WTF? Detecting Students who are Conducting Inquiry Without Thinking Fastidiously

Michael Wixon, Ryan S.J.d. Baker, Janice Gobert,
Jaclyn Ocumpaugh, & Matthew Bachmann

Worcester Polytechnic Institute, Worcester, Massachusetts
{mwixon,rsbaker, jgobert, jocumpaugh}@wpi.edu,
bachmann.matt@gmail.com

Abstract. In recent years, there has been increased interest and research on identifying the various ways that students can deviate from expected or desired patterns while using educational software. This includes research on gaming the system, player transformation, haphazard inquiry, and failure to use key features of the learning system. Detection of these sorts of behaviors has helped researchers to better understand these behaviors, thus allowing software designers to develop interventions that can remediate them and/or reduce their negative impacts on user outcomes. In this paper, we present a first detector of what we term WTF ("Without Thinking Fastidiously") behavior, based on data from the Phase Change microworld in the Science ASSISTments environment. In WTF behavior, the student is interacting with the software, but their actions appear to have no relationship to the intended learning task. We discuss the detector development process, validate the detectors with human labels of the behavior, and discuss implications for understanding how and why students conduct inquiry without thinking fastidiously while learning in science inquiry microworlds.

Keywords: student modeling, educational data mining, intelligent tutoring system, science inquiry, off-task behavior

1 Introduction

In recent years, there has been increasing awareness that the behavior of students learning from educational software can deviate in several ways from the behaviors expected by software designers. Traditional student modeling paradigms tend to assume that a learner is attempting to perform the designated task as intended, and that incorrect performance pertains solely to not knowing the skill [1-3]. However, other researchers have considered the various ways that student behavior may deviate from expected patterns. For example, students may game the system, attempting to succeed in an educational task by systematically taking advantage of properties and regularities in the system used to complete that task, rather than by thinking through the material [4]. Students also may transform the learning task to a different task entirely [5]. Additionally, students may engage in haphazard inquiry, whereby they get closer to and then further from the goal of the task [6], showing a lack of understanding of how to conduct inquiry. Finally, some students may engage in acts wholly disconnected from the goals of the learning system. For example, in an online learning environment in which students were expected to discover what disease is infecting a community of scientists, students instead spent their time in unrelated behaviors, such as placing bananas in the toilet [personal communication, Jennifer Sabourin]. In another example, students plotting points from a function in a Cognitive Tutor for high school mathematics may instead plot a smiley face.

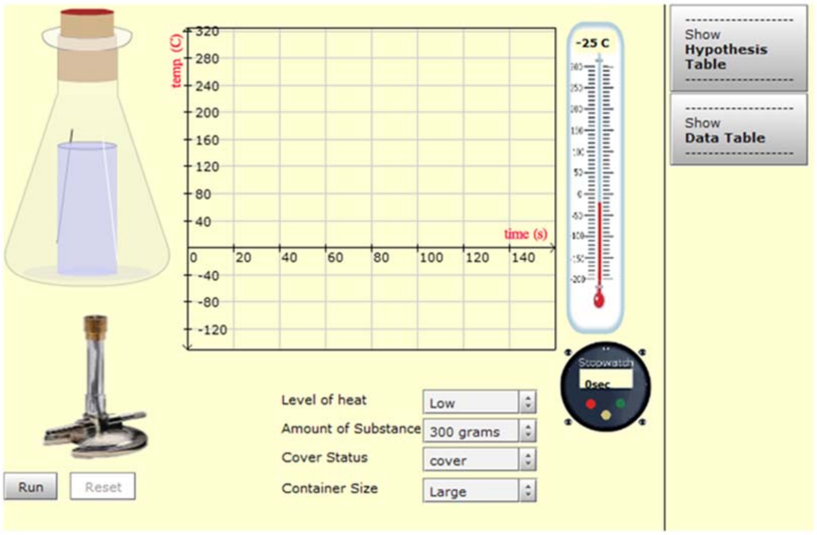
Rowe and his colleagues conceptualize this type of behavior as off-task [7], which they define as "behaviors that are clearly unrelated to the narrative and curriculum." We believe that there are important differences between this behavior and the type of behaviors typically considered to be off-task, whether within educational software [4] or non-computerized learning settings [8]. Whereas off-task behavior in previous accounts is seen as being completely disconnected from the learning task and environment, this "bananas in the toilet" behavior is disconnected from the learning task but occurs within the learning environment. Hence, we propose that this behavior be referred to instead as "WTF behavior." (WTF, of course, stands for "Without Thinking Fastidiously.") WTF behaviors may have negative impacts on learning, as off-task behavior does. However, to the extent that WTF behavior differs from off-task behavior, it may manifest differently in log files, necessitating detectors tailored to this behavior.

Within this paper, we present the first automated detector of WTF behavior, developed in the context of a science inquiry microworld in the domain of Phase Change, within the Science ASSISTments learning software [www.scienceassistments.org; 9-10]. This detector is generated using a combination of feature engineering and step regression, and is cross-validated at the student level (e.g. repeatedly trained on one group of students and tested on other students). We report this detector's effectiveness at identifying WTF behavior, analyze its internal features, and compare it to past detectors of other forms of disengagement.

2 Data Set

The data analyzed in this study were produced by 144 eighth graders (generally ages 12-14), who were using the Science ASSISTments' Phase Change microworld, within their science classes. All attended a middle school with a diverse population in a medium-sized city in central Massachusetts. The student population exhibits substantial economic and educational challenges: 20% of them qualified for free or reduced-price school lunches in the 2009-2010 school year and greater than 50% scored at or below "needs improvement" in the Science & Technology/Engineering portion of the Massachusetts Comprehensive Assessment System (MCAS).

Within the Phase Change microworld, shown in Figure 1, students observe and manipulate a simulation to conduct inquiry regarding the changes between solid, liquid, and gas. Specifically, students form hypotheses regarding the phenomenon, and test their hypotheses by running experiments within the simulation. They then interpret their data, warrant their claims, and communicate findings. At any point during their analysis, they may return to the experiment or hypothesizing phases.



Level of heat: Low
Amount of Substance: 300 grams
Cover Status: cover
Container Size: Large

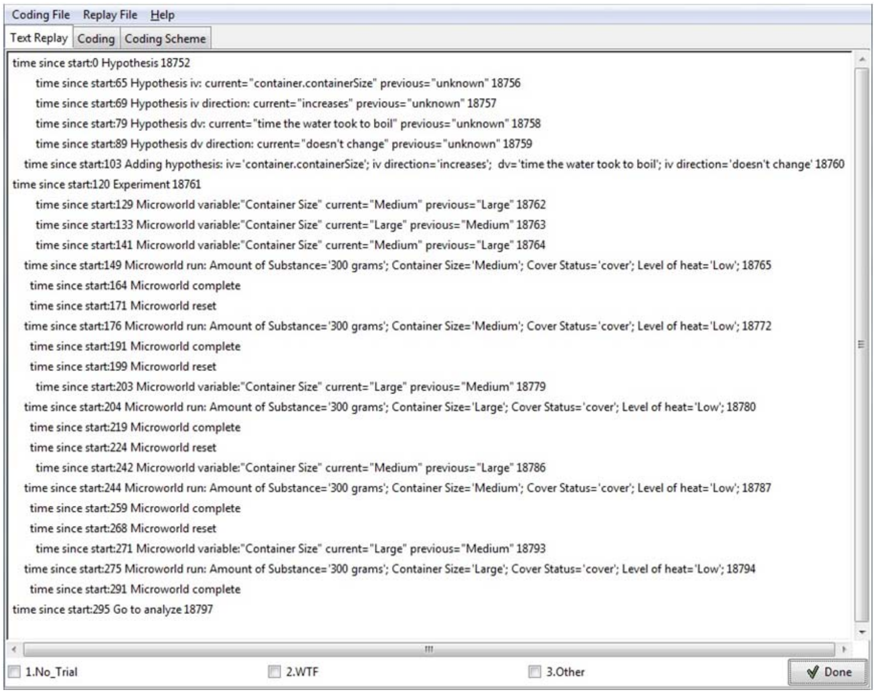
Fig. 1. A screen shot of the Phase Change microworld

Each of the 144 students completed at least one data collection activity in the phase change environment. In this paper, we focus on student actions in the hypothesizing and experimentation phases of the microworld. As students solved these tasks, their actions within the software were logged, including generating hypotheses, designing and running experiments, and switching between hypothesizing, experimentation, and other inquiry activities – for a total of 144,841 actions. Logs included the action type, the relevant simulation variable values, and the time stamp.

3 WTF Detector

3.1 Obtaining Ground Truth Labels of WTF Behavior Using Text Replays

The first step in our process of developing a data-mined detector of WTF behavior, is to develop ground truth labels, using text replays [9, 11]. In text replays, human coders are presented “pretty-printed” versions of log files (as shown in Figure 2). WTF behavior may be difficult to rationally define in log files (and rational detectors of this nature are difficult to validate for generalizability), but behavior that is completely disconnected from the learning task can be identified by humans relatively easily. In past cases, text replays have proved effective for providing ground truth labels for behaviors of this nature [9, 12-13]. Examples of WTF behavior in this data set include running the exact same experiment a large number of times (shown in Figure 2), toggling variable settings back and forth repeatedly, and changing large numbers of variables repeatedly. As can be seen, WTF behavior manifests in several ways, an interesting challenge for developing an automated detector of this construct.



```
time since start:0 Hypothesis 18752
    time since start:65 Hypothesis iv: current="container.containerSize" previous="unknown" 18756
    time since start:69 Hypothesis iv direction: current="increases" previous="unknown" 18757
    time since start:79 Hypothesis dv: current="time the water took to boil" previous="unknown" 18758
    time since start:89 Hypothesis dv direction: current="doesn't change" previous="unknown" 18759
  time since start:103 Adding hypothesis: iv='container.containerSize'; iv direction='increases';  dv='time the water took to boil'; iv direction='doesn't change' 18760
time since start:120 Experiment 18761
    time since start:129 Microworld variable:"Container Size" current="Medium" previous="Large" 18762
    time since start:133 Microworld variable:"Container Size" current="Large" previous="Medium" 18763
    time since start:141 Microworld variable:"Container Size" current="Medium" previous="Large" 18764
  time since start:149 Microworld run: Amount of Substance='300 grams'; Container Size='Medium'; Cover Status='cover'; Level of heat='Low'; 18765
   time since start:164 Microworld complete
   time since start:171 Microworld reset
  time since start:176 Microworld run: Amount of Substance='300 grams'; Container Size='Medium'; Cover Status='cover'; Level of heat='Low'; 18772
   time since start:191 Microworld complete
   time since start:199 Microworld reset
    time since start:203 Microworld variable:"Container Size" current="Large" previous="Medium" 18779
  time since start:204 Microworld run: Amount of Substance='300 grams'; Container Size='Large'; Cover Status='cover'; Level of heat='Low'; 18780
   time since start:219 Microworld complete
   time since start:224 Microworld reset
    time since start:242 Microworld variable:"Container Size" current="Medium" previous="Large" 18786
  time since start:244 Microworld run: Amount of Substance='300 grams'; Container Size='Medium'; Cover Status='cover'; Level of heat='Low'; 18787
   time since start:259 Microworld complete
   time since start:268 Microworld reset
    time since start:271 Microworld variable:"Container Size" current="Large" previous="Medium" 18793
  time since start:275 Microworld run: Amount of Substance='300 grams'; Container Size='Large'; Cover Status='cover'; Level of heat='Low'; 18794
   time since start:291 Microworld complete
time since start:295 Go to analyze 18797
```

Fig. 2. Text Replay Showing Student Running The Same Trial a Large Number of Times

In order to create text replays, the student data was segmented into “clips”, sequences of student behavior. In this paper, we segment student data by sequences of student data collection behavior (experimentation within the microworld), adopting the approach for doing so proposed in [9]. In this approach, a clip begins when a student enters the data collection phase and ends when the student leaves that phase. The typical order of student actions in Science ASSISTments is to create hypotheses, collect data, interpret data, warrant claims, and then communicate their findings, but a student can return to data collection after interpreting data. Thus, a clip may start either after the student makes a hypothesis and decides to collect data, or after the student attempts to interpret data and decides to collect more data.

Clips were coded individually, but not in isolation. That is, coders had access to all of the previous clips the same student produced within the same activity so that they could detect WTF behavior that might have otherwise been missed due to lack of context. For example, a student may repeatedly switch between hypothesizing and experimentation, running the exact same experiment each time. Although repeating the same experiment two or three times may help the student understand the simulation better, doing so more than twenty times might be difficult to explain except as WTF.

Two human coders (the 2nd and 5th authors) practiced coding WTF on two sets of clips which were excluded from use in detector development. In the first set of clips, they coded together and discussed coding standards. Next, the two coders separately each coded a second set of 200 clips independently. The two coders achieved acceptable agreement, with Cohen's [14] Kappa of 0.66.

Afterwards, the 2nd author coded 571 clips, which were used to develop the WTF detector. Since several clips could be generated per activity, a single, randomly chosen clip was tagged per student, per activity (however, not all students completed all activities, causing some student-activity pairs to be missing from the data set). This ensured all students and activities were approximately equally represented in this data set. Seventy of these clips were excluded from analysis, due to a lack of data collection actions on the student's part. Of the 501 clips remaining, 15 (3.0%) were labeled as involving WTF behavior, a proportion similar to the proportions of disengaged behavior studied in past detector development [cf. 12]. These 15 clips were drawn from 15 (10.4%) of the students (i.e., no student was coded as engaging in WTF behavior more than once).

3.2 Data Features

In order to develop an automated detector of WTF behavior from the log files, we distilled features of the data corresponding to the clips of behavior labeled by the coders. An initial set of 77 features was distilled using code that had been previously developed to detect student use of experimentation strategies and testing the correct hypothesis within Science ASSISTments [9]. As many of these features did not appear relevant to detecting WTF behavior and a greater number of features increases the risk of over-fitting [16], this set was manually reduced to 24 features without reference to the labeled data.

All of these 24 features corresponded to information about the set of actions involved in a specific clip and prior actions that provided context for the clip. The first four features involve overall statistics for the clip: (1) the total number of actions, (2) the average time between actions, (3) the maximum time between actions, and (4) the total number of experimental trials run by the student. The next three features were based on pauses: (5) the number of times a student paused the simulation during runs, (6) the average duration of student-initiated pauses of the simulation (i.e., total time spent paused, divided by number of pauses), and (7) the duration of the longest single instance when the student paused the system.

Ten more features relevant to the time elapsed during experimentation were used: (8) the total amount of time spent before running each experimental trial but after performing the previous action, (9) the average time spent by the student before running each experimental trial but after performing the previous action, (10) the standard deviation of the time spent by the student before running each experimental trial but after performing the previous action, and (11) the maximum time spent before running each experimental trial but after performing the previous action.

Several features related to resetting or pausing the experimental apparatus (or the absence of this action), were included. Pausing the simulation can be appropriate in many situations, but doing so large numbers of times may be an indicator of WTF behavior. These include: (12) the number of experimental trials run without either pauses or resets, (13) the average time spent by the student before running each experimental trial which was completed without being reset but after performing the previous action, (14) the number of trials where the system was reset, (15) the average time spent before running each experimental trial that were reset but after performing the previous action, and (16) the maximum time spent before running an experimental trial that was reset before completion but after performing the previous action.

The next set of features involved whether and how a student changed the variables while forming hypotheses. These included (17) the number of times a variable was changed, and three measures of the period of time that elapsed before the student changed a variable (measured from the previous action, whatever it was): (18) the sum total of time elapsed in all these periods, (19) the mean time elapsed across these periods, and (20) the standard deviation of time elapsed across these periods.

The final features consisted of changes to independent variables between experimental trials. These included: (21) the number of times an independent variable was changed during the experiment phase, and three measures of the period of time that elapsed before the student changed a variable (measured from the previous action, whatever it was), namely: (22) the sum total of time elapsed in all these periods, (23) the mean time elapsed across these periods, and (24) the standard deviation of time elapsed across these periods. These features regarding variable changes were useful as extremely large numbers of changes would not map to any reasonable experimentation strategy.

3.3 Detector Development

We attempted to fit detectors of WTF using 11 common classification algorithms, including Naïve Bayes, and J48 decision trees. The best model performance was achieved by the PART algorithm [17], an algorithm that produces rules out of C4.5 decision trees (essentially the same algorithm as J48 decision trees). The implementation of PART from WEKA [18] was run within RapidMiner 4.6 [19]. In this algorithm, a set of rules is built by repeatedly building a decision tree and making a rule out of the path leading to the best leaf node at each iteration. PART has not been frequently used in student modeling, but was used in [20] to predict student course success. These models were evaluated using a process of six-fold student-level cross-validation [21]. In this process, students are split randomly into six groups. Then, for each possible combination, a detector is developed using data from five groups of students before being tested on the sixth "held out" group of students. By cross-validating at this level, we increase confidence that detectors will be accurate for new groups of students.

Detectors were assessed using four metrics, A' [22], Kappa [14], precision [23], and recall [23]. A' is the probability that the detector will be able to distinguish a clip involving WTF behavior from a clip that does not involve WTF behavior. A' is equivalent to both the area under the ROC curve in signal detection theory and to W, the Wilcoxon statistic [22]. A model with an A' of 0.5 performs at chance, and a model with an A' of 1.0 performs perfectly. An appropriate statistical test for A' in data across students would be to calculate A' and standard error for each student for each model, compare using Z tests, and then aggregate across students using Stouffer's method. However, the standard error formula for A' [22] requires multiple examples from each category for each student, which is infeasible in the small samples obtained for each student in our data labeling procedure. Another possible method, ignoring student-level differences to increase example counts, biases undesirably in favor of statistical significance. Hence, statistical tests for A' are not presented in this paper.

The second feature used to evaluate each detector was Cohen's Kappa, which assesses whether the detector is better than chance at identifying which clips involve WTF behavior. A Kappa of 0 indicates that the detector performs at chance, and a Kappa of 1 indicates that the detector performs perfectly. Detectors were also evaluated using precision and recall, which indicate (respectively) how good the model is at avoiding false positives, and how good the model is at avoiding false negatives.

A' and Kappa were chosen because they compensate for successful classifications occurring by chance [24], an important consideration in data sets with unbalanced proportions of categories (such as this case, where WTF is observed 3.0% of the time). Precision and recall give an indication of the detector's balance between two forms of error. It is worth noting that unlike Kappa, precision, and recall (which only look at the final label), A' takes detector confidence into account.

4 Results

The detector of WTF behavior developed using the PART algorithm achieved good performance under 6-fold student-level cross-validation. As shown in Table 1, the detector achieved a very high A' of 0.979, signifying that it could distinguish whether or not a clip involved WTF behavior approximately 97.9% of the time. When uncertainty was not taken into account, performance was lower, though still generally acceptable. The detector achieved a Kappa value of 0.4, indicating that the detector was 40% better than chance. This level of Kappa is comparable to past detectors of other constructs effectively used in interventions [9, 12]. Kappa values in this range, combined with almost perfect A' values, suggest that the detector is generally good at recognizing which behavior is more likely to be “WTF”, but classifies some edge cases incorrectly. In general, the detector’s precision and recall (which, like Kappa, do not take certainty into account), were approximately balanced, with precision = 38.9% and recall = 46.7%. Thus, it is important to use fail-soft interventions and to take detector certainty into account when selecting interventions – but there is not evidence that the detector has strong bias either in favor of or against detecting WTF behavior.

Table 1. WTF Detector Confusion Matrix

	Clips Coded as WTF by Humans	Clips Coded as NOT WTF by Humans
Detector Predicted WTF	7	11 (false positives)
Detector Predicted NOT WTF	8 (false negatives)	475

The algorithm, when fit on the entire data set, generated the following final model. In running this model, the rules are run in order from the first rule to the last rule.

1) IF the total number of independent variable changes (feature 21) is seven or lower, AND the number of experimental trials run (feature 7) is three or lower, THEN **NOT WTF**.
2) IF the maximum time spent between an incomplete run and the action preceding it (feature 16) is 10 seconds or less, AND the total number of independent variable changes (feature 21) is eleven or less, AND the average time spent paused (feature 5) is 6 seconds or less, THEN **NOT WTF**.
3) IF the total number of independent variable changes (feature 21) is greater than one, AND the maximum time between actions (feature 3) is 441 seconds or less, AND the number of trials run without pauses or resets (feature 12) is 4 or less, THEN **NOT WTF**.

4) IF the total number of independent variable changes (feature 21) is 12 or less, THEN **WTF**.
5) IF the maximum time spent before running each experimental trial but after performing the previous action (feature 11) is greater than 1.8 seconds, THEN **NOT WTF**.
6) All remaining instances are classified as **WTF**.

As can be seen, this detector used 6 rules to distinguish WTF behavior, which employ 8 features from the data set. Four of the rules identify the characteristics of behavior that is NOT WTF, while only two identify the characteristics of WTF behavior. We discuss the implications of the specific rules in the following section.

5 Discussion and Conclusions

In this paper, we introduce a first automated detector that can identify when a student is completely disconnected from the learning task but is still actively using the learning environment. This behavior, which we term WTF behavior ("without thinking fastidiously"), has been reported in multiple online learning environments, but has not yet been modeled or studied to the degree that it merits. Our findings suggest that WTF behavior has prevalence similar to gaming the system, a behavior known to be associated with poor learning [4], and that it can be identified both by human coders and by an automated detector. This opens the possibility of studying how WTF behavior correlates with learning, identifying what factors lead students to engage in WTF behavior, and in turn, developing automated interventions designed to bring students back on track. Work along these lines is currently ongoing in our lab.

Examining the model of WTF behavior obtained provides some interesting implications about this type of behavior. Previous detectors of undesirable behavior have largely focused on identifying the specific undesirable behavior studied [cf. 12, 13, 25]. By contrast, the rules produced by the WTF detector are targeted more towards identifying what *is not* WTF behavior than identifying what *is* WTF behavior. Four of the six rules identify non-WTF behavior. Of the two rules identifying WTF behavior, one simply states that any behavior not captured by the first five rules can be considered WTF. As such, this model suggests that WTF behavior may be characterized by the absence of appropriate strategies and behaviors, in a student actively using the software, rather than specific undesirable behavior.

It is also worth discussing the data feature which is most frequently employed in the model rules: the number of times the student changes a simulation variable (feature 21). Though this feature is used in four of the six rules, there is not a clear pattern where frequently changing variables is simply either good or bad. Instead, different student actions appear to indicate WTF behavior in a student who frequently changes simulation variables, compared to a student who seldom changes simulation variables. Specifically, a student who changes variables many times without stopping to think before running the simulation is seen as displaying WTF behavior. By contrast, a student who changes variables fewer times is categorized as displaying WTF behavior if he or she runs a large number of experimental trials and also pauses the simulation for long periods of time. This may indicate that the student is running the simulation far more times than is warranted for the number of variables being changed, and that his or her pattern of pauses does not seem to indicate that he or she is using the time to study the simulation.

As mentioned earlier, one potential direction for future work is to study the individual differences and situational factors leading students to engage in WTF behavior. This behavior could be expected to emerge for several reasons, including attitudinal reasons such as not valuing the learning task, a goal orientation of work avoidance, or immediate affective states such as confusion, frustration, and boredom. A key first paper investigating this question is Sabourin et al. [26], which showed that when WTF behavior (termed off-task behavior) emerges among students displaying different affect, it has different implications about their affect later in the task. Students who engage in this behavior when they are confused later become bored or frustrated. By contrast, students who engage in this behavior when they are frustrated often become re-engaged. These findings suggest that intelligent tutors should offer different interventions, depending on the affective context of WTF behavior, but further research is needed to determine which strategies are most appropriate and effective for specific learning situations and for learners with specific characteristics. For example, a confused student engaging in WTF behavior may need additional support in understanding how to learn from the learning environment [27]. By contrast, a student who engages in WTF behavior due to boredom or because they do not value the learning task may require intervention targeted towards demonstrating the long-term value of the task for the student's goals [cf. 28].

Automated detectors such as the one presented here have a substantial role to play in understanding the causes of WTF behavior. In specific, these detectors will make it feasible to study WTF behavior across a greater number of situations [cf. 15], helping us to better understand the factors leading to WTF behavior. By understanding the causes of WTF behavior, and how learning software should respond to it, we can take another step towards developing learning software that can effectively adapt to the full range of students' interaction choices during learning.

Acknowledgements. This research was supported by grant "Empirical Research: Emerging Research: Using Automated Detectors to Examine the Relationships Between Learner Attributes and Behaviors During Inquiry in Science Microworlds", National Science Foundation award #DRL-100864. We also thank Michael Sao Pedro for assistance in data distillation and for helpful comments and suggestions, and we thank Arnon Hershkovitz and Adam Nakama for helpful comments and suggestions.

6 References

1. Corbett, A.T., Anderson, J.R.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction, 4, 253-278. (1995)
2. Martin, J., VanLehn, K.: Student assessment using Bayesian nets. International Journal of Human-Computer Studies, 42, 575-591. (1995)
3. Pavlik, P.I., Cen, H., Koedinger, K.R.: Performance Factors Analysis – A New Alternative to Knowledge Tracing. Proc. of the 14th International Conference on Artificial Intelligence in Education, 531-540. (2009)
4. Baker, R.S., 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)
5. Magnussen, R., Misfeldt, M.: Player transformation of educational multiplayer games. Proceedings of Other Players. Copenhagen, Denmark. (2004)
6. Buckley, B., Gobert, J., Horwitz, P. & O'Dwyer, L.: Looking inside the black box: Assessing model-based learning and inquiry in Biologica, 5, 2, 166-190. International Journal of Learning Technologies. (2010).
7. Rowe. J., McQuiggan, S., Robison, J., Lester, J.: Off-Task Behavior in Narrative-Centered Learning Environments. Proceedings of the 14th International Conference on AI in Education, 99-106. (2009)
8. Karweit, N., Slavin, R.E.: Measurement and Modeling Choices in Studies of Time and Learning. American Educational Research Journal, 18, 157-171 (1981)
9. 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. To appear in User Modeling and User-Adapted Interaction: The Journal of Personalization Research. (in press)
10. Gobert, J., Sao Pedro, M., Raziuddin, J.: Studying the Interaction Between Learner Characteristics and Inquiry Skills in Microworlds. Proceedings of the 9th International Conference on the Learning Sciences, 46-47. (2010)
11. Baker, R.S.J.d., Corbett, A.T., Wagner, A.Z.: Human Classification of Low-Fidelity Replays of Student Actions. Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems, 29-36. (2006)
12. Baker, R.S.J.d., de Carvalho, A. M. J. A.: Labeling Student Behavior Faster and More Precisely with Text Replays. Proceedings of the 1st International Conference on Educational Data Mining, 38-47. (2008)
13. Baker, R.S.J.d., Mitrovic, A., Mathews, M.: Detecting Gaming the System in Constraint-Based Tutors. Proceedings of the 18th Annual Conference on User Modeling, Adaptation, and Personalization, 267-278. (2010)
14. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46. (1960)

15. Baker, R.S.J.d., de Carvalho, A.M.J.A., Raspat, J., Aleven, V., Corbett, A.T., Koedinger, K.R.: (2009) Educational Software Features that Encourage and Discourage "Gaming the System". Proceedings of the 14th International Conference on Artificial Intelligence in Education, 475-482.
16. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (1997)
17. Frank, E., Witten, I. H.: Generating Accurate Rule Sets Without Global Optimization. Proceedings of the Fifteenth International Conference on Machine Learning, 144–151. (1998).
18. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kauffmann, San Francisco (1999)
19. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T.: YALE: Rapid Prototyping for Complex Data Mining Tasks. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 935-940. (2006)
20. Esposito, F., Licchelli, O., Semeraro, G.: Discovering Student Models in e-learning Systems. J. Universal Computer Science, 10(1), 47-57. (2004)
21. Efron, B. & Gong, G.: A leisurely look at the bootstrap, the jackknife, and cross-validation. American Statistician, 37, 36–48. (1983)
22. Hanley, J., McNeil, B.: The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology, 143, 29-36. (1982)
23. Davis, J., Goadrich, M.: The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning, 233-240. (2006)
24. Ben-David, A.: About the Relationship between ROC Curves and Cohen’s Kappa. Engineering Applications of Artificial Intelligence, 21, 874-882. (2008)
25. Cetintas, S., Si, L., Xin, Y.P., Hord, C.: Automatic Detection of Off-Task Behaviors in Intelligent Tutoring Systems with Machine Learning Techniques. IEEE Transactions on Learning Technologies, 3 (3), 228-236. (2009)
26. 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 of the 15th International Conference on Artificial Intelligence in Education, 534-536. (2011)
27. Roll, I., Aleven, V., McLaren, B.M., Koedinger, K.R.: Can help seeking be tutored? Searching for the secret sauce of metacognitive tutoring. Proceedings of the 13th International Conference on Artificial Intelligence in Education, 203-210. (2007)
28. Pekrun, R.: The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. Educational Psychology Review, 18 (4), 315-341. (2006)