Sensor-Free Affect Detection for a Simulation-Based Science Inquiry Learning Environment

Luc Paquette[1], Ryan S. Baker 2[1], Michael A. Sao Pedro[2], Janice D. Gobert[2], Lisa Rossi[3], Adam Nakama[2], Zakkai Kauffman-Rogoff[2]

[1]Teachers College, Columbia University, New York, NY
[2]Worcester Polytechnic Institute, Worcester, MA
[3]Georgia Institute of Technology, Atlanta, GA
paquette@tc.columbia.edu, baker2@exchange.tc.columbia.edu, mikesp@wpi.edu, jgobert@wpi.edu, lrossi@gatech.edu, nakama@wpi.edu, zakkai@gmail.com

Abstract. Recently, there has been considerable interest in understanding the relationship between student affect and cognition. This research is facilitated by the advent of automated sensor-free detectors that have been designed to "infer" affect from the logs of student interactions within a learning environment. Such detectors allow for fine-grained analysis of the impact of different affective states on a range of learning outcome measures. However, these detectors have to date only been developed for a subset of online learning environments, including problem-solving tutors, dialogue tutors, and narrative-based virtual environments. In this paper, we extend sensor-free affect detection to a science microworld environment, affording the possibility of more deeply studying and responding to student affect in this type of learning environment.

Keywords: Educational data mining, affect detection, affective computing

1 Introduction

It is well recognized that affect interacts with engagement and learning in complex ways [1, 2, 3, 4, 5, 6]. Learning software such as ITSs offer great opportunities to study those interactions due to their fine-grained interaction logs and their capacity to track students' actions at multiple levels. In recent years, this research has been facilitated by the use of sensor-free affect detectors that can automatically infer a range of student affective states from student interactions. Sensor-free detectors have been developed for three kinds of ITSs to date: problem-solving ITSs where answers are straightforward (e.g. [7, 8, 9]), dialogue tutors where the student iterates towards an answer (e.g. [10, 11]), and narrative-based virtual environments where the student explores a complex environment (e.g. [12]). One key finding is that, though the principles of affect detection are largely the same, the student behaviors associated with each affect often differ considerably based on the design of the learning environment being used. For instance, affect detection in problem-solving tutors tends to focus on timing, pauses, and patterns of errors, and the contexts in which they occur. In game-

like virtual environments such as Crystal Island, affect detectors have been built using counts of how many times the player engaged in meaningful actions such as viewing books, and whether the student has completed important milestones [12]. In dialogue tutors, affect detection tends to focus on the actual content of student dialogue acts and how the content changes over time. Given this coupling between student behaviors indicative of affective states and the learning environment in which they are demonstrated, it is important to study those behaviors in a broader range of learning environments to make sensor-free affect detection more feasible.

In this paper, we study how to automatically detect student affect in the Inq-ITS inquiry learning environment [13] in which students use simulation and support tools to engage in inquiry. We do this by using a combination of data mining and ground-truth labels that were obtained from field observations of affect. When compared with other systems, Inq-ITS's simulation microworlds offer a less constrained learning environment than problem-solving [7, 8, 9] or dialogue tutors [10, 11], allowing more exploratory behaviors. At the same time, simulation microworlds are more constrained than virtual environments, such as Crystal Island [12] and EcoMUVE [14], where students have a lot of freedom to explore the virtual world which can lead to a wider range of ways that affect can manifest in behaviors.

Prior research on affect in simulation microworlds has provided evidence of a range of different affective states associated with learning. For example, relatively high amounts of boredom, an undesirable affect associated with both gaming the system [1] and off-task behavior [15], has been observed in some simulation microworlds [15]. The availability of sensor-free affect detectors for this type of environment would enable more in-depth studies of similar relationships, providing a better understanding of how affect impacts learning in these rich learning contexts.

2 Inq-ITS Learning Environment

The Inq-ITS learning environment (formerly known as Science Assistments [13]) is a web-based environment in which students conduct inquiry with interactive simulations aligned to middle school Physical, Life, and Earth Science content described in the NGSS standards [16]. Activities have a driving question pertinent to a science topic, and require students to address the question by conducting an investigation using a simulation and other inquiry support tools.

For example, a driving question in a Phase Change activity asks students to determine if one of three factors (size of a container, amount of ice to melt, and amount of heat applied to the ice) affects various measurable outcomes (e.g., melting or boiling point). Students address this by conducting inquiry, i.e., formulating a hypothesis, collecting data to test it with the simulation, analyzing the data, warranting their claims, and communicating their findings. Before making a hypothesis, students can first explore the simulation. More information about Inq-ITS can be found in [13, 17].

3 Method

3.1 Data Collection

Data on student affect was collected from 326 students who conducted inquiry within the Inq-ITS system in 2011 in 11 different 8th grade classes from 3 schools in Massachusetts. Students came from a diverse population (Table 1).

Table 1. Demographic information for the three schools in our data set.

	First school	Second school	Third school	State average
Hispanic students	3%	6%	40%	10%
African-American students	0%	2%	17%	8%
Asian-American students	3%	12%	12%	6%
Caucasian Students	89%	79%	28%	76%
Students at or above proficient level on the MCAS science test	53%	63%	10%	39%
Students receiving reduced or free lunch	5%	16%	83%	34%

Four expert field observers coded student affect and engaged/disengaged behaviors while students used the software. Here, we focus on the affect codes. The observers based their judgment of a student's affect on the student's work context, actions, utterances, facial expressions, body language, and interactions with teachers or fellow students [cf. 18, 19]. Within an observation, each observer coded affect on five categories [1]: boredom, confusion, frustration, engaged concentration (the affect associated with the flow state [cf. 1]) and "?" (an affect different from the coding scheme and situations when coding was impossible/irrelevant such as when a student went to the bathroom).

The coders used the HART app for Google Android handheld computers [8], which implements the Baker-Rodrigo Observation Method Protocol (BROMP) [1, 20], a protocol for coding affect and behavior during use of educational software. All coding was conducted by the second, fifth, sixth, and seventh authors. These coders were previously trained by two expert coders. Pairs of coders achieved inter-rater reliability (Kappa) of 0.72 (second and sixth authors, affect), 0.60 (second and seventh, affect) and 0.60 (fifth author and additional expert coder, affect). This degree of reliability is on par with Kappas reported by past projects that have assessed the reliability of detecting naturally occurring emotional expressions [1, 18, 21, 22].

As mandated in BROMP [20], students were coded in a pre-chosen order, with each observation focusing on a specific student. To obtain the most representative indication possible of student affect, only the current student's affect was coded. At the beginning of each class, an ordering of observation was chosen based on the class layout and was enforced using the hand-held observation software. A total of 4155 observations were made across the 326 students. Each observation lasted up to twenty seconds, with observation time automatically coded by the handheld software. If affect and behavior were determined before twenty seconds elapsed, the coder moved to the next observation. If two distinct affective states occurred during a single observa-

tion, only the first state observed was coded. Each observation was conducted using peripheral vision or side-glances to reduce disruption [cf. 1, 20, 22, 23].

From the initial 4155 observations, 1214 (from 205 students, with an average of 5.92 observations per students and a standard deviation of 5.94) were used in the final analyses. Of the 2941 discarded observations: 1146 were coded as "?"; 331 were made while the student had been inactive for more than 5 minutes; and 1464 were made when the student was not currently involved in a science inquiry task (for example, when the student was answering other multiple-choice test questions [e.g. 24]). Within the 1214 remaining observations, the affective states had the following frequencies: engaged concentration was observed 896 times (82.50%), boredom 109 times (10.03%), confusion 44 times (4.05%), and frustration 38 times (3.50%).

3.2 Feature Distillation

In order to distill a feature set for our affect detectors, student actions within the software were synchronized to the field observations. During data collections, both the handheld computers and the Inq-ITS server were synchronized to the same internet NTP time server. Actions during the 20 seconds prior to data entry by the observer were considered as co-occurring with the observation. A total of 127 features were developed using the actions that co-occurred with or preceded the observation.

Our main feature set was based on the 73 features distilled by Sao Pedro et al. in [24], which looked at the different types of actions the students can make while they use Inq-ITS. Of the action types distilled in [24], we kept those that occurred in our data set: hypothesis variable changes, simulation variable changes, simulation pauses, incomplete trials run, complete trials run, all trials run and all relevant actions. We note that Sao Pedro et al. [24] did not include student interactions during the analysis stage of the inquiry process. We included analysis stage interactions to enable affect detection in that stage and created 7 new features to summarize those interactions.

To compute values for the previously described features, we accumulated lists of each type of relevant action during the 60 seconds prior to an observation to capture the student's behavior immediately before it. For each of those lists, like [24], we calculated the minimum, maximum, average, median, standard deviations, and sum of the time spent on each action, as well as a count of the number of actions in the list. Since some observations were made when the student had been inactive for more than 60 seconds, we repeated the same process to create a second set of features using lists from the 5 actions prior to the observation. This combination accounted for 112 of the features distilled from our dataset.

We created two features related to the time elapsed since the last student action: a binary feature indicating whether the student has been inactive for the last 60 seconds, a potential indicator of off-task behavior [cf. 8], and the time elapsed between the last action of the student and the moment of the observation.

Bayesian knowledge tracing (BKT) was used to distill features indicating whether students knew how to apply two inquiry skills, designing controlled experiments and testing stated hypothesis [24]. Three features were computed for each skill: the probability that the skill was known before the most recent practice opportunity, the proba-

bility the skill was known afterwards, and the probability that the student would correctly apply the skill on the most recent practice opportunity. In addition, we computed the ratio of positive and negative assessments during the last 5 student actions.

An additional 3 features were distilled in relation to the different stages of the inquiry process: whether the student had explored the microworld before making a hypothesis, whether the student had completed the current stage at least once in a past activity and the time elapsed so far during the current inquiry stage.

Finally, in the version of Inq-ITS (Science Assistments) for which the interaction data were collected, each time a student enters a stage for the first time for the current activity, the system shows a text box containing orienting instructions for each stage of inquiry. We created three features related to this text box: whether it is currently open, the time elapsed since it was opened (if it is still opened), and whether the student closed it during the 20-seconds of actions co-occurring with the observation.

3.3 Machine Learning Algorithms

We built separate detectors for four affective states: boredom, confusion, frustration, and engaged concentration for three stages of inquiry: hypothesizing, collecting data, and analyzing data. Thus, each affective state was predicted separately – e.g. BORED was distinguished from NOT BORED (i.e., all other affective states) within each inquiry stage (i.e., BORED/NOT BORED in hypothesizing, BORED/NOT BORED while collecting data, etc.). Separate detectors were created for each stage because they each have specific actions associated with its user interface. As such, the patterns of actions related to each affective state may differ between stages. For the specific case of engaged concentration, cases where students were off-task were considered NOT ENG. CONC., since this reflects engaged concentration with something other than learning or Inq-ITS (e.g. the day's classroom gossip). Also, no detectors were built for the "exploring" stage due to the low number of observations (only 23). Table 2 shows the frequency of each affective state.

Each detector was evaluated using leave-one-out student-level cross-validation. In this process, for each student, a detector is built using data from every other student before being tested on that student. By cross-validating at this level, we increase confidence that detectors we build with a specific feature set will be accurate for new students. In addition, re-sampling was used to make the class frequency more equal for detector development (e.g. 96.15% of the observations were labeled as "not frustrated" during hypothesizing). However, all performance calculations were made with reference to the original dataset, as in [12].

Table 2. Frequency of the affect observation across the four stages of inquiry.

	Hypothesizing	Experimenting	Analyzing
BORED	35 (11.22%)	43 (8.14%)	28 (7.98%)
CONFUSED	13 (4.17%)	19 (3.60%)	10 (2.85%)
FRUSTRATED	12 (3.85%)	13 (2.46%)	10 (2.85%)
ENG. CONC.	220 (70.51%)	390 (73.86%)	271 (77.21%)

We fit sensor-free affect detectors using three common classification algorithms that have been successful for building affect detectors in the past [8, 9]: J48 decision trees, JRip, and step regression (linear regression with a step function). By fitting the detectors using multiple algorithms, we can select the best algorithm for each affective state, as manifested in the relationship between the distilled features and the affect labels (linear, small clusters, etc.). Detector performance was assessed using two metrics: Cohen's Kappa [25] and A' computed as the Wilcoxon statistic [26]. Cohen's Kappa assesses the degree to which the detector is better than chance at identifying the student's affective state for a specific observation. A Kappa of 0 indicates that the detector performs at chance, and a Kappa of 1 indicates that the detector performs perfectly. A' is the probability that the algorithm will correctly identify whether an observation is an example of a specific affective state. A' is equivalent to the area under the ROC curve in signal detection theory, and is approximated by W [26]. A model with an A' of 0.5 performs at chance, and a model with an A' of 1.0 performs perfectly. A' was computed at the observation-level.

Feature selection for machine learning was conducted using two semi-automated procedures. First, we applied forward selection, a process in which the feature that most improves model performance is added repeatedly until adding additional features no longer improves performance. During forward selection, cross-validated Kappa and A' on the original non-resampled dataset were used. Kappa was used as the main performance metric for selecting a feature, but an alternate feature was selected when the model's A' was judged to be unusually low when compared to the value of Kappa. Then, backward elimination was applied on the sets of features generated by the forward selection algorithm to determine whether a simpler model could achieve better or equivalent performance, thereby reducing model over-fitting.

4 Results

We evaluate the degree to which the detectors for each construct within each inquiry stage can identify their respective affect. Detectors' performance over all four constructs and across all inquiry tasks was better than chance (A' = .50, Kappa = 0.0) and comparably well to past sensor-free detectors of affect. Table 3 shows the performance of the 12 detectors we built and provides a list of the features used in each detector. Descriptions of each feature (from F1 to F47) are provided in Table 4. The average student cross-validated Kappa was 0.354 and the average A' was 0.720. This is above the average Kappa of 0.296 and A' of 0.682 obtained in a study with similar validation [9] within the ASSISTments problem-solving ITS for math. The detectors described in [12] for a virtual environment achieved an average accuracy that was 16% better than the base rate (approximately comparable to a Kappa of 0.16). The detectors for Cognitive Tutor Algebra from [8] achieved an average Kappa of 0.30.

Another positive aspect of our detectors is that they were cross-validated at the student-level, and developed using a diverse population (Table 1). As such, it is likely that they will generalize to new students across the entire population of Inq-ITS users.

Table 3. Each of the models and their student-level cross-validated performances.

	Hypothesizing	Experimenting	Analyzing
BORED	J48 F2, F7, F31, F38 Kappa = 0.305 A' = 0.699	JRip F3, F15, F19, F35, F36 Kappa = 0.252 A' = 0.704	J48 F1, F10, F22, F37, F45, F47 Kappa = 0.438 A' = 0.767
CONFUSED	J48 F2, F14, F31, F45 Kappa = 0.327 A' = 0.704	JRip F2, F3, F9, F16, F20 Kappa = 0.355 A' = 0.777	J48 F20, F28, F33, F38 Kappa = 0.319 A' = 0.724
FRUSTRATED	JRip F2, F5, F31, F42, F44, F46 Kappa = 0.301 A' = 0.688	J48 F8, F11, F18, F30, F34, F39, F46 Kappa = 0.486 A' = 0.762	J48 F13, F23, F24, F26, F30, F32 Kappa = 0.379 A' = 0.729
CONCENTRATED	Step regression F3, F4, F6, F12, F29, F36, F43 Kappa = 0.336 A' = 0.715	J48 F17, F21, F27, F38, F41 Kappa = 0.313 A' = 0.638	Step regression F17, F23, F25, F34, F40 Kappa = 0.431 A' = 0.738

Table 4. List of all the features used in the final detectors.

F1: The number of hypothesis variables changed in the last 60 seconds.
F2: The mean of all time taken to change one of the hypothesis variable in the last 60 seconds.
F3: The sum of all time taken to change one of the hypothesis variable in the last 60 seconds.
F4: The number of hypothesis variable changed in the last 5 student actions.
F5: The maximum of all time taken to change one of the hypothesis variable in the last 5 student actions.
F6: The median of all time taken to change one of the hypothesis variable in the last 5 student actions.
F7: The standard deviation of all time taken for hypothesis variable changes in the last 5 student actions.
F8: The minimum of all time taken to change one of the simulation variable in the last 60 seconds.
F9: The maximum of all time taken to change one of the simulation variable in the last 60 seconds.
F10: The median of all the time taken to change the value of a simulation variable in the last 60 seconds.
F11: The mean of all time taken to change one of the simulation variable in the last 60 seconds.
F12: The sum of all time taken changing a simulation variable in the last 60 seconds.
F13: The mean of all time taken to change one of the simulation variable in the last 5 student actions.
F14: The sum of all the time spent on completed trials run in the last 60 seconds.
F15: The minimum of all the time taken executing an incomplete trial in the last 60 seconds.
F16: The number of incomplete trials run in the last 5 student actions.
F17: The sum of all time spent executing trials in the last 60 seconds.
F18: The maximum of all time spent executing a trial in the last 5 student actions.
F19: The sum of all time taken executing trials in the last 5 student actions.
F20: The number of simulation pauses in the last 5 student actions.
F21: The mean of all time spent on simulation pauses in the last 5 student actions.
F22: The mean of all the time taken to execute one of the analysis action in the last 60 seconds.
F23: The sum of all time taken to execute any analysis action in the last 60 seconds.
F24: The number of analysis actions amongst the last 5 student actions.
F25: The mean of all time taken to execute any analysis action in the last 5 student actions.
F26: The standard deviation of all time taken to execute any analysis action in the last 5 student actions.
F27: The number of relevant actions executed in the last 60 seconds.
F28: The minimum of all time taken to execute any relevant action in the last 60 seconds.
F29: The median of all time taken to execute any relevant action in the last 60 seconds.

F30: The standard deviation of all time taken to execute any relevant action in the last 60 seconds.
F31: The sum of all the time taken to execute any relevant action in the last 60 seconds.
F32: The number of relevant actions amongst the last 5 student actions.
F33: The median of all time taken to execute any relevant action in the last 5 student actions.
F34: The standard deviation of all time taken to execute any relevant action in the last 5 student actions.
F35: The probability of knowing how to design controlled exp. before the most recent practice opportunity.
F36: The probability of knowing how to design controlled exp. after the most recent practice opportunity.
F37: The probability of correctly designing a controlled exp. on the most recent practice opportunity.
F38: The probability of knowing how to test stated hypothesis before the most recent practice opportunity.
F39: Whether the student was inactive in the software for the last 60 seconds.
F40: The time elapsed since the last user action at the moment of the observation.
F41: Whether the student entered the exploration stage during this activity.
F42: Whether the student has completed the current stage at least once in a previous activity.
F43: The time elapsed since the start of the current stage.
F44: Whether the text box is currently opened.
F45: The time elapsed since the explanation text box was opened, if it is still opened.
F46: Whether the student closed the text box during the observation.
F47: The ratio of positive and negative assessments by the system for the last 5 student actions.

5 Discussion and Conclusion

In this paper, we presented 12 sensor-free detectors that detect boredom, confusion, frustration, and engaged concentration in the different stages of inquiry in the Inq-ITS environment [17]. This work represents the first automated sensor-free detectors of student affect in simulation microworlds built. Conducting affect detection in a simulation microworld such as Inq-ITS presents different challenges than in other online learning environments. The absence of action-by-action assessment of correctness as in problem-based tutors (e.g. [8]) and the lack of on-demand help (e.g. [8, 9, 10]) hinder the engineering of features similar to those that have proven effective in problem-solving tutors and dialogue tutors such as Cognitive Tutor [8], ASSISTments [9] and AutoTutor [10]. However, other features such as the time spent on different types of actions, the probability that the student knew two key skills [24], and whether the student was inactive in the last 60 seconds, proved useful for this challenge (Table 4).

The non-uniform user interface for the different stages of inquiry also proved to be an important consideration for the generation of affect detectors. Each stage has specific types of actions associated with it and thus patterns of actions related to each affect differ in each stage. This is a general problem for affect detection in learning environments where the student-computer interaction can change considerably from moment to moment. An additional challenge comes from having many observations that co-occur with actions from two stages. In those situations, the interpretation, as an indicator of a specific affect, might differ for the same type of actions depending on whether the action occurred shortly before changing stages or right after changing stage. For these reasons, we created different detectors for each stage of the inquiry process in Inq-ITS. As can be seen in Table 3, few of the best features for individual detectors were reused across multiple stages for the same affect. No features were reused across the BORED detectors, F2 and F20 were reused for CONFUSED, F30 and F46 for FRUSTRATED, and F17 for CONCENTRATED.

The detectors proposed in this paper can be used to study whether specific features of the Inq-ITS system have an impact on the occurrence of affective states. For example, a brief analysis indicates that in our dataset (collected on a prior version of Inq-ITS), 23 out of the 38 observations of frustration (60.53%) occurred when a text box was open or shortly after it was closed. This is more than one would expect as only 36.90% of all the observations matched this condition, and this feature has subsequently been changed in Inq-ITS.

By developing automated detectors that can identify boredom, confusion, frustration, and engaged concentration, we can take a step towards allowing Inq-ITS to effectively adapt to the full range of student's interaction choices during learning and develop interventions that target very specific kinds of disengaged behaviors, as has been successfully done to improve learning in other systems [as in 27, 28, and 29] to offset negative affect states such as boredom.

Acknowledgments. This research was supported by National Science Foundation grant DRL #1008649 awarded to Janice Gobert & Ryan Baker. We thank Michael Wixon, Ermal Toto, and Francis McGeever for their help in pre-processing the data and Jaclyn Ocumpaugh for BROMP training.

References

1. Baker, R.S.J.d., D'Mello, S.K., Rodrigo, M.M.T., 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), 223-241, 2010.
2. Baker, R.S.J.d., et al.: The Dynamics Between Student Affect and Behavior Occurring Outside of Educational Software. *Proceedings of ACII 2011*, 14-24, 2011.
3. D'Mello, S.K., Taylor, R., Grasser, A.C.: Monitoring Affective Trajectories During Complex Learning. *Proceedings of the 29th Annual Cognitive Science Society*, 203-208, 2007.
4. Dragon, T., et al.: Viewing Student Affect and Learning Through Classroom Observation and Physical Sensors. Proceedings of ITS *2008*, 29-39, 2008.
5. Lee, D.M., Rodrigo, M.M., Baker, R.S.J.d., Sugay, J., Coronel, A.: Exploring the Relationship Between Novice Programmer Confusion and Achievement. *Proceedings ACII 2011*, 175-184, 2011.
6. Sabourin, J., Rowe, J., Mott, B., Lester, J.: When Off-Task in On-Task: The Affective Role of Off-Task Behavior in Narrative-Centered Learning Environments. *Proceedings AIED 2011*, 534-536, 2011.
7. Conati, C., Maclaren, H.: Empirically Building and Evaluating a Probabilistic Model of User Affect. *UMUAI*, 19, 267-303, 2009.
8. Baker, R.S.J.d., et al.: Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. *Proceedings of EDM 2012*, 126-133, 2012.
9. Pardos, Z., Baker, R.S.J.d., San Pedro, M.O.Z., Gowda, S.M., Gowda, S.: Affective States and State Tests: Investigating how Affect Throughout the School Year Predicts End of Year Learning Outcomes. *Proceedings of LAK 2013*, 117-124, 2013.
10. D'Mello, S.K., Craig, S.D., Witherspoon, A.W., McDaniel, B.T., Graesser, A.C.: Automatic Detection of Learner's Affect from Conversational Cues. *UMUAI*, 18, 45-80, 2008.

11. Litman, D.J., Forbes-Riley, K.: Recognizing Student Emotions and Attitudes on the Basis of Utterances in Spoken Tutoring Dialogue with Both Humans and Computer-Tutors. Speech Communication, 48 (5), 559-590, 2006.
12. Sabourin, J., Mott, B., Lester, J.: Modeling Learner Affect with Theoretically Grounded Dynamic Bayesian Networks. *Proceedings ACII 2011*, 286-295, 2011.
13. Gobert, J., Sao Pedro, M., Baker, R., Toto, E., Montalvo, O.: Leveraging Educational Data Mining for Real Time Performance Assessment of Scientific Inquiry Skills within Microworlds. *JEDM*, 4 (1), 111-143, 2012.
14. Metcalf, S.J., Kamarainen, A., Grotzer, T.A., Dede, C.J.: Ecosystem Science Learning via Multi-User Virtual Environments. *AERA Conference*, 2011.
15. Hershkovitz, A., Baker, R.S.J.d., Gobert, J., Nakama, A.: A Data-Driven Path Model of Student Attributes, Affect, and Engagement in a Computer-Based Science Inquiry Microworld. *Proceedings of the ICLS*, 2012.
16. NGSS Lead States: *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press, 2013.
17. Sao Pedro, M., Baker, R., Gobert, J., Montalvo, O., Nakama, A.: Leveraging Machine-Learned Detectors of Systematic Inquiry Behavior to Estimate and Predict Transfer of Inquiry Skill. *UMUAI*, 23, 1-39, 2013.
18. Bartel, C.A, Saavedra, R.: The Collective Construction of Work Group Moods. *Administrative Science Quarterly*, 45, 197-231, 2001.
19. Planalp, S., DeFrancisco, V.L., Rutherford, D.: Varieties of Cues to Emotion in Naturally Occurring Situations. *Cognition and Emotion*, 10 (2), 137-153, 1996.
20. Ocumpaugh, J., Baker, R.S.J.d., Rodrigo, M.M.T.: *Baker-Rodrigo Observation Method Protocol (BROMP) 1.0 Training Manual version 1.0*. Technical Report, New York, NY: EdLab, Manila, Philippines: Ateneo Laboratory for the Learning Sciences, 2012.
21. Litman, D.J., Forbes-Riley, L.: Recognizing Student Emotions and Attitudes on the Basis of Utterances in Spoken Tutoring Dialogues with Both Human and Computer Tutors. *Speech Communication*, 48 (5), 559-590, 2006.
22. Rodrigo, M.M.T., et al.: Comparing Learners' Affect While Using an Intelligent Tutoring Systems and a Simulation Problem Solving Game. *Proceedings of ITS 2008*, 40-49, 2008.
23. Baker, R.S.J.d., 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, 2004.
24. Sao Pedro, M., Baker, R., Gobert, J., Montalvo, O., Nakama, A.: Levaraging Machine-Learned Detectors of Systematic Inquiry Behavior to Estimate and Predict Transfer of Inquiry Skill. *User Modeling and User-Adapted Interaction*, 23, 1-39, 2013.
25. Cohen, J.: A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20 (1), 37-46, 1960.
26. Hanley, J., McNeil, B.: The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, 143, 29-36, 1982.
27. Woolf, B.P., Arroyo, I., Cooper, D., Burleson, W., Muldner, K.: Affective Tutors: Automatic Detection of and Response to Student Emotion. *Advances in Intelligent Tutoring Systems*, 207-227, 2010.
28. Lehman, B.A., et al.: Inducing and Tracking Confusion with Contradictions During Complex Learning. *IJAIED*, 22 (2), 85-105, 2013.
29. Rai, D., et al.: Repairing Deactivating Negative Emotions with Student Progress Pages. *Proceedings of AIED 2013*, 795-798, 2013.