Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra

Ryan S.J.d. Baker
Worcester Polytechnic Institute
Worcester, MA USA
rsbaker@wpi.edu

Sujith M. Gowda
Worcester Polytechnic Institute
Worcester, MA USA
sujithmg@wpi.edu

Michael Wixon
Worcester Polytechnic Institute
Worcester, MA USA
mwixon@wpi.edu

Jessica Kalka
Carnegie Mellon University
Pittsburgh, PA USA
jessi.kalka@gmail.com

Angela Z. Wagner
Carnegie Mellon University
Pittsburgh, PA USA
awagner@cmu.edu

Aatish Salvi
Worcester Polytechnic Institute
Worcester, MA USA
aatishsalvi@gmail.com

Vincent Aleven
Carnegie Mellon University
Pittsburgh, PA USA
aleven@cs.cmu.edu

Gail W. Kusbit
Carnegie Mellon University
Pittsburgh, PA USA
gkusbit@cs.cmu.edu

Jaclyn Ocumpaugh
Worcester Polytechnic Institute
Worcester, MA USA
jocumpaugh@wpi.edu

Lisa Rossi
Worcester Polytechnic Institute
Worcester, MA USA
lrossi@wpi.edu

ABSTRACT

In recent years, the usefulness of affect detection for educational software has become clear. Accurate detection of student affect can support a wide range of interventions with the potential to improve student affect, increase engagement, and improve learning. In addition, accurate detection of student affect could play an essential role in research attempting to understand the root causes and impacts of different forms of affect. However, current approaches to affect detection have largely relied upon sensor systems, which are expensive and typically not physically robust to classroom conditions, reducing their potential real-world impact. Work towards sensor-free affect detection has produced detectors that are better than chance, but not substantially better—especially when subject to stringent cross-validation processes. In this paper we present models which can detect student engaged concentration, confusion, frustration, and boredom solely from students' interactions within a Cognitive Tutor for Algebra. These detectors are designed to operate solely on the information available through students' semantic actions within the interface, making these detectors applicable both for driving interventions and for labeling existing log files in the PSLC DataShop, facilitating future discovery with models analyses at scale.

Keywords

Educational data mining, affective computing, affect detection, boredom, engaged concentration, frustration, confusion, intelligent tutoring system

1. INTRODUCTION

In recent years, the log data collected through educational software such as intelligent tutoring systems has been a major resource for the educational data mining community [cf. 12; 33]. In specific, it has been possible to study changes in student learning and engagement over long periods of time by developing models using approaches such as classification or knowledge engineering and applying the models to larger data sets, a process termed “discovery with models.” Examples of this research include work to understand which models best predict student learning with an intelligent tutoring system [29; 38], work to find prerequisites within a curriculum [37], work to study the differences in engagement over the course of an entire year between urban, rural, and suburban schools [10], and work to study the differences in disengaged behavior between different tutor lessons [3].

Fewer research studies have focused on affect/academic emotions [30]. It is known that affect interacts with engagement and learning in complex fashions [cf. 7; 11; 21; 22; 26; 35]. However, research of this nature has largely been limited to relatively brief time-windows, on the order of a small number of lab sessions or field sessions. This limitation is due to the methods used in conducting these studies: self-report [cf. 1; 18; 35], retrospective emote-aloud protocols [cf. 19], field observations [cf. 7; 11; 22], and video observations [19, 21]. Each of these methods has been shown to produce replicable assessments of relevant academic affect, but each method also has limitations in terms of large-scale applicability. Specifically, self-report can disrupt naturally affective processes, and retrospective emote-aloud protocols, like observational methods, are expensive to conduct at large-scale.

A method with the potential to address this limitation is automated detection of affect. Researchers have been investigating affect detection from physiological sensors or vocal patterns for over a decade, and have produced successful detectors for a range of emotions. Such work is reviewed in detail by Calvo and D'Mello [14, 15]. In the domain of educational research, researchers have used sensors to develop detectors for several affective constructs. Litman and Forbes-Riley have found that features of students' voices while engaging in vocal dialogues with tutors can predict students' emotions [27]. D'Mello and Graesser have shown that a combination of body language and facial features, in combination with student interaction with the learning software, can be used to detect learner affect [20]. Muldner, Burleson, and VanLehn have shown that a combination of sensors can be used to detect student delight while learning [28]. Finally, Arroyo and colleagues have shown that sensor-based approaches to affect detection can work in urban schools and classrooms, enabling real-time adaptation to students' affect in an authentic learning setting [1].

However, approaches relying upon sensors are limited in application to data sets for which sensors were present. This limits applicability for schools, where sensor breakage can present a challenge for long-term use. In addition, sensor cost can be an economic challenge for schools, and internet connections may not have sufficient bandwidth to log full physiological sensor data for retrospective analysis.

Hence, to achieve maximum utility of an affect detector for retrospective discovery with models analysis, it is necessary to detect affect without reference to any sensor data. Ideally, such detection will be conducted solely with the type of log file data already being collected at large scale, such as the data being collected in the PSLC DataShop repository [cf. 24].

D'Mello and colleagues presented a first paper on an affect detector developed solely from log files [19]. Modeling student affect in the AutoTutor intelligent tutoring system in a laboratory study, they achieved decent agreement to ground-truth labels provided by human video coders. Their model successfully distinguished frustration from the neutral state approximately 40% better than the base rate (e.g. Kappa = approximately 0.4), and distinguished boredom, confusion, and flow from the neutral state approximately 20% better than chance. However, there were a few limitations in this pioneering study that need to be addressed to make sensor-free detectors of affect maximally useful. First, the detectors' best performance was achieved when distinguishing between specific affective states and the neutral state (e.g. all other affective states were discarded from the data set). The detectors achieved relatively poorer performance (Kappa = 0.163) when attempting to distinguish affective states from each other. Second, they re-sampled the data to eliminate imbalance between classes, and validated their models on the re-sampled data. Re-sampling is an appropriate method for generating unbiased classifiers, but the resultant models should ideally be tested on a non-resampled data set to verify detector effectiveness for future application of the models to data with natural class distribution. Third, their models were cross-validated at the observation level, rather than the student level, providing less information on detector generalizability to new students. Within this paper, we attempt to build on the methods in this pioneering research, while addressing these limitations.

A second paper developing non-sensor-based detectors was presented by Conati and Maclaren, who conducted a laboratory study of affect in the game Prime Climb [18]. In this paper, detectors using a combination of questionnaire and log data were used to predict self-reports of student affect, using a Bayesian framework. The cross-validation in this paper was conducted at the student level, giving information on model applicability to new students. Also, affect was compared using a median-split on binary distinctions (such as the distinction between joy versus distress), avoiding bias that may stem from discarding data that is neither the current affective state being detected nor the neutral state.

As in D'Mello et al. [19], Conati and Maclaren re-sampled the data to eliminate imbalance between classes during training. They validated their models using both the re-sampled distribution and the original distribution [18]. For the re-sampled data, their model was 32% better than the base rate at distinguishing between joy and distress, and 6% better than the base rate at distinguishing between admiration and reproach. However, their models achieved accuracy below the base rate when applied to the original distribution. This result indicates the challenge of achieving appropriate cross-validated performance for unbalanced constructs that are only indirectly reflected in student interaction within learning software.

A third paper developing sensor-free affect detectors was presented by Sabourin, Mott, and Lester, who studied the affect of students using the Crystal Island narrative-centered learning environment, [34]. In this classroom study, as in Conati and Maclaren's laboratory research [18], detectors based on a combination of questionnaire and log data were used to predict self-reports of student affect, using a Bayesian framework. As in [18], cross-validation was conducted at the student level. In addition, all relevant data was considered in model development and evaluation, and models were evaluated using the original data distribution rather than a re-sampled distribution. Their model was 38% better than the base rate at identifying focused students, and 24% better than the base rate at identifying curious students. It was less successful at identifying students who were confused (19% better than base rate), frustrated (14% better than base rate), bored (10% better than base rate), excited (8% better than base rate), although the detectors were better than the base rate for every construct except anxiety (3% worse than base rate). The only limitation for broad applicability of these models is the use of questionnaire measures, which require that a new student be given the same questionnaires for the model to be applied to that student.

A fourth paper, by Lee and colleagues [40], presented a sensor-free detector of confusion in a programming development environment. This detector achieved a very high student-level cross-validated Kappa of 0.86, but it is not clear if this detector was assessing the affective state of confusion or the more general experience of a student having difficulty with the material.

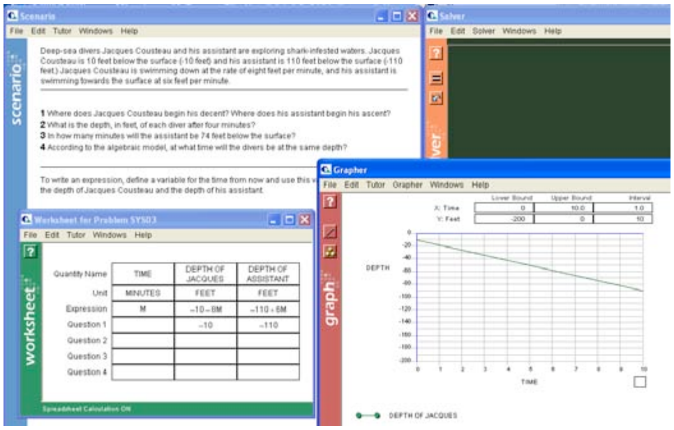
Within this paper, we build automated detectors of affect for Cognitive Tutor Algebra I, a widely used learning environment. In doing so, we restrict ourselves to the data generally available for this learning environment in the PSLC DataShop [cf. 24], making it feasible to apply the resultant detectors to hundreds of thousands of hours of student data. Ground-truth labels are obtained using field observations of affect [7] conducted using a handheld app for the Android platform, and then synchronized with log files. The detectors are constructed using only log data from student actions within the software occurring at the same time as or before the observations. Affect is known to have different prevalence following specific behaviors [cf. 7; 11; 35], suggesting that a detector that takes this information into account may be more effective than one that does not. By using only information from before and during the observation, our detectors can be used for fail-soft interventions, as well as discovery with models analyses.

2. METHODS

2.1 Data Collection

Data on student affect was collected from 89 students who were using Cognitive Tutor Algebra I as part of their regular mathematics curriculum. The students were using a lesson on systems of algebraic equations. Cognitive Tutors are a popular type of interactive learning environment now used by around half a million students a year in the USA. In Cognitive Tutors, students solve problems with exercises and feedback chosen based on a model of which skills the student possesses. Cognitive Tutor Algebra has been shown to significantly improve student performance on standardized exams and tests of problem-solving skill [25].

Each of the students studied in this paper were enrolled in one of four classes in a high school in rural Western Pennsylvania. In this school, 67% of students are rated as proficient or higher on the PSSA standardized exam, moderately higher than the state average. Students in this school are 96% Caucasian, typical in rural schools in this region, but higher than the state average. 18% of students are eligible for free or reduced-price lunch, approximately half of the state average. Students studied were approximately balanced in terms of gender.


Scenario
File Edit Tutor Windows Help
scenario
Solver
File Edit Solver Windows Help
ver
Worksheet for Problem SYS03
File Edit Tutor Windows Help
worksheet
Spreadsheet Calculation ON
Grapher
File Edit Tutor Grapher Windows Help
graph
DEPTH
0
-20
-40
-60
-80
-100
-120
-140
-160
-180
-200
0
1
2
3
4
5
6
7
8
9
10
TIME
DEPTH OF JACQUES

Deep-sea divers Jacques Cousteau and his assistant are exploring shark-infested waters. Jacques Cousteau is 10 feet below the surface (-10 feet) and his assistant is 110 feet below the surface (-110 feet.) Jacques Cousteau is swimming down at the rate of eight feet per minute, and his assistant is swimming towards the surface at six feet per minute.

1 Where does Jacques Cousteau begin his decent? Where does his assistant begin his ascent?
2 What is the depth, in feet, of each diver after four minutes?
3 In how many minutes will the assistant be 74 feet below the surface?
4 According to the algebraic model, at what time will the divers be at the same depth?

To write an expression, define a variable for the time from now and use this v
the depth of Jacques Cousteau and the depth of his assistant.

Quantity Name	TIME	DEPTH OF JACQUES	DEPTH OF ASSISTANT
Unit	MINUTES	FEET	FEET
Expression	M	-10 - 8M	-110 + 6M
Question 1		-10	-110
Question 2			
Question 3			
Question 4			

	Lower Bound	Upper Bound	Interval
X: Time	0	10.0	1.0
Y: Feet	-200	0	10

Figure 1: The Systems of Equations A lesson, from Cognitive Tutor Algebra I, used in this study.

Two expert field observers coded student affect and engaged/disengaged behaviors as students used the learning software. In this paper, we focus solely on the affect codes, as models of relevant engaged and disengaged behaviors were already available for this tutoring system (see discussion of features below). The coders used software on a Google Android handheld computer, which implemented an observation protocol developed specifically for the process of coding behavior and affect during use of educational software, replicating the protocol in [7]. All coding was conducted by the fourth and fifth authors. These two coders were previously trained in coding behavior and affect by the first author and have achieved inter-rater reliability with the first author of 0.72 (first and fourth authors, affect) and 0.83 (first and fifth authors, behavior [cf. 6]) in previous research conducted with students using other learning environments. This degree of reliability is on par with Kappas reported by past projects that have assessed the reliability of detecting naturally occurring emotional expressions [7; 13; 27; 32].

Observations were conducted in the school's computer laboratory, where students typically use the Cognitive Tutor software. Students were observed across 2 class days. Students were coded in a pre-chosen order, with each observation focusing on a specific student, in order to obtain the most representative indication of student affect possible. At the beginning of each class, an ordering of observation was chosen based on the computer laboratory's layout, and was enforced using the hand-held observation software. Setting up observations took a few minutes at the beginning of each class. A total of 408.51 minutes of observations were conducted across sessions, across the two coders. During this time, 763 observations were conducted across all students, not counting observations of students who were not logged into the software or not present in the classroom, for an average of 8.57 observations per student (SD = 2.84).

Each observation lasted up to twenty seconds, with observation time automatically coded by the handheld observation software. If affect and behavior were determined before twenty seconds elapsed, the coder moved to the next observation. Typically, each student observation involved 5 taps to the handheld screen, with the coder choosing affect and behavior codes from a pair of pop-up menus, and then clicking to confirm their selection. As such, data entry by an experienced coder took approximately 3 seconds per observation.

Each observation was conducted using peripheral vision or side-glances to reduce disruption. That is, the observers stood diagonally behind the student being observed and avoided looking at the student directly [cf. 5; 7; 32], in order to make it less clear when an observation was occurring. This method of observing using peripheral vision was previously found to be successful for assessing student behavior and affect, achieving good inter-rater reliability [cf. 6, 7; 32]. To increase tractability of both coding and eventual analysis, if two distinct affective states were seen during a single observation, only the first state observed was coded. Any affect of a student other than the student currently being observed was not coded.

The observers based their judgment of a student's state or behavior on the student's work context, actions, utterances, facial expressions, body language, and interactions with teachers or fellow students. These are, broadly, the same types of information used in previous methods for coding affect [e.g. 13], and in line with Planalp et al's [31] descriptive research on how humans generally identify affect using multiple cues in concert for maximum accuracy rather than attempting to select individual cues. The judgments of affect were based on a sub-set of the coding scheme used in [7; 21], selected based on importance for learning. Within an observation, each observer coded affect with reference to five categories:

- Boredom
- Confusion
- Engaged concentration (the affect associated with the flow state [cf. 7])
- Frustration
- "?" (which refers to any affect outside the coding scheme, including eureka, delight, and surprise. It also includes cases where it was impossible to code affect, such as when a student went to the bathroom or the software crashed.)

Delight and surprise were removed from the earlier coding scheme in [7; 21], due to the relative rarity of these affective states in prior research [e.g. 7; 21; 32].

Within the observations, the affective states had the following frequencies: boredom was observed 5.9% of the time, engaged concentration was observed 84.5% of the time, frustration was observed 0.9% of the time, and confusion was observed 1.8% of the time. The remaining observations were coded as "?". This distribution of affect is in line with prior studies – engaged concentration is typically the most common affect in classroom learning [cf. 7; 10; 34]. However, confusion, which tends to be relatively rare in most cases, was somewhat less frequent than has been typically seen in previous classroom studies [cf. 7; 10; 34].

2.2 Feature Distillation

In order to distill a feature set for detectors of affect, student actions within the software were synchronized to the field observations. Only the types of data available in standard PSLC DataShop log files [cf. 24] were used, towards producing detectors that could be applied retrospectively to existing data at scale.

During data collection, both the handhelds and the educational software server were synchronized to the same internet-time server. Actions during the twenty seconds prior to data entry by the observer were collected as a clip.

A total of 58 features were developed using the student's behavior both during and prior to the 20-second window. Some features were completely about the current action, such as whether it was correct or not. Other features, such as the number of previous actions on the current skill that involved help requests, involved data from the student's past performance. These 58 features were aggregated across the actions within the clip using mean, min, max and sum aggregators, hence a total of 232 features were used in the development of the detectors. Features involving past behavior (such as the number of previous actions on the current skill that involved help requests) are likely to have little change during the course of a clip, but were aggregated in the same fashion for simplicity of implementation.

Using both features on the current clip and features involving past data has the potential to help us detect affect more effectively, as there is evidence that the prevalence of specific affective states is different following specific behaviors [7; 10; 35] during real-world learning.

Features were drawn from two sources:

- Features developed during our group's past work to develop behavior detectors in Cognitive Tutors [cf. 2; 4; 8], averaged across actions in the clip (or min or max across actions), or across actions prior to the clip.
- Prior models of disengaged and engaged behaviors previously developed for this tutor or related tutors [cf. 2; 4; 8; 36]. Engaged and disengaged behaviors are known to precede and co-occur with affect, giving potential leverage for detecting affect.

Examples of features used can be seen in Table 2.

2.3 Machine Learning Algorithms

Each affective state was predicted separately – e.g. BORED was distinguished from NOT BORED (e.g. all other affective states), FRUSTRATED was distinguished from NOT FRUSTRATED (e.g. all other affective states), and so on. This resulted in four detectors, one for boredom, confusion, engaged concentration, and frustration respectively.

Each detector was evaluated using six-fold student-level cross-validation [cf. 17; 34]. In this process, students are split randomly into six groups. Then, for each possible combination, a detector is developed using data from five groups of students before being tested on the sixth "held out" group of students. By cross-validating at this level, we increase confidence that detectors will be accurate for new students.

For each construct being detected, a separate student-level cross-validation was conducted, which stratified students based on the dependent variable. This procedure was used in order to guarantee that each fold had a representative number of observations of the majority and minority class. In addition, for unbalanced classes, re-sampling was used to make the class frequency more equal for detector development. However, all goodness calculations were made with reference to the original data set, as in Sabourin et al. [34].

We attempted to fit sensor-free affect detectors using eight common classification algorithms that have been successful for past educational data mining problems, including J48 decision trees, step regression, JRip, Naïve Bayes, and REP-Trees.

Feature selection for machine learning algorithms was conducted using forward selection, where the feature that most improves model goodness is added repeatedly until adding additional features no longer improves model goodness. During feature selection, cross-validated kappa on the original (e.g. non-re-sampled) data set was used as the goodness metric. Prior to feature selection, all features with cross-validated kappa equal to or below zero in a single-feature model were omitted from consideration, as a check on over-fitting.

Detector goodness was assessed using two metrics: Cohen's Kappa [17] and A' [23]. Cohen's Kappa assesses the degree to which the detector is better than chance at identifying which clips involve a specific affective state. A Kappa of 0 indicates that the detector performs at chance, and a Kappa of 1 indicates that the detector performs perfectly. For example, a Kappa of 0.31 would indicate that the detector is 31% better than chance. A' is the probability that the algorithm will correctly identify whether a specific affective state is present or absent in a specific clip. A' is equivalent to both the area under the ROC curve in signal detection theory, and to W, the Wilcoxon statistic [23]. A model with an A' of 0.5 performs at chance, and a model with an A' of 1.0 performs perfectly. In these analyses, A' was computed at the level of clips, rather than students, using the AUC (area under the curve) approximation.

3. RESULTS

Detector performance for all four constructs studied was better than chance (see discussion of cross-validation methodology in the previous section), but left room for improvement. Full results are shown in Table 1. For engaged concentration, the best algorithm was K*. The engaged concentration detector achieved an A' of 0.71 and a Kappa of 0.31. For confusion, the best algorithm was JRip. The confusion detector achieved an A' of 0.99 and a Kappa of 0.40. For frustration, the best algorithm was REPTree. The frustration detector achieved an A' of 0.99 and a Kappa of 0.23. For boredom, the best algorithm was Naïve Bayes. The boredom detector achieved an A' of 0.69 and a Kappa of 0.28.

Several of these detectors showed an imbalance between A' and Kappa. Imbalance of this nature typically indicates a detector which is better at getting the relative order between classes correct (in its confidence estimates) than at drawing an optimal line between classes. Using detectors of this nature, whether for intervention or discovery with models analyses, will be more effective if confidence is taken into account.

Features automatically selected for each of the detectors during machine learning are listed in Table 2. Full detail on models, including runnable versions of the models (for RapidMiner 4.6) can be found in the PSLC DataShop [24], in data set "Baker – Closing the Loop on Gaming – Hopewell Spring 2011", at (https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=479). This data set also includes all data used in the analysis, distilled features used to develop the detectors, data from the field observations synchronized with the student interaction data, and the detector's predictions for actions not initially labeled through the field observations.

The features for engaged concentration involve actions where the student was more likely to have a history of few errors and help requests on the skills in the clip. When a student is in engaged concentration and requests help, they typically do not follow it up with an error (e.g. they read the hints carefully; while they might also have been gaming the tutor's hints, gaming typically results in some proportion of errors, as students do not read hints carefully). These features suggest a student who is closely engaged and working effectively.

	A'	Kappa
Engaged Concentration	0.71	0.31
Confusion	0.99	0.40
Frustration	0.99	0.23
Boredom	0.69	0.28
Average Across Constructs	0.85	0.30

Table 1: The goodness of each final model, under student-level cross-validation, for the original data set.

The features for confusion suggest a struggling student [cf. 26]. A confused student is more likely to have a pattern of slower actions after making two errors and tend to have a history of more incorrect actions and help requests. Furthermore, his or her correct actions are relatively more likely to represent guesses, using the contextual guess model from [4]. On the other hand a student who is not confused tends to be able to successfully answer 5 items in rows, working slowly.

Engaged Concentration
The minimum number of previous incorrect actions and help requests for any skill in the clip.
Among the skills involved in the clip, the minimum value for previous incorrect actions and help requests for that skill.
The duration (in seconds) of the fastest action in the clip.
The percentage of clip actions involving a hint followed by an error.

Boredom
The average time the student took to respond on the current step prior to the clip, averaged across all the actions with a clip.
The average time the student took to respond, unitized across time taken by all students on the same problem steps, within sequences of three actions in a row.
The maximum product of the probability of moment-by-moment learning P(J) [9], and the probability of guess P(G) calculated using the contextual guess model [4] for any action in the clip. This can be interpreted as actions where the student learned after guessing.
The maximum number of previous incorrect actions and help. requests for any skill in the clip.

Confusion
The percentage of clip actions involving actions taking longer than 5 seconds after two incorrect answers.
The percentage of actions in the clip that were hint requests.
The minimum number of previous incorrect actions for any skill in the clip.
The maximum product of the probability of guess P(G) as computed using contextual guess model [4], across sequences of three actions in a row.
The average time the student took to respond, unitized across time taken by all students on the same problem steps, within sequences of five actions in a row that were correct.

Frustration
The percent of past actions on the skills involved in the clip that were incorrect.
Were there any actions in the clip where the student made a wrong answer rather than requesting help when their probability of knowing the skill was under 0.7?

Table 2. The features in the final detectors of each construct.

The features for frustration involve incorrect actions and help avoidance. In particular, frustrated students tend to have a history of past incorrect actions and help requests. Curiously, frustrated students are more likely to avoid help and make errors when they do not know the skill. It is unclear whether this behavior is a result of frustration, or whether it is perhaps a cause of frustration.

The features for boredom are interestingly different than the features for other constructs. Bored students were more likely to guess than other students. Interestingly, though, they were also relatively likely to learn from their guesses. Compared to other students, bored students were relatively less likely to have a history of many errors and help requests. In addition, students who were bored had a past history of working slowly, and worked slowly while they were bored, across multiple actions within the tutor software.

4. DISCUSSION AND CONCLUSIONS

In this paper, we have presented automated detectors that are a step towards identifying student affect solely from log files, in a Cognitive Tutor for Algebra. These detectors are better than chance at identifying engaged concentration, confusion, frustration, and boredom, among a population of students using the Cognitive Tutor as part of their regular mathematics classes.

These detectors achieve goodness values that are moderately better than past values obtained through sensor-free detectors, when averaged across constructs. In this study, the average detector Kappa was 0.30. The detectors closest in validation to this study within D'Mello et al. [19], albeit in a different domain, achieved an average Kappa of 0.16. The detectors closest in validation to this study within Conati & Maclaren [18] achieved an average accuracy below the base rate, and detectors validated on re-sampled data achieved an average accuracy that was 19% better than the base rate (approximately comparable to Kappa of 0.19). The detectors in Sabourin et al. [34], validated in the same fashion as these detectors, achieved an average accuracy that was 16% better than the base rate (approximately comparable to Kappa of 0.16). Individual detectors from previous studies performed better than the detectors presented here (e.g. frustration in [19], focused/engaged concentration in [34]), but on the average the detectors presented here performed better than detectors presented in previous papers. While comparison of model goodness obtained in different software platforms, age groups, and populations should be done with caution, the detectors presented here appear to represent further progress towards effective, sensor-free detectors of affect.

It is possible that at least part of this progress is the result of a greater degree of feature engineering in this detector's development, including the use of features previously used to detect disengaged behaviors, and existing models of several potentially relevant constructs such as guessing [4]. These results suggest that by using both the detectors of disengaged behaviors known to be associated with affect as features and the features used to produce those detectors, increased detector goodness can be obtained with acceptable construct validity.

At the same time, our affect detectors are clearly still imperfect. These new features have only achieved 30% of potential progress towards perfect detection, and, while perfect detection is probably infeasible (after all, even expert coders only achieve Kappa values around 0.6 or 0.7), there is clearly substantial room for improvement. Further work should consider further feature engineering, and potentially alternate methods for aggregating data. Continued improvement in terms of feature engineering may be supported by further research on the behaviors that correspond to specific affective states [cf. 7; 11; 35].

In addition, there is considerable work needed in the area of cross-validation. The detectors presented here are developed and validated for a single, fairly homogenous population. As such, their validity for the broad and diverse population of learners using Cognitive Tutor Algebra in the USA has not yet been fully established. Likewise, the detectors are developed within the context of a single Cognitive Tutor lesson. As such, the detector's validity for new curricular materials has not been established. The detectors may indeed be generalizable and usable in new contexts, as past detectors of disengaged behaviors have often been found to be (for instance, in their use within the detectors presented within this paper), but establishing generalizability will be an important area of future work.

One positive note for the applicability of these detectors in other populations and domains is that the behaviors identified by each detector have reasonable construct validity, suggesting that the detectors may be less accurate in these contexts, but are unlikely to provide meaningless predictions. For this reason, it may still be appropriate to use these detectors in discovery with models analyses, with the expectation that the strength of correlations may be reduced, but that findings with high strength are unlikely to be wholly spurious. The detectors can also be used immediately in the development of detectors of other constructs, as behavior detectors were used here. In these cases, the validity of the detectors is shown by their relevance to detecting other constructs. Thus, though the detectors are imperfect, they still may prove a useful component for EDM research. As these detectors predict affect solely using log file data, they can be applied to existing data from Cognitive Tutor Algebra in the PSLC DataShop and elsewhere. As hundreds of thousands of students use this software each year, we believe that many analyses can be accomplished with these detectors and look forward to working with colleagues to accomplish this goal.

Similarly, it may be possible to incorporate these detectors into the Cognitive Tutor software for fail-soft interventions, which could be used to advance learning outcomes.

In the long-term, detectors of this nature are likely to provide a useful tool for understanding and automatically adapting to differences in learner affect. We see the work here as an incremental step, following on the pioneering work of D'Mello and colleagues [19], Conati and Maclaren [18], and Sabourin, Mott, and Lester [34] towards this goal.

5. ACKNOWLEDGMENTS

This research was supported by grant "Toward a Decade of PSLC Research: Investigating Instructional, Social, and Learner Factors in Robust Learning through Data-Driven Analysis and Modeling," National Science Foundation award #SBE-0836012, and by grant #OPP1048577 from the Bill & Melinda Gates Foundation. We also thank Leigh Ann Sudol, Ma. Mercedes T. Rodrigo, and Sidney D'Mello for helpful comments and suggestions.

6. REFERENCES

[1] Arroyo, I., Woolf, B.P., Cooper, D., Burleson, W., Muldner, K., and Christopherson, R. Emotion Sensors Go To School. In V. Dimitrova, R. Mizoguchi, B. Du Boulay, and A. Graesser (eds.) *14th International Conference on Artificial Intelligence In Education*, 2009.

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

[3] Baker, R.S.J.d. Differences Between Intelligent Tutor Lessons, and the Choice to Go Off-Task. *Proceedings of the 2nd International Conference on Educational Data Mining,* pp. 11-20, 2009.

[4] Baker, R.S.J.d., Corbett, A.T., and Aleven, V. More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, pp. 406-415, 2008.

[5] Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., and Wagner, A.Z. Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game The System". *Proceedings of ACM CHI 2004: Computer-Human Interaction*, pp. 383-390, 2004.

[6] Baker, R.S.J.d., Corbett, A.T., and Wagner, A.Z. Human Classification of Low-Fidelity Replays of Student Actions. *Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems,* pp. 29-36, 2006.

[7] Baker, R.S.J.d., D'Mello, S.K., Rodrigo, M.M.T., and Graesser, A.C. Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive-Affective States during Interactions with Three Different Computer-Based Learning Environments. *International Journal of Human-Computer Studies,* 68 (4), pp. 223-241, 2010.

[8] Baker, R.S.J.d., and de Carvalho, A.M.J.A. Labeling Student Behavior Faster and More Precisely with Text Replays. *Proceedings of the 1st International Conference on Educational Data Mining,* pp. 38-47, 2008.

[9] Baker, R.S.J.d., Goldstein, A.B., Heffernan, N.T. Detecting Learning Moment-by-Moment. To appear *in International Journal of Artificial Intelligence in Education,* in press.

[10] Baker, R.S.J.d., and Gowda, S.M. An Analysis of the Differences in the Frequency of Students' Disengagement in Urban, Rural, and Suburban High Schools. *Proceedings of the 3rd International Conference on Educational Data Mining,* pp. 11-20, 2010.

[11] Baker, R.S.J.d., Moore, G., Wagner, A., Kalka, J., Karabinos, M., Ashe, C., and Yaron, D. The Dynamics Between Student Affect and Behavior Occurring Outside of Educational Software. *Proceedings of the 4th bi-annual International Conference on Affective Computing and Intelligent Interaction*, pp. 14-24, 2011.

[12] Baker, R.S.J.d., and Yacef, K. The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining,* 1 (1), pp. 3-17, 2009.

[13] Bartel, C.A., and Saavedra, R. The collective construction of work group moods. *Administrative Science Quarterly,* 45, pp. 197-231, 2001.

[14] Calvo, R.A., and D'Mello, S.K. Affect Detection: An Interdisciplinary Review of Models, Methods, and their Applications. *IEEE Transactions on Affective Computing,* 1 (1), pp. 18-37, 2010.

[15] Calvo, R.A., and D'Mello, S.K. eds. New Perspectives on Affect and learning technologies. New York: Springer. 2011.

[16] Cleary, J.G., and Trigg, L.E. K*: An Instance- based Learner Using an Entropic Distance Measure. *Proceedings of the 12th International Conference on Machine learning,* pp. 108-114, 1995.

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

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

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

[20] D'Mello, S.K., and Graesser, A.C. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-adapted Interaction*. 20 (2), pp. 147-187, 2010.

[21] D'Mello, S.K., Taylor, R., and Graesser, A.C. Monitoring Affective Trajectories during Complex Learning. In D.S. McNamara & J.G. Trafton (eds.) *Proceedings of the 29th Annual Cognitive Science Society*, pp. 203-208, 2007.

[22] Dragon, T., Arroyo, I., Woolf, B.P., Burleson, W., El Kaliouby, R., and Eydgahi, H. Viewing Student Affect and Learning through Classroom Observation and Physical Sensors. *International Conference on Intelligent Tutoring Systems 2008,* pp. 29-39, 2008.

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

[24] Koedinger, K. R., Baker, R., Cunningham, K., Skogsholm, A., Leber, B., and Stamper, J. A data repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy, and R.S.J.d. Baker (eds.) *Handbook of Educational Data Mining*. pp. 43-55, Boca Raton, FL: CRC Press, 2010.

[25] Koedinger, K.R., and Corbett, A.T. Cognitive Tutors: Technology bringing learning science to the classroom. In K. Sawyer (ed.) *The Cambridge Handbook of the Learning Sciences*, pp. 61-78, Cambridge, UK: Cambridge University Press, 2006.

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

[27] Litman, D.J., and Forbes-Riley, K. Recognizing Student Emotions and Attitudes on the Basis of Utterances in Spoken Tutoring Dialogues with both Human and Computer Tutors. *Speech Communication*, 48 (5), pp. 559-590, 2006.

[28] Muldner, K., Burleson, B., and VanLehn, K. "Yes!": Using tutor and sensor data to predict moments of delight during instructional activities. *Proceedings of the International Conference on User Modeling and Adaptive Presentation (UMAP'10),* pp. 159-170, 2010.

[29] Pardos, Z.A., and Heffernan, N. T. Using HMMs and bagged decision trees to leverage rich features of user and skill from

an intelligent tutoring system dataset. To appear in the *Journal of Machine Learning Research W & CP*, in press.

[30] Pekrun, R., Goetz, T., Titz, W., and Perry, R. P. Academic emotions in students' self-regulated learning and achievement: A program of quantitative and qualitative research. *Educational Psychologist*, 37, pp. 91-106, 2002.

[31] Planalp, S., DeFrancisco, V. L., and Rutherford, D. Varieties of cues to emotion in naturally occurring situations. *Cognition and Emotion*, 10 (2), pp. 137-153, 1996.

[32] Rodrigo, M.M.T., Baker, R.S.J.d., D'Mello, S., Gonzalez, M.C.T., Lagud, M.C.V., Lim, S.A.L., Macapanpan, A.F., Pascua, S.A.M.S., Santillano, J.Q., Sugay, J.O., Tep, S., and Viehland, N.J.B. Comparing Learners' Affect While Using an Intelligent Tutoring Systems and a Simulation Problem Solving Game. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, pp. 40-49, 2008.

[33] Romero, C., and Ventura, S. Educational Data Mining: A Review of the State-of-the-Art. *IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 6, pp. 601-618, 2010.

[34] Sabourin, J., Mott, B., and Lester, J. Modeling Learner Affect with Theoretically Grounded Dynamic Bayesian Networks. *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction*, pp. 286-295, 2011.

[35] Sabourin, J., Rowe, J., Mott, B., and Lester, J. When Off-Task in On-Task: The Affective Role of Off-Task Behavior in Narrative-Centered Learning Environments. *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, pp. 534-536, 2011.

[36] Shih, B., Koedinger, K.R., and Scheines, R. A response time model for bottom-out hints as worked examples. *Proceedings of the 1st International Conference on Educational Data Mining*, pp. 117-126, 2008.

[37] Vuong, A., Nixon, T., and Towle, B. A Method for Finding Prerequisites Within a Curriculum. *Journal of Educational Data Mining*, pp. 211-216, 2011.

[38] Yu, H.-F., Lo, H.-Y., Hsieh, H.-P., Lou, et al. Feature Engineering and Classifier Ensemble for KDD Cup 2010. *Proceedings of the KDD Cup 2010 Workshop*, pp. 1–16, 2010.