Automated Detection of Proactive Remediation by Teachers in Reasoning Mind Classrooms

William L. Miller
Reasoning Mind,
Houston, TX
wlmiller@gmail.com

Ryan S. Baker
Teachers College,
Columbia University,
New York, NY
baker2@
exchange.tc.
columbia.edu

Matthew J. Labrum,
Karen Petsche,
Yu-Han Liu
Reasoning Mind,
Houston, TX

Angela Z. Wagner
Human-Computer
Interaction Institute,
Carnegie Mellon
University, Pittsburgh, PA
awagner@cmu.edu

ABSTRACT
Among the most important tasks of the teacher in a classroom using the Reasoning Mind blended learning system is proactive remediation: dynamically planned interventions conducted by the teacher with one or more students. While there are several examples of detectors of student behavior within an online learning environment, most have focused on behaviors occurring fully within the context of the system, and on student behaviors. In contrast, proactive remediation is a teacher-driven activity that occurs outside of the system, and its occurrence is not necessarily related to the student's current task within the Reasoning Mind system. We present a sensor-free detector of proactive remediation, which is able to distinguish these activities from other behaviors involving idle time, such as on-task conversation related to immediate learning activities and off-task behavior.

1. INTRODUCTION
In recent years, researchers in learning analytics and educational data mining have been successful at detecting a range of student behaviors during the use of online and blended learning systems, including whether the student is gaming the system [1], engaging in behaviors not related to the learning task [2], exploring the learning environment [3], or avoiding help [4]. These detectors in turn have supported both automated intervention [5,6] and "discovery with models" analyses [1,7,8,9]. As these behaviors are manifested entirely within student interaction with the learning system, it is feasible to detect these behaviors solely from logs of student interaction with the system.

LAK '15, March 16 - 20, 2015, Poughkeepsie, NY, USA

ACM 978-1-4503-3417-4/15/03□$15.00
http://dx.doi.org/10.1145/2723576.2723607

In recent years, this work has been extended to also include detection of behaviors not entirely occurring within the system, such as off-task behavior outside the learning system [10], and students' affective states [9,11]. These results indicate that log files contain a great deal of information that can be used for inference about constructs and behaviors beyond just student behaviors within the learning system.

In this paper, we demonstrate that interaction log files can also be used to make inference about on-task, education-related interactions between a student and an instructor, completely outside of the learning environment. In this paper, we focus on proactive remediation, dynamically planned intervention by the teacher with one or more students. In a proactive remediation, the teacher decides to provide help to one or more students on a topic that they are not currently struggling with, based on evidence that the student(s) need to learn that topic. The teacher plans such interventions using formative assessment data provided by the tutoring system. Proactive remediation is different from the traditional view of on-task conversations in blended learning [10], where the student discusses the current material being presented in the tutor, with another student or the teacher.

In this paper, we describe the construction of a detector of proactive remediation for the Reasoning Mind Genie 2 system [12]. The Reasoning Mind Genie 2 system is a blended learning mathematics curriculum for elementary and middle school students (current offerings cover the second through the sixth grades), which is implemented within classrooms with a teacher present. 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. 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.

The teacher's role within the Reasoning Mind classroom is a crucial one; he or she provides vital support to students beyond what is provided by the online system. Reasoning Mind provides extensive professional development both in mathematical content knowledge

and in effectively running a Reasoning Mind classroom. Within the professional development materials, teachers are taught that one of the most important activities of a teacher during a Reasoning Mind classroom session is proactive remediation. The Genie 2 system provides rich and detailed student metrics to the teacher, distilled using learning analytics; the teacher is trained to use these data to plan one-on-one and small group interventions with their students. In order to further support the teachers in these activities, teachers are assigned implementation coordinators, who answer any questions the teacher may have as well as helping them develop classroom strategies. Implementation coordinators also visit teacher classrooms throughout the year to observe and give feedback to the teacher on how to teach with Reasoning Mind more effectively. These implementation coordinators have been able to help teachers develop proactive remediation strategies and other strategies for supporting students, but their visits and services are resource-intensive, and difficult to scale. By automatically detecting proactive remediation, it may be feasible to determine how much and when teachers engage in this behavior, towards giving a greater degree of feedback to teachers without having to send an implementation coordinator to the school for a day.

At a basic level, one would expect proactive remediation to look similar to on-task conversation and off-task behavior in the student log files; all three are likely to involve extended periods of student inactivity within the system. However, the log activities leading up to a proactive remediation are likely to be quite different than those leading to off-task behavior and on-task conversation (as when comparing off-task behavior and on-task conversation to each other [cf. 10]), enabling us to distinguish proactive remediation from other events and behaviors. It would be ideal to use a combination of log data on both student and teacher interactions in developing these detectors; however, the teacher interactions are not yet fully instrumented. In this paper, therefore, we study proactive remediation working solely from student log files.

2 METHODS

2.1 Data Set

A detector of proactive remediation by teachers was 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.

A recent study using the BROMP (Baker-Rodrigo-Ocumpaugh Monitoring Protocol [14]) protocol for quantitative field observations found evidence that students find Reasoning Mind highly engaging; specifically, this study found high rates of on-task behavior and engaged concentration among students working in the Reasoning Mind system [13]. The same method was used in this study to observe students in a total of six schools for a different purpose, to develop an automated detector of proactive remediation. The BROMP protocol has been used in a variety of contexts; as of this writing, there are 129 BROMP-certified coders.

Expert field observers coded student affect and engaged/disengaged behaviors as students used the learning software, using the BROMP protocol. The coders used the HART app on a Google Android handheld computer [11], which enforced the BROMP protocol [14], an observation protocol developed specifically for the process of coding behavior and affect during use of educational software.

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 [15,18], 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 [15,18]. To increase tractability of both coding and eventual analysis, if two distinct behaviors were seen during a single observation, only the first behavior 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 state or behavior 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 [16], and in line with Planalp et al.'s [19] descriptive research on how humans generally identify affect using multiple cues in concert for maximum accuracy rather than attempting to select individual cues. Within an observation, each observer coded behavior with reference to five categories:

- On-Task
- Off-Task
- Proactive Remediation
- On-Task Conversation
- "?" (which refers to any behavior outside of the coding scheme or any case where it was impossible to code student behavior)

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 achieved inter-rater reliability with the trainer of 0.66, 0.72, and 0.83, during training, on par with past projects [15,16,17,18].

To increase the probability of model generalizability, data was collected 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.

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%

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, coded by three observers across six separate days.

After construction of the detectors, they were applied to the log data for 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.

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. 11, 9], 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 (also called discovery with models [cf. 7]). For example, the probability that the student knows the current skill (from Bayesian Knowledge Tracing [20]), student carelessness [21], and features of the student's moment-by-moment learning graph [22,8] 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 a detector of proactive remediation; examples are given in Table 2.

2.3 Machine Learning Algorithms

Detectors were built for PROACTIVE REMEDIATION, described above. Detector evaluation was by ten-fold student-level evaluation, 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, for each possible combination. Cross validation at this level reduces concerns about over-fitting to specific students, and ensures the generalizability of detectors to new students.

Because proactive remediation is a relatively rare occurrence (proactive remediation represented about 0.8% of all observations), data were re-sampled (e.g. cloning data within the minority class) to have more equal class frequencies before machine learning techniques were applied. 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′ [23] (see description below) as measured on the original data set. Features are added in the way 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 in a single-feature model were removed from the set of possible features.

A′ and Cohen's Kappa [24] were used to assess detector goodness. A′ is the probability that, given one example from each class (i.e. PROACTIVE REMEDIATION and NOT PROACTIVE REMEDIATION), the model can correctly 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 [23]. 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 [25], available from http://www.columbia.edu/~rsb2162/edmtools.html. Cohen's kappa is a measure of the degree to which the detector is better than chance at identifying which clips involve the behavior of interest. A Kappa of 0 indicates performance at chance (according to the base rate), and a Kappa of 1 is perfect performance; intermediate values indicate how much better (as a percentage) the detector is than chance.

3. RESULTS

The detector for proactive remediation appeared successful in terms of A′, with an overall cross validated A′ of 0.90. When Kappa was calculated, the results initially appeared poor -- Kappa was 0.06. This difference was sufficiently surprising that we re-checked A′ by hand and found that it had been computed correctly.

Upon further examination, we noted that the average confidence for proactive remediation was 2.1%, while the average confidence for the other examples was 0.6%. As such, almost all clips were assessed by the detector as having confidence below 50%. Confidences that are systematically too low or high can be adjusted post-hoc by rescaling or by allowing the threshold to vary (these are mathematically

equivalent). In this case, if we choose an optimal threshold of 0.013947 (using the evolutionary equation solver in Microsoft Excel), a much better Kappa of 0.65 is achieved.

As such, the model is successful at distinguishing proactive remediation, and can be used for binary decision-making, but cannot be used with a threshold of 0.5. It must be used with a custom threshold.

When this detector was applied to the labels for on-task conversation and off-task behavior, it yielded (non-cross-validated) A′ values of 0.66 and 0.59, respectively, indicating that the detector is less successful at distinguishing proactive remediation from those specific behaviors than it is in general, but that it is still better than chance at distinguishing proactive remediation from these other two behaviors that involve pauses in the system.

Table 2 shows the model for proactive remediation. Two of the three features in this model involve actions on which students take longer than other students (possibly indicating that they are not at the computer while those actions wait for input – also seen in off-task detection [cf. 10]. The other feature involves student performance on the items in the clip (students who see an item template multiple times in a row are receiving that item template multiple times because they are answering incorrectly: in particular, current poor student performance is a negative predictor of proactive remediation, perhaps differentiating proactive remediation (which is not based on the activity immediately preceding the action) from on-task conversation (which likely is in many cases).

Table 2. The final proactive remediation detector.

Coefficient	Feature
+0.007	The maximum value across the clip of the difference between the normalized action duration for the given action and the average normalized action duration for the previous two hours for the current student.
-0.003	The maximum number of times an item template in the clip has been seen consecutively (up to and including the clip). Students are often given new variants of the same item template if they are incorrect on their first attempt.
+0.002	The maximum value across the clip of the difference between consecutive changes in the normalized action duration for strings of three consecutive actions with the clip.
+0.007	(constant value)

4. ANALYSIS

The detector of proactive remediation can be used for both intervention and discovery with models analyses [7]. As an example of a potential use of these detectors, we consider the degree to which teacher attendance in Reasoning Mind professional development sessions (which inform teachers on how to effectively use the Reasoning Mind software to decide when to perform proactive remediation and effective methods for carrying it out), is correlated with how often the teacher conducts proactive remediation. To perform this analysis, proactive remediation model raw predictions were computed for all clips in the data set, and then averaged for each teacher. The data contained actions for 462 students across nine teachers. Attendance data were compiled for RM's Best Practices Workshops (BPWs). BPWs are workshops covering various areas of teacher professional development in the context of the Reasoning Mind classroom, including a significant focus on proactive remediation. For each of the nine teachers represented in the data set, BPW attendance for the 2011-2012 and 2012-2013 academic years was tallied and correlated to the proactive remediation average.

Fig. 1 shows this relationship; the two quantities are positively correlated, with $R^2 = 0.50$. The three teachers with the lowest numbers of BPWs attended were first year teachers (and thus could not have attended BPWs in 2011-2012). If those three teachers are excluded, the relationship is stronger, with a steeper slope and $R^2 = 0.73$. This analysis indicates that the occurrence of proactive remediation is indeed increased by participation in BPWs, a goal of that program.

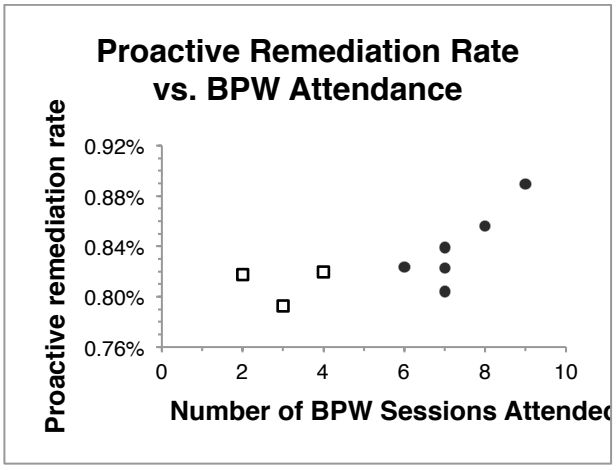

Figure. 1. Average proactive remediation rate (per teacher) vs. number of Best Practices Workshops attended in academic years 2011-2012 and 2012-2013. First-year teachers are indicated by an open square.

5. CONCLUSION

In this paper, we have constructed an automated, sensor-free detector of proactive remediation within the Reasoning Mind mathematics curriculum. This model achieves detector goodness of A′ = 0.90, and also achieves Kappa of 0.65 after post-hoc threshold adjustment. This detector represents another demonstration of the power of interaction log files from online and blended learning systems to support a wide range of inferences about learning. Past work has demonstrated that student behaviors that occur outside of the learning system can be detected [e.g. 10], and that student affect can be inferred [9, 11]. The work presented here indicates that student log files can even be used to distinguish which types of student-teacher interactions are occurring. It is likely that the detector would be even more successful if it incorporated log files from teacher behaviors; teacher data use is not currently instrumented, but could be. It would be interesting to study how much this type of additional instrumentation could contribute to inferring this behavior.

In this paper, we used the proactive remediation detector in a simple discovery with models analysis, which showed that attending Reasoning Mind's teacher professional development is associated with an increase in the occurrence of proactive remediation in teachers' classes, which is one of the goals of the professional development sessions.

Future goals for this work involve bringing these detectors on-line for real-time detection of this behavior; in particular, this detector has the potential to eventually be part of a system of comprehensive, automated detectors of teacher efficacy. As a result, we will be able to create automated interventions for teachers that encourage effective classroom practices, as well as providing this information to implementation coordinators, supporting better teacher practice at lower cost than is currently feasible.

7. ACKNOWLEDGMENTS

We would like to thank George Khachatryan for helpful discussions and suggestions, and Mark Riedl for suggesting the term "proactive remediation". We would also like to thank 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.

8.REFERENCES

1. Baker, R. S., Corbett, A. T., Koedinger, K. R. 2004. Detecting Student Misuse of Intelligent Tutoring Systems. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems,* pp.531-540.
2. Rowe, J., McQuiggan, S., Robison, J., Lester, J. 2009. Off-Task Behavior in Narrative Centered Learning Environments. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence in Education* (AIED-09), pp.99-106.
3. Amershi, S., Conati, C. 2009. Combining Unsupervised and Supervised Machine Learning to Build User Models for Exploratory Learning Environments. *Journal of Educational Data Mining* 1(1), 18-71.
4. Aleven, V., McLaren, B., Roll, I., Koedinger, K. 2006. Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence and Education* (16), 101-128.
5. Arroyo, I., Ferguson, K., Johns, J., Dragon, T., Meheranian, H., Fisher, D., Barto, A., Mahadevan, S., Woolf, B. P. 2007. Repairing Disengagement With Non-Invasive Interventions. In *Proceedings of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, pp.195-202.
6. Roll, I., Aleven, V., McLaren, B. M., Koedinger, K. R. 2011. Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction.* 21(2), 267-280.
7. Hershkovitz, A., Baker, R. S. J. d., Gobert, J., Wixon, M., Sao Pedro, M. 2013. Discovery with Models: A Case Study on Carelessness in Computer-based Science Inquiry. *American Behavioral Scientist.* 57(10), 1479-1498.
8. Hershkovitz, A., Baker, R. S. J. d., Gowda, S. M., Corbett, A. T. 2013. Predicting Future Learning Better Using Quantitative Analysis of Moment-by-Moment Learning. In *Proceedings of the 6th International Conference on Educational Data Mining*, pp.74-81.
9. Pardos, Z. A., Baker, R. S. J. d., San Pedro, M. O. C. Z., Gowda, S. M., Gowda, S. M. 2013. 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*, pp.117-124.
10. Baker, R. S. J. d. 2007. Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems. In *Proceedings of ACM CHI 2007: Computer-Human Interaction*, pp.1059-1068.
11. Baker, R. S. J. d., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J., Rossi, L. 2012. Towards Sensor-free Affect Detection in Cognitive Tutor Algebra. In *Proceedings of the 5th International Conference on Educational Data Mining*, pp.126-133.
12. Khachatryan, G., Romashov, A., Khachatryan, A., Gaudino, S., Khachatryan, J., Guarian, K., Yufa, N. 2014. 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, 24* (3), 333-382.
13. Ocumpaugh, J., Baker, R. S. J. d., Gaudino, S., Labrum, M. J.,Dezendorf, T. 2013. Field Observations of Engagement in Reasoning Mind. In *Proceedings of the 16th International Conference on Artificial Intelligence and Education*, pp.624-627.
14. Ocumpaugh, J., Baker, R. S. J. d., Rodrigo, M. M. T. 2012. *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.
15. Baker, R. S. J. d., D'Mello, S. K., Rodrigo, M. M. T., Graesser, A. C. 2010. Better to Be Frustrated than Bored. T*he. International Journal of Human-Computer Studies.* 68(4), 223-241.
16. Bartel, C. A., Saavedra, R. 2009. The collective construction of work group models. *Administrative Science Quarterly* 1(1), 3-17.
17. Litman, D. J., Forbes-Riley, K. 2006. Recognizing Student Emotions on the Basis of Utterances in Spoken Tutoring Dialogues with both Human and Computer Tutors. *Speech Communication,* 48(5), 559-590.
18. 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., Viehland, N. J. B. 2008. 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,* pp.40-49.
19. Planalp, S., DeFrancisco, V. L., Rutherford, D. 1996.Varieties of cues to emotion in naturally occurring situations. *Cognition and Emotion,* 10(2), 137-153.
20. Corbett, T., A., Anderson, J. 1995. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction,* 4, 253-278.
21. San Pedro, M., Baker, R., Rodrigo, M. 2011. Detecting Carelessness through Contextual Estimation. In *Proceedings of 15th International Conference on Artificial Intelligence in Education,* pp.304-311.
22. Baker, R. S. J. d., Goldstein, A. B., Heffernan, N. T. 2011. Detecting Learning Moment-by-Moment. *International Journal of Artificial Intelligence in Education,* 21(1-2), 5-25.
23. Hanley, J., McNeil, B. 1982. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology,* 143, 29-36.
24. Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20(1), 37-46.
25. Baker, R. S. J. d., Corbett, A. T., Aleven, V. (2008)x . 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,* pp.406-415.