Boredom Across Activities, and Across the Year, within Reasoning Mind

William L. Miller
Reasoning Mind
2000 Bering Dr. Ste. 300
Houston, TX 77057
neal.miller@reasoningmind.org

Ryan S. Baker
Teachers College, Columbia University
525 W. 120th St
New York, NY 10027
1-212-678-8329
baker2@exchange.tc.columbia.edu

Matthew J. Labrum
Reasoning Mind
2000 Bering Dr. Ste. 300
Houston, TX 77057

Karen Petsche
Reasoning Mind
2000 Bering Dr. Ste. 300
Houston, TX 77057

Angela Z. Wagner
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213

ABSTRACT
Since the beginning of educational assessment, there has been interest in modeling more than just what a student knows. However, the ease of measuring relatively simple knowledge and skills through multiple-choice tests has led to a field that has largely focused on what is easy to measure. In recent years, there has been increasing interest in automated assessment of students in a broader range of contexts, and for a broader range of constructs, than traditional assessment achieves. In this paper, we present automated assessment that can infer boredom, a key non-cognitive factor during student learning. We study this construct in the context of Reasoning Mind, a blended learning system that integrates a range of learning activities, embedding them into story-based curricula. We asses student boredom across the year using sensor-free automated detectors developed using a combination of quantitative field observations and data mining, validated to generalize across students and school contexts. We then apply these detectors to data from the entire cohort of 70,000 students who used Reasoning Mind during the 2012-2013 school year. We demonstrate the use of the detectors by investigating how student boredom while using the Reasoning Mind blended learning program changes over the course of the year, and how different objectives and activities result in more or less student boredom. We find that while there is essentially no relationship between student boredom and the time of year, it is strongly related to the objectives in the self-paced Reasoning Mind curriculum; in particular, boredom on Reasoning Mind objectives is bimodal in character, with some objectives inducing considerable boredom while others do not.

Categories and Subject Descriptors
K.3.1 [**Computers and Education**]: Computer Uses in Educations – *computer-managed instruction.*

General Terms
Algorithms, Measurement

Keywords
Boredom, Affect detection, Curriculum

1. INTRODUCTION

Since the beginning of educational assessment, there has been interest in modeling more about a student than just what they know. For instance, as early as 1948, the first President of Educational Testing Services suggested measuring personal drive, motivation, conscientiousness, interpersonal skill, and interest [23], and there were serious attempts to measure personality and motivation starting from the 1960s [22].

However, the ease of measuring relatively simple knowledge and skills through multiple-choice tests has led to a field that has largely focused on what is easy to measure. While criticism of this type of measure has been prevalent since the 1960s (e.g. [17]), the clean psychometric properties of multiple-choice and ease of validation have made it the dominant choice for educational assessment to this day (see, for instance, [18, 19, 26]). As a result, educational assessment tends to focus on what is easily measured through multiple-choice tests.

In recent years, there has been increasing interest in automated assessment of students in a broader range of contexts, and for a broader range of constructs, than traditional assessment achieves (cf. [3, 9, 25, 35, 39, 40]).

One construct that has emerged as a focus of research in the last decade is boredom. Boredom is of particular importance due to evidence that this affective state is associated with negative learning outcomes [31, 34], more so than other affective states such as frustration and confusion [11, 30]. Boredom can be detected in students using physical sensors (cf. [12]); however, recent work by Baker et al. [4] and Pardos et al. [30] has resulted in the construction of detectors of student boredom based only upon interactions between students and educational software. We apply these techniques to study boredom in the Reasoning Mind Genie 2 system.

The Reasoning Mind Genie 2 system [21] is a self-paced blended learning mathematics curriculum for elementary and middle school students (current offerings cover the second through the sixth grades), which is implemented within classrooms. Reasoning Mind combines extensive professional development, a rigorous curriculum drawing from successful curricular design in Russia, and a game-like, internet-based interface. Student learning in Reasoning Mind takes place in "RM City," a virtual city where students engage in learning activities in different "buildings." The primary mode of study for students is "Guided Study," wherein they are guided by a pedagogical agent named "Genie" through a series of learning objectives. Interspersed with the instructional and problem-solving modes are Speed Games, which are timed

mental math problems designed to help students develop fluency in fundamental operations. It is used by approximately 100,000 students a year, primarily in the Southern United States. The fifth and sixth grade curricula are "core" curricula; they replace the traditional mathematics class and are generally used for the students' entire scheduled mathematics instruction time, usually 3-5 days per week for 45-90 minutes each day.

Given the key importance of engaging students and avoiding boredom for longer-term student outcomes [22, 36], a great deal of time and expense has been dedicated to making the Reasoning Mind content engaging. Instruction and problems are presented alongside colorful graphics and entertaining stories featuring a cast of characters with which the student grows familiar. A recent study using Quantitative Field Observations (QFOs) in the context of Reasoning Mind classroom found evidence that students find the system highly engaging; students had unusually high rates of on-task behavior and engaged concentration [27].

However, like any curriculum, Reasoning Mind has room for improvement. Reducing negative affect is a key goal, with boredom a particular focus due to its association with worse student outcomes. Sensor-free boredom detectors such as those described above can infer student affect moment-by-moment, allowing detailed analysis of which contexts are associated with improved student affect. As such, they provide a measurement which is formatively useful not just for assessing the student, but for assessing the curriculum as well. For instance, Doddannara and colleagues [13] have studied which features of the design of Cognitive Tutors are associated with differences in student boredom. In studying these issues, however, it is important to be certain that differences in affect seen are due to the content and not just to the time of year; for instance, Beck [6] found that hasty guessing (a behavior associated with boredom [2]) was more frequent later in the year than earlier in the year.

As such, in this paper, we present a method for automated, sensor-free assessment of boredom within Reasoning Mind, developed and validated according to current best-practices. We then apply these detectors to an entire year of data of Reasoning Mind usage, and use the detectors to assess how student affective profile varies across the school year and in different objectives within the Reasoning Mind fifth grade curriculum.

2. METHODS

2.1 Data Set

Detectors of student boredom were constructed based on field observations of students in Reasoning Mind and log data from the Reasoning Mind system, which was synchronized to the field observations.

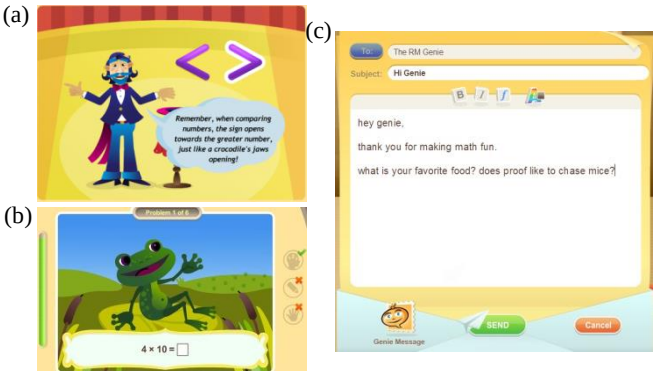

Figure 1 Examples of Reasoning Mind Genie 2 content: (a) Theory content, (b) a Speed Game problem, (c) an example of writing a letter to the "Genie".

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, and on boredom in specific. The coders used the HART app on a Google Android handheld computer, which enforced the BROMP protocol [29], an observation protocol developed specifically for the process of coding behavior and affect during use of educational software. As of this writing, there are 63 coders certified in the BROMP protocol, and the BROMP protocol has now been used in dozens of papers (see review in [21]). All coding was conducted by the third, fourth, and fifth authors. These three coders were previously trained in coding behavior and affect using the BROMP protocol, and all achieved inter-rater reliability with the trainer of greater than 0.6 during training, on par with past projects [cf. 2, 5, 24, 33].

Observations were conducted during the student's regular math class, where students typically use the Reasoning Mind software. 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 behavior 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 handheld observation software. Setting up observations took a few minutes at the beginning of each class.

Each observation lasted up to twenty seconds, with observation time automatically coded by the handheld observation software. If behavior was determined before twenty seconds elapsed, the coder moved to the next 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 [2, 33], in order to make it less clear when an observation was occurring. This method of observing was previously found to be successful for assessing student behavior and affect, achieving good inter-rater reliability [2, 33], and forms the basis of the BROMP protocol. To increase tractability of both coding and eventual analysis, if two distinct affective states were seen during a single observation, only the first affective state observed was coded. Any behavior involving a student other than the student currently being observed was not coded.

The observers based their judgment of a student's affect on the student and teacher's work context, actions, utterances, facial expressions, body language, and interactions with others in the room. These are, broadly, the same types of information used in previous methods for coding affect [5], in line with Planalp et al.'s [32] descriptive research on how humans generally identify affect using multiple cues in concert for maximum accuracy rather than attempting to select individual cues. Because detecting affect is a complex problem involving several interconnected cues, the BROMP protocol takes advantage of humans' intuitive ability to determine affect in others, rather than proscribing exhaustive definitions of what an observer should consider. Within an observation, each observer coded affect with reference to five categories, drawn from [2]:

- Concentrating
- Bored
- Frustrated
- Confused

Table 1. Regions and demographic information for schools included in this study.

	Region	Free/Reduced Price Lunch	White	African-American	Hispanic
1	Texas (Urban)	85%	1%	84%	13%
2	Texas (Urban)	79%	3%	32%	63%
3	Texas (Urban)	96%	1%	10%	88%
4	Texas (Suburban)	48%	24%	50%	17%
5	Texas (Suburban)	33%	52%	24%	16%
6	West Virginia (Rural)	51%	80%	16%	1%

- "?" (which refers to any affect outside of the coding scheme, such as delight, or any case where it was impossible to code student affect)

To increase the probability of model generalizability (cf. [28]), data was collected across the span of several months from a diverse sample of students, representative of the population currently using Reasoning Mind. Five of the six schools were in the Texas Gulf Coast region. Three of these Texas schools were in urban locations and served economically disadvantaged populations (defined as a high proportion of students receiving free or reduced lunch); of these three, two served predominantly African-American student populations, and one served a predominantly Hispanic student population. The other two schools in this region were in suburban locations, one serving mostly White students, and the other with a mix of student ethnicities; both of these schools had a lower proportion of economically disadvantaged students. The sixth school was a rural school in West Virginia, with an economically disadvantaged, majority White population. See Table 1 for more detailed information about the observed schools.

These observations were synchronized with the system logs of the students working through the Reasoning Mind system, by synchronizing both the HART application and the Reasoning Mind system to the same internet time server, leading to synchronization error of under 1 second. The resulting data set consisted of 4891 distinct observations of student behavior for 408 students (mean = 12.0 observations per student, standard deviation = 6.9), coded by three observers across six separate days.

2.2 Feature Distillation

For each observation, a clip was computed from the log data which matched as closely as possible to the observation (20 seconds before observation entry time to observation entry time) (cf. [4, 30]), facilitated by the log synchronization procedure discussed above. Using the student's activities both within the twenty-second window and preceding it (but not using the future), 93 features were developed. Some features – for example, whether an action was correct or not, or how long the action took – were computed for each action in the clip and then aggregated across the clip (see next paragraph for details). Others – for example, the fraction of previous attempts on the current skill the student has gotten correct – are based on the student's complete activity from the beginning of the school year. A third category involves the results of other models applied to the student log (a form of discovery with models (cf. [15])). For example, the probability that the student knows the current skill (from Bayesian Knowledge Tracing [10]), student carelessness [37], and features of the student's moment-by-moment learning graph [3, 16] were all included as features.

These 93 features were then aggregated across actions in the clip by a variety of methods, depending on the nature of the feature: mean, min, max, standard deviation, sum, presence (for example, '1' if there was any "problem" item type in the clip), count, and proportion (by count or by time; for example, what proportion of the actions in the clip were "problem" item types, and what proportion of the time within the clip was spent on "problems"). The result was a total of 278 features used to develop the boredom detector; examples are given in Table 3.

2.3 Machine Learning Approach

Detectors were built for each of the affective states described above; for reasons of scope, the current paper focuses on the detector for boredom (the other detectors achieved comparable or higher model goodness than boredom). Detector evaluation was conducted using ten-fold student-level cross-validation, whereby students were randomly split into ten groups and a detector was developed using data from nine of the groups and then tested on the remaining group of students, for each possible combination. Cross-validation at this level reduces concerns about over-fitting to specific students, and increases confidence that the detectors will generalize to new students.

Data were re-sampled to have more equal class frequencies before machine learning techniques were applied (cf. [13]). However, all calculations of model goodness were performed on the original data set.

Four algorithms were tried: JRip, J48 decision trees, step regression, and Naïve Bayes. We found that step regression – linear regression turned into a binary classifier with a step function applied at a pre-chosen threshold – was most successful.

Feature selection was via forward selection. In this selection scheme, features are added one at a time (starting from the empty set), each time selecting the feature that most improves cross-validated detector goodness. For the purposes of feature selection, detector goodness was defined as the value of A′ [14] (see description below) as measured on the original data set. Features are added until no single feature can be added to further improve the goodness of the detector. To reduce the potential for over-fitting, a first pass was performed in which any feature that yielded A′ below 0.5 (chance) in a single-feature model were removed from the set of possible features.

A′, sensitivity, and specificity were used to assess detector goodness. A popular alternative, Cohen's Kappa, is not recommended for highly skewed data such as seen in this data set – [20]. A′ is the probability that, given one example from each class (i.e. BORED and NOT BORED), the model can correctly

Table 2. Confusion matrix for boredom detector.

		Detector	
		BORED	NOT BORED
Truth	BORED	163	151
	NOT BORED	796	1187

identify which is which. It is mathematically equivalent to the area under the ROC curve (AUC) used in signal detection and to W, the Wilcoxon statistic [14]. A value of 0.5 for A′ indicates performance exactly at chance, and a value of 1 indicates perfect performance. In these analyses, A′ was calculated at the level of clips, rather than students. A′ was calculated using Baker et al.'s "Simple A"' calculation code [1], available from http://columbia.edu/~rsb2162/edmtools.html. This code calculates this metric using the Wilcoxon approach; the alternative approach of integrating the area under the curve using calculus leads to bugs for special cases in all known implementations.

3. RESULTS OF DETECTOR VALIDATION

The detector of boredom achieved cross-validated A′ = 0.64, moderately better than chance. A confusion matrix for this detector is shown in Table 2. The detector was somewhat better at getting relative ordering (A′) than making absolute distinctions; the sensitivity of the detector was 0.52, and the specificity was 0.60. The low sensitivity compared to specificity implies that the detector is somewhat conservative, favoring a low false positive rate at the expense of a higher false negative rate; however, given the low occurrence of boredom, the overall false negative rate is still quite low. The detector captures a substantial amount of occurrence of boredom, and the goodness measures described make it appropriate for fail-soft interventions – interventions which are not detrimental when applied to students who are not bored.

The final detector is shown in Table 3. Boredom is positively correlated with the standard deviation of problem correctness (indicating the student was somewhat erratic across the clip) and to the minimum slip parameter (from Bayesian Knowledge Tracing) on skills in the clip (indicating that the student is working through material that students tend to get right if they know it). It is also negatively correlated with both the number of actions within the clip that occurred a Speed Game and the fraction of the clip duration spent in a Speed Game, which students may find more entertaining or exciting than other content.

4. Analysis

After construction of the detectors, they were applied to the log data for the observed classes for the entire 2012-2013 academic year; this data set was comprised of 2,974,944 actions by 462 students, including 54 students who were not present when the classes were observed, either because they were absent or because they transferred into the class after the observations were performed. While these detectors are likely to work for the full sample of Reasoning Mind students given the diverse population used to build them, applying them only to the original population is a conservative choice that increases the probability of model validity where used.

Table 3. The final boredom detector.

Coefficient	Feature
+0.212	The standard deviation, across the clip, of student correctness (1 or 0) on each action.
-0.013	The number of actions in the clip that occurred on Speed Game items.
-0.070	The fraction of the total clip duration spent on Speed Game items.
-0.073	The number of actions in the clip on items where the answer input was made by selecting an item from a drop-down list.
+0.290	The minimum slip parameter (P(S) in Bayesian Knowledge Tracing) on skills in the clip.
-0.260	The standard deviation, across the clip, of the action duration, normalized across all students, times the presence (1) or absence (0) of a hint request on the previous action.
+0.123	Y-intercept.

The boredom detector developed here, by virtue of its goodness values that are above chance but fairly weak, is better for aggregate discovery with models analyses [15] than for individually-targeted interventions selected with a single cut-off. In using these detectors within discovery with models analyses, the extra information contained in the probability estimates, and the superior performance of A′ than sensitivity and specificity indicates that it is preferable to use exact probabilities from the model rather than resolving the probabilities into binary predictions.

We use the boredom detector to investigate student affect across the school year and across activities in the curriculum, as discussed in the Introduction. A plot of boredom vs. date is shown in Figure 2(a). Note that, with the exception of a spike in boredom early in the school year, the plot is relatively featureless. There is a very weak negative correlation (-0.06) between boredom on a given action (aggregated by student) and the date on which it occurred (e.g. how far along the year it was). Predicting boredom by date during the year (controlling for student to avoid violating independence assumptions) results in a model that is statistically significant, $R^2 = 0.015$, $p < 0.001$.

Because the curriculum is self-paced, students are working through different content at any given point in the year. Average boredom vs. Reasoning Mind objective is plotted in Figure 2(b). The Reasoning Mind Basic II (fifth grade) curriculum is made up of 45 objectives, indicated in the plot at 5.01 through 5.45. Each student sees the objectives in the same order. Reasoning Mind objectives are highly interconnected, and each objective builds on those preceding it. This figure and the previous are scaled with the same y-axis to make clear the difference in variation observed; note that the range of average boredom values observed is much larger when averaged over objectives than over days, despite that fact that each data point in the objectives plot contained significantly more individual student actions. Attempting to predict boredom by objective (controlling for student to avoid violating independence assumptions) yields a model which is statistically significant, $p < 0.001$. The resultant model has a model with a correlation of 0.343.

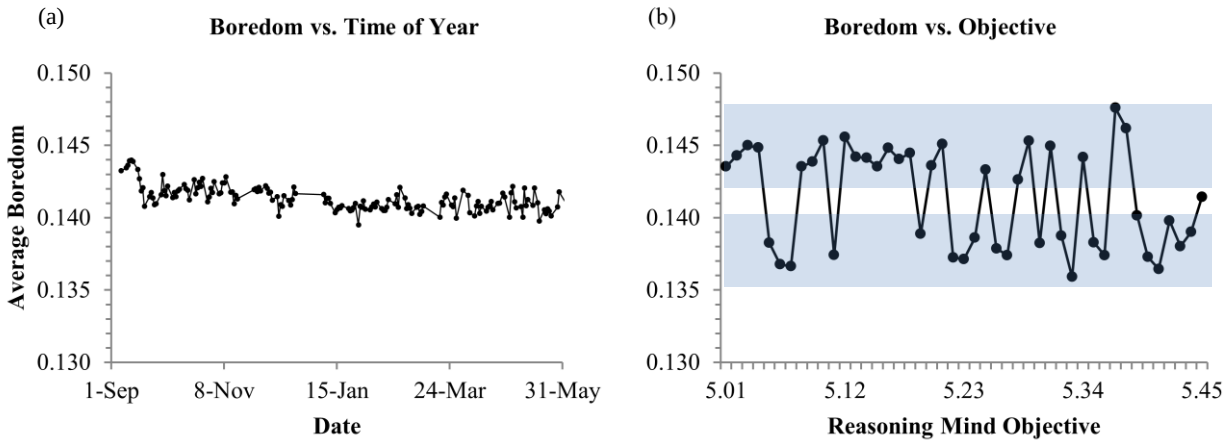

Figure 2. (a) Average boredom across all student actions for each day between September 1, 2012 and May 31, 2013 with at least 1,000 student actions. (b) Average boredom across all student actions in each of the 45 Reasoning Mind Basic II objectives. 5.01 is the first objective of the 5th grade curriculum, and so on. The two clusters of boredom are highlighted by blue bands.

Furthermore, the average boredom by objective shows a clear bimodal character: the average boredom value seems to alternate between two groups, one, comprising 21 objectives, with an average boredom of 0.138 and the other, comprising 23 objectives, with an average boredom of 0.145. The final objective (5.45) falls roughly in the middle of the two extremes (average boredom = 0.141). Assignment of student actions to groups based on objective (as described above, discarding objective 5.45) yields a moderate correlation between objective "group" (high, low) and boredom of 0.31 and a Cohen's *d* value of 0.67, indicating that the difference in boredom between the two groups is 0.67 standard deviations. Including both the date and the objective in a generalized linear fit to the data (controlling for the student to avoid violating independence assumptions) results in a small improvement in the in fit (R^2 = 0.1182 vs. 0.1177 for objective alone). Both objective and date are statistically significant ($p < 0.001$) in such a model: for both objective and date, $p<0.001$. An analysis of the two groups of objectives did not reveal obvious explanatory differences; however, further analysis is ongoing to determine what may explain the grouping.

Because student work in the Reasoning Mind Genie 2 system is self-paced, the objectives worked on by students quickly diverged as the year went on. Thus, early in the school year, when most students are on the first four lessons (all of which were in the "high" boredom group) there was a peak in boredom, but as students diverged, boredom converged to an average value. Note also that the source of the mild negative correlation between time of year and boredom is now revealed; the "high" boredom lessons are more concentrated in the early part of the curriculum (16 of the first 21 objectives) than in the latter half (only 7 in the remaining 24). Because students are, on average, farther into the curriculum near the end of the school year, the result is a decrease in boredom over time.

5. Conclusions

In this paper, we have constructed automated, sensor-free detectors of student affect within the Reasoning Mind curriculum. These detectors achieve moderate goodness, with an A′ value of 0.64. We apply these detectors to the entire school year of data (having validated them on data collected across the year). We then use the detectors to study student boredom across the school year, and across learning objectives.

When the detector of student boredom was applied across the school year, a weak negative correlation (-0.06) with date was found. However, when student activity was grouped by objective rather than by date, a distinct bimodal character in student boredom was found, which divided the curriculum into two similar-sized groups.

This assessment of student boredom across Reasoning Mind objectives allows the evaluation of the student and the curriculum at a qualitatively different level than simple multiple-choice assessments of simple knowledge and skills. Though affect has been studied through Likert scale items [31], it would not be feasible to administer these items at the scale needed to study the research questions investigated here. This research is only feasible with next-generation assessment such as automated affect detectors.

As such, a key future goal for research will be to identify the characteristics of low boredom and high boredom objectives. One method to doing so is to conduct a systematic analysis of the design features and content of each objective (cf. [13]). By understanding which lessons are most boring and why, it will become possible to iteratively improve the "high boredom" lessons in Reasoning Mind Basic II and to guide future development of Reasoning Mind curriculum – reducing boredom, and potentially improving both learning and long-term outcomes.

6. ACKNOWLEDGMENTS

We would like to thank George Khachatryan for helpful discussions and suggestions, Maria Nosovskaya and Caitlin Watts for their support in data collection, and the Bill and Melinda Gates Foundation for their generous support for this collaboration.

7. REFERENCES

[1] 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. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (2008), 406-415.

[2] Baker, R.S.J.d., D'Mello, S.K., Rodrigo, M.M.T., and Graesser, A.C. Better to Be Frustrated than Bored: The. *International Journal of Human-Computer Studies*, 68, 4 (2010), 223-241.

[3] Baker, R.S.J.d., Goldstein, A.B., and Heffernan, N.T. Detecting Learning Moment-by-Moment. *International Journal of Artificial Intelligence in Education*, 21, 1-2

(2011), 5-25.

[4] Baker, R.S.J.d., Gowda, S.M., Wixon, M. et al. Towards Sensor-free Affect Detection in Cognitive Tutor Algebra. In *Proceedings of the 5th International Conference on Educational Data Mining* (2012), 126-133.

[5] Bartel, C.A. and Saavedra, R. The collective construction of work group models. *Administrative Science Quarterly*, 1, 1 (2009), 3-17.

[6] Beck, J. E. Engagement tracing: using response times to model student disengagement. *Artificial intelligence in education: Supporting learning through intelligent and socially informed technology*, 125, 88 (2005).

[7] Calvo, R. Latent and Emergent Models in Affective Computing. *Emotion Review*, 2, 3 (2010), 288-289.

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

[9] Conati, C. and Maclaren, H. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, 19, 3 (2009), 267-303.

[10] Corbett, T., Albert, and Anderson, John R. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4 (1995), 253-278.

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

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

[13] Doddannara, L., Gowda, S., Baker, R. S. J. d., Gowda, S., and de Carvalho, A. M. J. B. Exploring the relationships between design, students' affective states, and disengaged behaviors within an ITS. In *Proceedings of the 16th International Conference on Artificial Intelligence and Education* (2013), 31-40.

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

[15] Hershkovitz, A., Baker, R.S.J.d., Gobert, J., Wixon, M., and Sao Pedro, M. Discovery with Models: A Case Study on Carelessness in Computer-based Science Inquiry. *American Behavioral Scientist*, 57, 10 (2013), 1479-1498.

[16] Hershkovitz, A., Baker, R.S.J.d., Gowda, S.M., and Corbett, A.T. Predicting Future Learning Better Using Quantitative Analysis of Moment-by-Moment Learning. In *Proceedings of the 6th International Conference on Educational Data Mining* (2013), 74-81.

[17] Hoffman, B. (1962). Towards less emphasis on multiple-choice tests. *The Teachers College Record*, 64(3), 183-183.

[18] Hoover, H.D., Dunbar, S.B., and Frisbie, D. A. (2008) Iowa Tests of Basic Skills Form C. Rolling Meadows, IL: Riverside Publishing.

[19] International Association for the Evaluation of

Educational Achievement (IEA) (2011). TIMMS 2011 Assessment. Available from http://nces.ed.gov/TIMSS/educators.asp

[20] Jeni, L. A., Cohn, J. F., & De La Torre, F. Facing Imbalanced Data--Recommendations for the Use of Performance Metrics. In *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)* (2013) 245-251.

[21] Khachatryan, G., Romashov, A., Khachatryan, A., Gaudino, S., Khachatryan, J., Guarian, K., and Yufa, N.. Reasoning Mind Genie 2: An Intelligent Learning System as a Vehicle for International Transfer of Instructional Methods in Mathematics. *International Journal of Artificial Intelligence in Education* (2014) 333-382.

[22] Leading, L. L., & Kyllonen, P. C. (2005). The Case for Noncognitive Assessments.

[23] Lemann, N. (1995, September). The great sorting. *Atlantic Monthly*, 276(3), 84-100.

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

[25] Matthew, C. T., & Stemler, S. E. (2013). Assessing mental flexibility with a new word recognition test. *Personality and Individual Differences*, 55(8), 915-920.

[26] New York State Education Department (2010). New York State Regents Exam. Available from http://www.nysedregents.org/

[27] Ocumpaugh, J., Baker, R.S.J.d., Gaudino, S., Labrum, M.J., and Dezendorf, T. Field Observations of Engagement in Reasoning Mind. In *Proceedings of the 16th International Conference on Artificial Intelligence and Education* (2013), 624-627.

[28] Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., and Heffernan, C. Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Educational Technology* (in press).

[29] Ocumpaugh, J., Baker, R.S.J.d., and 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, 2010.

[30] Pardos, Z.A., Baker, R.S.J.d., San Pedro, M.O.C.Z., Gowda, S.M., and Gowda, S.M. Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (2013), 117-124.

[31] Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., & Perry, R. P. (2010). Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology*, 102(3), 531.

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

[33] Rodrigo, M.M.T., Baker, R.S.J.d., D'Mello, S. et al. Comparing Learners' Affect While Using an Intelligent Tutoring Systems and a Simulation Problem Solving Game. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (2008), 40-49.

[34] Rodrigo, M. M. T., Baker, R. S. J. d., Lagud, M. C. V. et al. Affect and Usage Choices in Simulation Problem-Solving Environments. *Artificial Intelligence in Education*, 158 (2007), 145-152.

[35] Sabourin, J., Mott, B., and Lester, J. C. Modeling learner affect with theoretically grounded dynamic Bayesian networks. In *Affective Computing and Intelligent Interaction*. Springer, Berlin Heidelberg, 2011.

[36] San Pedro, M.O.Z., Baker, R.S.J.d., Bowers, A.J., and Heffernan, N.T. Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. In *Proceedings of the 6th International Conference on Educational Data Mining* (2013), 177-184.

[37] San Pedro, M.O.C., Rodrigo, M.M., and Baker, R.S.J.d. The Relationship between Carelessness and Affect in a Cognitive Tutor. In *Proceedings of the 4th bi-annual International Conference on Affective Computing and Intelligent Interaction* (2011).

[38] Shih, B., Koedinger, K., and Scheines, R. A Response Time Model for Bottom- Out Hints as Worked Examples. In *Proceedings of the First Educational Data Mining Conference* (2008).

[39] Srikant, S., & Aggarwal, V. (2013). Automatic Grading of Computer Programs: A Machine Learning Approach. In *12th International Conference on Machine Learning and Applications (ICMLA)* (2013), 85-92.

[40] Ventura, M., Shute, V., and Zhao, W, The relationship between video game use and a performance-based measure of persistence. *Computers & Education*, (2013), 52-58.