds – if a student was inactive for the entire 20 seconds, the student was coded as being inactive. If two distinct behaviors were seen during an observation, only the first behavior observed was coded. In order to avoid affecting the current student's behavior if they became aware they were being observed, the observer viewed the student out of peripheral vision while appearing to look at another student.

The observations involved a total of six tutor lessons, involving a range of tutor topics. One lesson, on creating and interpreting scatterplots of data, was used in all three years. The observational data from this set of studies was used to validate that there was a significant relationship between gaming the system and poorer learning (Baker, Corbett, Koedinger, and Wagner, 2004), and to develop a detector of gaming behavior that transferred between tutor lessons without retraining (Baker et al, to appear).

Current Study, using Low-Fidelity Observations

We conducted a study on the effectiveness of low-fidelity observations, using log file data from one of the previous high-fidelity observational studies, a 2005 study on an unmodified version of the scatterplot tutor lesson. We only investigated gaming behavior, since this behavior has already been studied in considerable depth using high-fidelity observations, allowing a clear comparison. Additionally, gaming behavior is entirely expressible in the student's actions within the tutor (by contrast to, for instance, talking off task), though the richer data accessible in live co-located observations may provide additional leverage for identifying gaming.

In the low-fidelity study, two coders coded a set of clips of student behavior. These coders were the same two coders who had conducted the majority of the observations in the previous high-fidelity observations. They were also the same two coders who conducted the inter-rater reliability session in 2004, which determined the inter-rater reliability of the high-fidelity observations. These coders coded overlapping sets of clips: Both coder A and coder B coded the same 318 clips; coder A coded an additional 273 clips. Hence, there were a total of 909 classifications made during the low-fidelity study.

It was not possible to exactly sync the low-fidelity clips with the previous high-fidelity classroom observations, since exact times were not recorded for the earlier high-fidelity observations. Hence, the clips classified were chosen as follows: For each clip, one student action (entering an answer or requesting help) was chosen at random. This action became the first action in the clip. Subsequent actions were added to the clip, in the order they occurred in the log file, until adding an action would make the clip more than 20 seconds long (20 seconds was the length of the live observations).

The clips were shown to the observers in the format shown in Figure 2.

Time Taken

The 909 classifications made by the two coders, using text replays, were conducted in approximately 2 hours and 20 minutes. Since the two coders could code at their desks on their personal computer, and could start and stop whenever they liked, there was minimal logistical time associated with conducting the classifications (time spent programming the coding system was not counted, much as time spent negotiating an observation schedule with teachers and principals is not counted in the time taken for the classroom observations). Overall, this worked out to around 9 seconds per text replay classification.

Within the 2005 high-fidelity study, around 5 hours of classification time and 3 hours of logistical time were devoted to collecting 488 observations. (In actuality, half the class was using a different tutor lesson on percents, and observations were collected for both lessons in the same session – for comparability, since half the class was using the other tutor lesson, we simply halved the total amount of classification and logistical time from that study). This works out to around 1 minute per classification, an amount in line with the 1.3 minutes per

Table 2. Estimations of the actual time per classification, in three prior studies using high-fidelity observations, and in the current study using text action descriptions

Study	Total Time Spent Classifying (approx)	Session Logistical Time (approx)	Number of Classifications	Theoretical Classification Time	Actual Time per Classification
de Vicente and Pain 2002	6 hours	minimal	85	N/A	4.2 minutes
Baker et al. 2004	7.5 hours	6 hours	563	20 seconds	1.3 minutes
Craig et al. 2004	20 hours	8.5 hours	?	30 seconds	>5 minutes
2005 high-fidelity (live observation) study	5 hours	3 hours	488	20 seconds	1 minute
Text replays	2.3 hours	minimal	909	N/A	9 seconds

observation calculated for the earlier study conducted in 2003 (Baker et al., 2004). The moderate decrease in time per observation from 2003 to 2005 suggests that the observers may have become more efficient at writing down their observations and moving on to the next student from year to year.

Overall, then, text action description observations require about 15% as much time to conduct as live co-located observations. This is a considerable gain in terms of speed; but if that gain comes at substantial cost in terms of accuracy, higher-fidelity observations may still be superior.

Consistency Measures

In the 2004 inter-rater reliability session, two coders conducted live co-located observations on the same student at the same time. In order to do this, the two observers observed the same student out of peripheral vision, but from different angles. The observers moved from left to right; the observer on the observed student's left stood close behind the student to the left of the observed student, and the observer on the observed student's right stood further back and further right, so that the two observers did not appear to hover around a single student. The two observers began and ended each observation at the same time, through hand signals.

The two observers in the 2004 inter-rater reliability session made 49 simultaneous classifications; considering solely whether a behavior was coded as gaming or not gaming, Cohen's (1960) κ was 0.83, indicating very high agreement between these two observers. Within the text action description observations, the two observers separately coded the same 318 clips. Within these 318 clips, Cohen's κ was 0.58, indicating only moderate agreement. The overall rate of gaming codes in the two conditions was comparable (6.6% versus 5.3%), validating that Cohen's κ can validly be compared, between methods.

The greater degree of agreement in the live co-located observations suggests that this observational method is generally more accurate than text replays. However, text action description observations still had a high enough degree of agreement to suggest that they may be useful for capturing behavior. Additionally, reliability concerns can be addressed by using multiple coders for text replays. Given the differences in speed, it would be possible to have three coders code every clip, and still be twice as fast as live co-located observations.

Agreement Between Methods

To some extent, how internally consistent individual low-fidelity observations are is less important than whether they accurately capture the target behavior. We can assess this in two fashions: first, by comparing the data from low-fidelity observations to the data from high-fidelity observations, and second, by comparing both to a "gold-standard" indicator of each student's gaming frequency – the predictions made by the machine-learned gaming detector.

In both cases, we will need a more distilled measure than individual observations (since individual observations are not synched between methods). To this end, we compute an estimated gaming frequency for each student, according to each of the two measures, as the number of gaming classifications divided by the total number of classifications. We can then compare these estimated gaming frequencies, between methods.

Given the incomplete agreement of the classifications based on text replays, it is possible that greater accuracy will be obtained by having two observers code each clip. To test this possibility, we will analyze both the set of 591 clips coded by observer A, and the set of 318 clips coded by both observers A and B. In the single-observer case, between 1 and 25 clips were classified for each student; in the two-observer case, between 1 and 15 clips were classified for each student. In both cases, we eliminated any student with 4 or fewer total text replay classifications (occurring due to sampling error, or using the tutor only briefly), since any estimated frequency based on such a small number of clips would be quite imprecise.

There was a correlation of 0.32 between the estimated gaming frequencies calculated using live observations, and the estimated gaming frequencies calculated using text replays by two observers, significantly better than chance, $F(1,48)= 5.76$, $p=0.02$. There was a very similar correlation of 0.31 between the live-observation gaming frequencies and the estimated gaming frequencies calculated using only observer A's text action description observations, again significantly better than chance, $F(1,47)=5.11$, $p=0.03$.

These results suggest that the text replay observations produce results reasonably similar to live observations, in terms of how often each student is assessed to be gaming. The correlation is not perfect – this may be due to differences between the behavior captured in each technique, or it may be due to the amount of variation which naturally exists due to time-sampling and the lower reliability of the text replays. Additionally, it appears that there is not a large difference between using data from a single observer and using data from two observers, for text replays.

Agreement With "Gold-Standard" Metric

We can also investigate how much each observational method varies from capturing "true gaming behavior" by comparing each observational method's results to an alternate gold standard. This gold standard is the machine-learned detector of gaming, which looks at every action the student makes across their whole history of using the tutor (Baker et al, to appear). The current version of the detector was originally trained using data from live co-located observations in multiple tutor lessons from 2003 to 2005, and has been verified to transfer effectively between tutor lessons without re-training (Baker et al, to appear). Since this detector was trained on live co-located observational data, including the 2005 live co-located scatterplot observations, it may bias in favor of the live co-located observations; nonetheless, it provides an additional test of each observational method's accuracy.

The live co-located observational data achieves an excellent correlation of 0.54 to the predictions made for each student by the gaming detector. The text replay data for both observers achieves a very similar correlation of 0.57 to the predictions made for each student by the gaming detector. The text replay data for only observer A also achieves a similar correlation of 0.59 to the gaming detector's predictions.

The correlation between the text observations (either single-observer or two-observer) and the gaming detector is significantly higher than the correlation between the text observations and the live observations, respectively $t(47) = 2.49$, $p=0.02$, $t(48) = 2.15$, $p=0.04$, for a two-tailed Hotelling's t-test (Walker and Lev, 1953). Similarly, the correlation between the live observations and the gaming detector is marginally significantly higher than the correlation between the live observations and the (two-observer) text observations, $t(48) =1.91$, $p=0.06$.

Thus, live and text observations correlate better to the gold standard than they correlate to one another. This suggests that the differences between the two methods have more to do with natural variation due to sampling than actual differences between the behaviors captured. And, as before, it does not appear that there is a substantial difference between using data from a single observer and using data from two observers, for text replay observations.

DISCUSSION AND CONCLUSIONS

Within this paper, we have presented a theoretical framework for differences in the fidelity of human classifications, which we define as how broad a band of data an observer can use when making inferences. We have also shown that, at least in the case of assessing whether students are gaming the system, a very low-fidelity observational technique (text descriptions of a sequence of student actions) has lower internal reliability than a higher-fidelity technique (live co-located observations), but nonetheless correlates as well to a gold-standard definition of gaming behavior as the higher-fidelity technique does. The low-fidelity and high-fidelity predictions are also correlated to each other, but not as well as either is correlated to the gold-standard definition of gaming, perhaps because of sampling variation. The low-fidelity observations were also more than 5 times faster to conduct than the high-fidelity observations.

Thus, low-fidelity observations using text replays appear to offer considerable advantages as a technique for studying student behavior: They are considerably faster to conduct than higher-fidelity observational techniques, achieve comparable accuracy, and can be conducted on the type of simple log-files which are collected by most tutor research groups. Because they can be conducted on standard log files, they can be used for retrospective analyses on existing corpuses of log tutor data, and do not require conducting a special observational experiment.

Another potential advantage, not explored in this paper, is that the set of low-fidelity observations an observer codes can be selected by processes other than randomly selected 20 or 30 second contiguous clips – for instance, it would be very feasible to analyze a student's behavior on a specific skill across every opportunity to practice that skill in a week of tutor usage. Such analyses may prove useful for tracking behaviors that occur sporadically over time; such analyses are considerably more difficult with higher-fidelity techniques such as live observations or video replays (where isolating the proper video segments becomes a significant task in itself).

It is important to note, however, that there are potential limitations to the applicability of low-fidelity observations. First, low-fidelity observations can only be used to detect behaviors which occur entirely within the

Table 3. Correlations between predicted frequency of gaming, from each method. All correlations are significantly higher than chance.

Study	Live Observations	Gaming detector
Live observations	.	0.54
Text replays (both observers)	0.32	0.57
Text replays (observer A)	0.31	0.59

tutoring system. For example, the original observations conducted by Baker et al. were also used to examine the frequency of off-task behavior such as talking off-task; text observations would almost certainly be unable to distinguish between talking off-task and talking on-task with the teacher.

Additionally, it is not clear to what degree low-fidelity observations will be successful for detecting affect, as opposed to specific behaviors such as gaming. Detecting affect may depend on subtleties in student behavior that are only capturable through higher-fidelity observation techniques. Determining what types of classification can successfully be conducted with low-fidelity observations will therefore be an important area of future work. Nonetheless, the higher speed and convenience associated with low-fidelity observations suggest that they may be a powerful tool for the analysis and development of tutoring systems that can respond in sophisticated ways to differences in student behavior and affect.

ACKNOWLEDGEMENTS

We would like to thank Ido Roll and James Fogarty for helpful suggestions and ideas. This work was funded by NSF grant REC-0437794 to "IERI: Learning-Oriented Dialogs in Cognitive Tutors: Toward a Scalable Solution to Performance Orientation".

REFERENCES

Adolphs, R. (2006) How do we know the minds of others? Domain-specificity, simulation, and enactive social cognition. *Brain Research,* 1079, 25-35.

Aleven, V., McLaren, B.M., Roll, I., Koedinger, K.R. (2004) 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.

Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., Roll, I. (to appear) Generalizing Detection of Gaming the System Across a Tutoring Curriculum. To appear in *Proceedings of the 8th International Conference on Intelligent Tutoring Systems.*

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

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

Conati, C., Maclaren, H. (2005) Data-driven refinement of a probabilistic model of user affect. *Proceedings of UM2005 User Modeling*, 40-49.

Craig, S.D., Graesser, A.C., Sullins, J., Gholson, B. (2004) Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media,* **29,** 3, 241-250.

de Vicente, A., Pain, H. (2002) Informing the detection of the students' motivational state: an empirical study. *Proceedings of the Sixth International Conference on Intelligent Tutoring Systems* (2002), 933-943.

Fogarty, J., Hudson, S.E., Atkeson, C.G., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J.C., Yang, J. (2005) Predicting human interruptibility with sensors. *ACM Transactions on Computer-Human Interaction (TOCHI),* **12,** 1, 119-146.

Jacob, R. (1995) Eye tracking in advanced interface design. In Baroeld, W., and Furness, T. (Eds.) *Advanced Interface Design and Virtual Environments.* Oxford, UK: Oxford University Press, 258-288.

Karweit, N., Slavin, R.E. (1982) Time-On-Task: Issues of Timing, Sampling, and Definition. *Journal of Experimental Psychology,* 74 (6), 844-851.

Lloyd, J.W., Loper, A.B. (1986) Measurement and Evaluation of Task-Related Learning Behavior: Attention to Task and Metacognition. *School Psychology Review,* 15 (3), 336-345.

Russom J.E., Johnson, E.J., Stephens, D.L. (1989) The validity of verbal protocols. *Memory and Cognition,* 17 (6), 759-769.

Walker, H.M, Lev, J. (1953) *Statistical Inference.* New York: Henry Holt & Co.