CHAPTER 14 –Assessing the Disengaged Behaviors of Learners

Ryan S.J.d. Baker[1], **Lisa M. Rossi**[2]
[1]Columbia University Teachers College, [2]Worcester Polytechnic Institute

Introduction

In recent years, an increasing number of models have been published that can infer if a learner is behaviorally disengaged while working within an interactive learning environment, and can conduct inference using features of data focused on learner interaction with the learning system. In this chapter, we discuss some of the behaviors that have been shown to be amenable to this type of modeling, including off-task behavior, gaming the system, and carelessness. We also consider some of the algorithms and approaches that have been found to be particularly effective. We contemplate the relative merits of knowledge engineering and data-mining approaches for this type of model, and focus on the key validity concerns that must be addressed for these types of models to be used with confidence in a comprehensive framework such as GIFT.

Gaps

Over the last decades, adaptive computerized instruction has become increasingly effective at assessing the knowledge state of a learner (Corbett & Anderson, 1995; Martin & VanLehn, 1995; Shute, 1995; Pavlik et al., 2009; Pardos et al., 2011), supporting automated decisions about which content to assign to students through the implementation of strategies such as mastery learning, where a student is assigned content for a specific skill or knowledge component until demonstrating mastery (Corbett, 2001).

However, despite recent advances in the assessment of student disengagement (discussed in this chapter), and a small number of successful cases of models of disengaged behavior being used in learning interventions (Baker et al., 2006; Walonoski & Heffernan, 2006; Arroyo et al., 2007), adaptive computerized instruction is generally not as adaptive to engagement as it is to student knowledge.

There are likely multiple reasons for this. First, engagement models thus far have had to be created for specific learning environments, with only moderate similarity for models of the same construct created for different learning environment. By contrast, creating knowledge models often involves applying one of a small set of known algorithms to a data set in a standard fashion (Pardos et al., 2011). This often requires some knowledge engineering and rational modeling to create mappings between items and cognitive skills, but that process is relatively well known and has been conducted for a large range of learning environments. Second, validating models of engagement is more challenging than validating knowledge models; knowledge models are often validated in terms of whether predictions of future student behavior are correct (Corbett & Anderson, 1995; Pavlik et al., 2009; Pardos et al., 2011), but validating an engagement model beyond face validity requires collecting human labels of data in terms of the target construct, typically involving external coders (Baker et al., 2004; Baker, 2007b; Cetintas et al., 2010; Walonoski & Heffernan, 2006; Wixon et al., 2012). These labels often must be collected at considerable scale to ensure generalizability to a large and diverse target population.

Emerging Concept, Model, or Method

One of the challenges with modeling engagement within the context of adaptive computerized instruction is deciding which dimension(s) of engagement to model. Fredricks, Blumenfeld, and Paris (2004) have proposed that engagement be studied as a multifaceted construct, with behavioral, affective, and cognitive dimensions. These dimensions can be understood as follows: behavioral engagement centers on the action of participation in an educational interaction (including academic, social, and extracurricular activities), affective engagement focuses on both positive and negative emotional reactions (with regard to teachers, peers, or academics), and cognitive engagement is based on investment at a cognitive level and thoughtfulness. Within these dimensions, there are many constructs, e.g., many behaviors that indicate engagement or disengagement. Each of these constructs can also be defined in a range of ways. This multifaceted view of engagement imposes added complexity to our ability to infer disengagement, as it broadens the breadth of behaviors necessary to detect in order to achieve a full multidimensional picture of a learner's engagement. Fortunately, however, not all dimensions (or aspects of each dimension) need to be detected in order to support effective intervention. In addition, specific behaviors impact learning outcomes and longer-term engagement in different ways, and some are more important to identify and adapt to than others, depending on the learning context.

Despite these complexities, opportunities exist for a combination of rich logs of student interaction and EDM methods to be used in concert to create richer detectors of student disengagement with regard to all three facets of the construct.

In this chapter, we look at work that models the behavioral dimensions of engagement, focusing on work to identify behavioral disengagement in the types of adaptive computerized instruction being integrated with the GIFT framework (such as intelligent tutors, simulations, and serious games). We discuss the following behaviors: off-task behavior, gaming the system, carelessness, and WTF behavior (Wixon, Baker, Gobert, Ocumpaugh & Bachmann, 2012), which have been detected with validated models operating on data from learner interactions with adaptive computerized instruction (e.g., no physical sensors).

Modeling Off-Task Behavior

The first automated detector of whether a student is off-task within adaptive computerized instruction was published in Baker (2007b), which presented a machine-learned model to detect off-task behavior of students using an ITS for middle-school mathematics. Off-task behaviors were defined as behaviors that do not involve the system or learning task (including off-task conversation, off-task solitary behavior, and inactivity), building off past work studying off-task behavior within traditional classroom settings (Karweit & Slavin, 1982; Lahaderne, 1968; Lee, Kelly, and Nyre, 1999). These models were built using data from quantitative field observations conducted in middle-school classrooms by trained field coders, as training data. Models were validated by conducting cross-validation at the student level. Data features included in the model were the following:

- Details of student actions, such as whether the action was correct and the type of user interaction used in the current problem step (e.g., choosing from a set of options, inputting an answer, or plotting points; some types of interactions naturally take longer than others).

- Whether it was the student's first attempt at the problem.

- Time taken to complete a problem, expressed in three ways: (1) how many seconds the action took, (2) how many standard deviations faster or slower the action took compared to other students, and (3) time taken in the last three or five actions expressed as the sum of standard deviations faster or slower than other students.

In order to model off-task behavior, Latent Response Models (LRM) were used as the statistical basis for all detectors in this study, as they are able to easily and naturally integrate multiple data sources at different grain sizes. This framework included one observable level (assessing how frequently each student is off-task) and two hidden (latent) levels. The detector determines the proportion of time spent off-task by making a binary assessment as to whether each individual student action is off-task and determining the percentage of actions which are assessed to be off-task for each student. Model selection for the multiple-feature models were validated by adding parameters to the model until the parameter worsened the model's performance in a student-level tenfold cross-validation. It was found that the best-fitting multiple-parameter model fit the data with a cross-validated correlation of 0.55.

A second automated detector of whether a student is off-task was developed by Cetintas and colleagues (2010), who added data on mouse movement to the features used in Baker (2007b), modeling student off-task behavior within a math tutoring program designed to help elementary school students with learning disabilities and/or emotional disorders learn problem solving skills for Equal Group and Multiplicative Compare problems. This more multimodal approach extended past work that had not considered mouse movement data. Their approach also includes personalization of the detector to account for inter-user variability of behavior. The personalized version of the model incorporates data from a student's past trials on each problem, which are used to generate a student-specific version of each feature while predicting that student's behaviors for the current trial. The same general approach to coding students as on-task or off-task as was used in Baker (2007b) was used in this work. Quantitative field observations were conducted in the classrooms by trained field coders and synchronized with the log data. Students were observed sequentially (to avoid observer bias) and were coded as off-task if they were observed doing any of the following for more than 30 seconds: talking with another student about anything other than the subject material, being inactive, or exhibiting off-task solitary behavior.

Within this work, ridge regression (Hoerl & Kennard, 1970) was used to estimate model parameters, and it was found that this approach (including mouse movement data and student-specific models) led to better detection of off-task behavior than approaches lacking one or more of these features.

Modeling Gaming the System

Gaming the system is defined as attempting to succeed in an education environment by exploiting properties of the system rather than by learning the material and trying to use that knowledge to answer correctly (Baker et al., 2006). The first automated detectors of gaming the system were presented by Aleven et al. (2004) and Baker, Corbett, and Koedinger (2004).

One category of gaming detectors was developed using knowledge engineering (Beal, Qu & Lee, 2006; Buckley, Gobert & Horwitz, 2006; Shih, Koedinger & Scheines, 2008), where rational analysis is applied by a human analyst to derive a model that can be applied. Within knowledge engineering approaches, there is typically no gold standard used to validate models; models and their parameters are generated based on the judgment of the researcher(s), although in some cases research models are compared to learning outcomes. The first such detector was presented in Aleven et al. (2004) and refined within Aleven et al. (2006). This detector was developed using a knowledge engineering approach and took the form of a set of simple production rules. The detector was initially developed in the context of data from a Cognitive Tutor on geometry, but has since been applied to other intelligent tutors. Within this model, gaming behaviors were defined by a pair of rules, *Clicking Through Hints,* which consists of requesting a hint, and then requesting another hint too rapidly to read the first hint (defined as under 5 seconds), and *Try-Step Abuse,* where a student response took under 7 seconds. In related work, Beck (2005) looked at quick responses on difficult items, parameterizing "quick" and "difficult" and fitting values for these parameters based on data. Similar knowledge engineered models have been presented by Beal et al. (2006), Johns & Woolf (2006), Gong, Beck, and Heffernan (2010), and Muldner et al. (2011).

A second category of gaming detectors was developed using a combination of human labels of gaming behavior and data mining/machine learning methods, where a model is trained to infer what the human coder's labels are. The first such detector was presented in Baker, Corbett, and Koedinger (2004) and refined within Baker et al. (2008). This detector identified gaming the system and distinguished between gaming behaviors characterizing unsuccessful students and gaming behaviors characterizing more successful students (the primary distinction was in terms of when gaming occurs, with unsuccessful students gaming on difficult material and successful students gaming on easier material). This detector was built for students using a Cognitive Tutor for middle school, using quantitative field observations and tutor log data collected across a total of four tutor lessons used by over 400 students in two separate school districts. Data features used in the model included time expressed in terms of how many seconds the action took, how many standard deviations faster or slower the action took compared to other students, and time taken in the last three or five actions expressed as the sum of standard deviations faster or slower than other students. Details about the interaction were also included regarding the learning system's assessment of the action (as correct; incorrect, indicating a known bug; or a help request), the type of interface widget involved in the action, and whether this attempt was the student's first attempt. The automated detectors, developed using a LRM framework (Maris, 1995) integrated the field observations and tutor logs, at different grain sizes, into a single model. In a gaming detector's outermost/observable layer, the gaming detector assessed how frequently each of *n* students is gaming the system. The gaming detector's assessments for each student were then compared to the observed proportions of time each student spent gaming the system. The detector was validated to be able to generalize to new students and new tutor lessons.

In later work to detect gaming using machine-learning and human labels, Walonoski and Heffernan (2006) improved on the very coarse-grained label synchronization in Baker's work by using time windows of five different sizes (30 seconds, and 1, 2, 4, and 6 min), increasing the degree to which the model was fine-grained. Next, work by Baker and de Carvalho (2008) achieved 20-second-level synchronization by using text replays rather than quantitative field observations, for human labeling. Text replays are log files presented in textual form, and represent a segment of student behavior during a pre-selected duration of time or length. It has been shown that text replays have good inter-rater reliability and agree well with prediction made by models generated using quantitative field observation data (Baker, Corbett & Wagner, 2006). Using labeling at this label made it possible to use off the shelf classifiers, in this case J48 Decision Trees, an open-source implementation of C4.5 Decision Trees (Quinlan, 1993). In further work, this approach was extended to a constraint-based tutor for database programming (Baker, Mitrovic & Mathews, 2010) and a handheld app was developed to support synchronization of field observations with similar precision to what can be achieved with text replays (Ocumpaugh et al., 2012).

Modeling Carelessness

The construct of carelessness has been defined in two ways: as an error made on a task that the student already knows (Clements, 1982), or as impulsive and/or hurried actions (Maydeu-Olivares & D'Zurilla, 1995). San Pedro and colleagues (2011a) identified Clements's definition of carelessness as being the same as the contextual probability of slipping on a problem or problem step in an intelligent tutor, a construct that it was previously shown can be inferred through manipulating a BKT model (Baker, Corbett & Aleven, 2008). Based on this theoretical link, San Pedro and colleagues (2011a) manipulated the internal structure of a BKT algorithm in order to develop a model of carelessness, doing so using log files produced within a Cognitive Tutor for Scatterplots. The model of carelessness was developed by first obtaining ground-truth labels using future knowledge to drive a machine-learned model that can predict careless errors without using future data. Then, a model that only uses data from the past was created using sixfold student-level cross-validation linear regression modeling. Creating this model also serves a function of smoothing extreme estimates. The model achieved A' and BIC' values, which indicated that the detector performed better than chance. The same approach was replicated within Science ASSISTments, a set of scientific simulations that scaffold student inquiry processes and assess students' inquiry skill (Hershkovitz et al., 2013).

This model has been validated to work for new students (Baker, Corbett & Aleven, 2008), new populations in different countries (San Pedro et al., 2011a), and additional tutoring systems (Hershkovitz et al., 2011). It has been also been shown to predict student post-test score even when controlling for student knowledge (Baker et al., 2010).

Modeling WTF Behavior/Off-Task Behavior within environment

Rowe and colleagues (2009) reported that students sometimes engage in behavior within online learning that seems unrelated to the student's learning task, giving the example of students climbing on top of (in-game) buildings or putting (in-game) bananas in an (in-game) toilet. They identified this construct as off-task behavior within the learning environment. In 2012, Wixon and colleagues (2012) argued that this term may obscure considerable differences in why students engage in this behavior compared to traditional off-task behavior (where the student ceases to work on the task at all), as well as differences in impact, and suggested the alternate term of WTF behavior.

Rowe and colleagues proposed an operational definition as behaviors that are clearly unrelated to the narrative and curriculum, and built a knowledge-engineered model to infer this construct in a narrative-centered learning environment called Crystal Island, in which students solve scientific puzzles presented through interactive story scenarios. Their definition when expanded by Sabourin et al. (2011) consisted of: interactions with in-game objects that are not relevant to the scientific puzzle, moving a task-related object to an unrelated location, spending too much time in an irrelevant location, or exceeding a height achievable by normal navigation.

Wixon and colleagues (2012) developed a data-mined automated detector of WTF behavior, within the context of Science ASSISTments. Within this environment, WTF behavior involves behaviors such as changing variable values many times without running trials, or rapidly pausing and unpausing a simulation. Ground-truth labels for this behavior were developed using text replays, and then a set of features were distilled using code that had been previously developed to detect student use of experimentation strategies and hypothesis testing within Science ASSISTments. Eleven common classification algorithms were attempted to fit detectors of WTF, and the best model performance was achieved by the Projective Adaptive Resonance Theory Model (PART) algorithm (Frank & Witten, 1998), which produces rules out of C4.5/J48 decision trees. The models were evaluated using a process of sixfold student-level cross-validation, and the detectors were assessed using four metrics: A', Kappa, precision, and recall.

Use of Detectors in Intervention and "Discovery with Models" Analyses

Once detectors of disengaged behavior have been developed, they can be used in two fashions: within "Discovery with Models" analyses to understand the relationship that the disengaged behavior has to other constructs, and within interventions, by embedding the detectors in running software to drive adaptation, and using them to change the system's behavior.

Automated detectors of off-task behavior, gaming the system, carelessness, and WTF behavior have been used in several "discovery with models" analyses. Early analyses on gaming the system indicated that gaming was associated with poorer learning (Baker et al., 2004; Beck, 2005), although fast responses are positively correlated with learning if correctness is not taken into account (e.g., Aleven et al., 2006). This research was followed up by work by Cocea, Hershkovitz, and Baker (2009), who studied whether off-task behavior and gaming the system had an immediate impact on learning or a more aggregate impact on learning, and found evidence that off-task behavior was not associated with worse performance in the short-term, but that it led to the student having fewer opportunities to practice the skill, leading to smaller learning gains over time. By contrast, gaming the system was found to be associated with worse performance in the short-term as well as the long-term. Both Rowe et al. (2009) and Sabourin and colleagues (2011) found evidence that WTF behavior is associated with lower learning gains.

Baker (2007a) used an automated detector of gaming the system to determine whether differences in the frequency of student gaming were better predicted by tutor content than by which student was using the software. Interestingly, knowledge-engineered models have produced the opposite finding, that students were better predictors of gaming behavior than the lesson (Gong et al., 2010; Muldner et al., 2011), a contrasting finding which has not entirely been reconciled, although recent collaborative work between some of the authors of Baker (2007a) and Muldner et al. (2011) suggest that this contrast may be because the different detectors identify different behavior in general. Baker et al. (2009) and Baker (2009) followed up the finding in Baker (2007a) by studying which differences in lesson features predict the degree to which students will go off-task or game the system in a lesson, by combining data from a taxonomy of differences between 22 lessons with assessments of how often a set of students was off-task or gaming in each lesson, across the course of a year. They found that several features were associated with gaming, including ineffective or overly abstract hints, unclear toolbar icons or problem flow, and the lack of (interest-increasing) extraneous text in problem statements.

Detectors of off-task behavior, gaming the system, and carelessness were used in Baker and Gowda (2010) to study the differences in the proportion of these behaviors between students in an urban, rural, and suburban school, across an entire year of usage of a Cognitive Tutor for Geometry. They found that urban school students go off-task and are careless significantly more than rural and suburban school students, and also that gaming the system was most prominent in the urban school. In work within a single population, Rowe and colleagues (2009) found that WTF behavior was significant more common among male students than female students.

San Pedro and colleagues (2011b) used a machine-learned detector of student carelessness to study the relationship between carelessness and affect in high school students using a Cognitive Tutor for Scatterplots. It was found that errors made by students who are confused or bored are less likely to be careless errors. The negative correlation between confusion and carelessness increases in magnitude as students used the tutor more, even as confusion itself decreases in frequency. This suggests that students who were struggling the most and remained confused were less likely to become careless. It was also found that students displaying engaged concentration were more likely to make careless errors, a finding which seems strange but which may be consistent with offline findings in Clements (1982) that successful students are more likely to become careless. Affect was also studied in relation to WTF behavior by Sabourin and colleagues (2011), who investigated the affective role of this category of behavior in Crystal Island. They found that no emotional states were more likely than chance to lead to WTF behavior. However, it was found that students who had remained on-task after reporting confusion were more likely to report feeling focused next, while students who went off-task (in the environment) after reporting confusion were more likely to report boredom or frustration next. It was also found that frustrated students who went off-task (in the environment) were more likely to report feeling focused next, while confused students who went off-task (in the environment) were more likely to report a negative emotion next. By contrast, frustrated students who remained on-task were also more likely to report boredom next. This suggests that this type of behavior may be beneficial to frustrated students by allowing them to distance themselves from the problem.

Research has also been conducted to model the relationship between disengaged behavior and motivational variables. Baker (2007b) found that off-task behavior was associated with disliking computers, disliking mathematics, passive-aggressiveness, and lack of educational self-drive. Baker and Walonoski et al. (2008) investigated which student behaviors, motivations, and emotions are associated with gaming the system, across multiple studies with two different systems. They found that gaming the system was associated with disliking the software's subject matter, lacking self-drive, disliking computers and the learning environment, believing that mathematics ability is innate, and believing that the tutor is not helpful for learning. Beal, Qu, and Lee (2008), using the gaming detectors from Beal, Qu, and Lee (2006), found that students with low math self-concept were most likely to engage in guessing gaming behavior. Hershkovitz and colleagues (2013) studied the relationship between carelessness and motivation within Science ASSISTments, finding carelessness is higher in students characterized by high levels of learning goal orientation and academic efficacy (in the case of academic efficacy, replicating off-line results by Clements [1982]), and high levels of both performance-approach and performance-avoid goals. By contrast, carelessness was lower in students having neither learning nor performance goals. Hershkovitz et al. (2011) also found that students with performance goals demonstrated an increase in carelessness earlier within the set of trials than students with learning goals. On the other hand, students who lacked either type of goal demonstrated consistently higher carelessness over trials.

Detectors of this type have also been used to drive automated interventions, with the goal of improving student engagement and learning. One of the first examples of this can be seen in Baker et al. (2006), where an automated detector of gaming the system was embedded into an interactive agent (similar to non-player characters [NPCs] in games), who displayed negative emotion when students gamed, and who provided supplementary exercises designed to support students in learning material bypassed via gaming. This intervention improved student learning and reducing gaming behavior in the United States (Baker et al., 2006), although its results did not replicate in the Philippines (Rodrigo et al., 2012). Another intervention using gaming detection was developed by Arroyo and colleagues (2007), who provided meta-cognitive messages to students on the negative impact of gaming, combined with visualizations of students' recent gaming behavior. This intervention also reduced gaming and improved learning. Off-task behavior detection, carelessness detection, and WTF behavior detection has not yet been used as the basis of automated intervention, although research projects along these lines are currently underway (Inventado & Numao, 2012).

Discussion, Recommendations, and Future Research

As this chapter indicates, the last several years have seen considerable work in modeling a range of forms of student disengagement, including gaming the system, off-task behavior, WTF behavior, and carelessness. These detectors have been developed through a range of approaches, and a consensus appears to be emerging that disengaged behaviors can be assessed in a range of different online learning environments.

As such, there is an opportunity for a framework such as GIFT to incorporate models of this type. Historically, the ITS field has been better at developing these types of models and using them within discovery with models analyses than it has been in using them to modify tutor pedagogy and adaptivity (although successful examples of the latter exist, particularly for gaming the system – cf. Baker et al., 2006; Walonoski & Heffernan, 2006; Arroyo et al., 2007; Roll et al., 2011). This limitation is holding back the potential of these approaches to provide information that can be used to reengage learners and enhance learning.

A key step that could facilitate incorporation of these types of models into GIFT would be to incorporate tools that support detector-building into the GIFT framework. Several types of tools have been developed to support this process, but the tools have been developed by a variety of research groups and are not well integrated with specific ITSs, or for that matter, with one another. For example, tools have been developed for automated feature generation by a wide variety of research groups, but have been often scoped for use with a single learning environment. One exception is found in the EDM Workbench (Rodrigo et al., 2012), which can generate a range of features for data in the format used by the PSLC DataShop (Koedinger et al., 2010), but can only do so post-hoc. Still, this system's feature generation could be extended and used as a basis for feature generation within the GIFT framework. Ideally, feature generation should be conducted both on existing data sets, and at run time within a tutor, in order to facilitate both the creation and use of automated detectors of disengagement. A tool such as the EDM Workbench could be incorporated into the GIFT framework via creating explicit API-level links between GIFT and the EDM Workbench, where the EDM Workbench could pull data directly from GIFT, and export detectors back to GIFT for use in the user model. In addition, the feature generation code in the EDM Workbench could be integrated into the GIFT framework, so that distilled features could be directly used by detectors exported to GIFT's user model. Simply making these tools available to GIFT users is useful but not sufficient. The process of importing data from GIFT to the EDM Workbench is time consuming if formats do not match, and an unnecessary step compared to the direct approach possible with API-level links. Similarly, the process of manually taking detectors and associated feature distillers from the EDM Workbench and building them into GIFT's user model is challenging if integration is not created between these tools. These issues are not unique to the EDM Workbench, but are general to the problem of using tools for feature distillation or detector building to enhance GIFT. It is worth noting also that tools like the EDM Workbench are useful for modeling a range of behaviors, beyond the disengaged behaviors discussed in this chapter.

A second opportunity is to integrate tools for labeling data in terms of engagement into the GIFT framework. There are currently tools for collecting both text replays (discussed above) and quantitative field observations of disengagement that could be integrated into GIFT. The EDM Workbench offers support for conducting text replays, though in a less visually attractive format than tools designed explicitly for conducting text replays in a single learning system. Tools for generating tailored text replays have also been designed for a variety of learning environments. Integrating text replays into GIFT would be a useful step towards a framework that can incorporate engagement detectors into a wider number of ITSs. Similarly, integrating code into GIFT for conducting field observation of student disengagement, such as the HART app for Android (Ocumpaugh, Baker & Rodrigo, 2012), would support the development of detectors for ITSs using GIFT. Another potential opportunity can be seen in work to develop detectors of carelessness. The development of detectors of carelessness relies upon initial estimates of carelessness computed using student knowledge models. Extending the knowledge modeling in GIFT to produce carelessness labels would be a valuable step towards incorporating this type of adaptation capacity into GIFT.

Through these steps, it will become easier to build automated detectors of student disengagement into the GIFT framework. Doing so will make it feasible to conduct further research on how these models can best be used to reengage learners toward developing understanding in the field as to how ITSs can best adapt to differences in student engagement.

References

Aleven, V., McLaren, B. M., Roll, I. & Koedinger, K. R. (2004). Toward tutoring help seeking: Applying cognitive modeling to meta-cognitive skills. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems* (pp. 227-239).

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.

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 13th International Conference on Artificial Intelligence in Education* (pp. 195-202).

Baker, R. S. J. d. (2007a). Is gaming the system state-or-trait? Educational data mining through the multi-contextual application of a validated behavioral model. In *Complete On-Line Proceedings of the Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling 2007* (pp. 76-80).

Baker, R. S. J. d. (2007b). Modeling and understanding students' off-task behavior in intelligent tutoring systems. In *Proceedings of ACM CHI 2007: Computer-Human Interaction* (pp. 1059-1068).

Baker, R. S. J. d. (2009). Differences between intelligent tutor lessons, and the choice to go off-task. In *Proceedings of the 2nd International Conference on Educational Data Mining* (pp. 11-20).

Baker, R. S. J. d., Corbett, A. T. & Aleven, V. (2008). 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).

Baker, R. S. J. d., Corbett, A. T., Gowda, S. M., Wagner, A. Z., MacLaren, B. M., Kauffman, L. R., Mitchell, A. P. & Giguere, S. (2010). Contextual slip and prediction of student performance after use of an intelligent tutor. In *Proceedings of the 18th Annual Conference on User Modeling, Adaptation, and Personalization* (pp. 52-63).

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).

Baker, R. S. J. d., Corbett, A. T., Koedinger, K. R., Evenson, S. E., Roll, I., Wagner, A. Z., Naim, M., Raspat, J., Baker, D. J. & Beck, J. (2006). Adapting to when students game an intelligent tutoring system. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 392-401).

Baker, R. S., Corbett, A. T., Koedinger, K. R. & Wagner, A. Z. (2004). Off-task behavior in the Cognitive Tutor classroom: When students "game the system". In *Proceedings of ACM CHI 2004: Computer-Human Interaction* (pp. 383-390).

Baker, R. S. J. d., Corbett, A. T., Roll, I. & Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction, 18*(3), 287-314.

Baker, R. S. J. d., Corbett, A. T. & Wagner, A. Z. (2006). Human classification of low-fidelity replays of student actions. In *Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems* (pp. 29-36).

Baker, R. S. J. d. & de Carvalho, A. M. J. A. (2008). Labeling student behavior faster and more precisely with text replays. In *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 38-47).

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". In *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 475-482).

Baker, R. S. J. d. & Gowda, S. M. (2010). An analysis of the differences in the frequency of students' disengagement in urban, rural, and suburban high schools. In *Proceedings of the 3rd International Conference on Educational Data Mining* (pp. 11-20).

Baker, R. S. J. d., Mitrovic, A. & Mathews, M. (2010). Detecting gaming the system in constraint-based tutors. In *Proceedings of the 18th Annual Conference on User Modeling, Adaptation, and Personalization* (pp. 267-278).

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A. & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research, 19*(2), 185-224.

Beal, C. R., Qu, L. & Lee, H. (2006). Classifying learner engagement through integration of multiple data sources. In *Proceedings of the 21st National Conference on Artificial Intelligence* (pp. 2-8).

Beal, C. R., Qu, L. & Lee, H. (2008). Mathematics motivation and achievement as predictors of high school students' guessing and help-seeking with instructional software. *Journal of Computer Assisted Learning, 24*, 507-514.

Beck, J. (2005). Engagement tracing: using response times to model student disengagement. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education* (pp. 88-95).

Buckley, B., Gobert, J. & Horwitz, P. (2006). Using log files to track students' model-based inquiry. In *Proceedings of the Seventh International Conference of the Learning Sciences* (pp. 57-63).

Cetintas, S., Si, L., Xin, Y. P. & Hord, C. (2010). Automatic detection of off-task behaviors in intelligent tutoring systems with machine learning techniques. *IEEE Transactions on Learning Technologies, 3*(3), 228-236.

Clements, M. (1982). Careless errors made by sixth-grade children on written mathematical tasks. *Journal for Research in Mathematics Education, 13*(2), 136-144.

Cocea, M., Hershkovitz, A. & Baker, R. S. J. d. (2009). The impact of off-task and gaming behaviors on learning: Immediate or aggregate? In *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 507-514).

Corbett, A. (2001). Cognitive Computer Tutors: Solving the Two-Sigma Problem. In M. Bauer, P. J. Gmytrasiewicz & J. Vassileva (Eds.), *Proceedings of the 2001 International Conference on User Modeling* (pp. 137-147). Berlin: Springer.

Corbett, A. T. & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*, 253-278.

Frank, E. & Witten, I. H. (1998). Generating accurate rule sets without global optimization. In *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 144-151).

Fredricks, J. A., Blumenfeld, P. C. & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59-109.

Gong, Y, Beck, J. E. & Heffernan, N. T. (2010). Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures. In *Proceedings of the 10th International Conference on Intelligent Tutoring Systems* (pp. 35-44).

Hershkovitz, A., Baker, R. S. J. d., Gobert, J. & Wixon, M. (2011). Goal orientation and changes of carelessness over consecutive trials in science inquiry. In *Proceedings of the 4th International Conference on Educational Data Mining* (pp. 315-316).

Hershkovitz, A., Baker, R. S. J. d., Gobert, J., Wixon, M. & Sao Pedro, M. (in press). Discovery with models: A case study on carelessness in computer-based science inquiry. *American Behavioral Scientist*.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Inventado, P. S. & Numao, M. (2012). Supporting student self-regulation in unsupervised learning environments. In *Proceedings of the 20th International Conference on Computers in Education* (pp. 1-4).

Johns, J. & Woolf, B. (2006). A dynamic mixture model to detect student motivation and proficiency. In *Proceedings of the 21st National Conference on Artificial Intelligence* (pp. 163-168).

Karweit, N. & Slavin, R. E. (1982). Time-on-task: Issues of timing, sampling, and definition. *Journal of Experimental Psychology, 74*(6), 844-851.

Koedinger, K. R., Baker, R. S. J. d., Cunningham, K., Skogsholm, A., Leber, B. & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy & R. S. J. d. Baker (Eds.), *Handbook of Educational Data Mining* (pp. 43-56). Boca Raton, FL: CRC Press.

Lahaderne, H. M. (1968). Attitudinal and intellectual correlates of attention: A study of four sixth-grade classrooms. *Journal of Educational Psychology, 59*(5), 320-324.

Lee, S. W., Kelly, K. E. & Nyre, J. E. (1999). Preliminary report on the relation of students' on-task behavior with completion of school work. *Psychological Reports, 84*, 267-272.

Maris, E. (1995). Psychometric Latent Response Models. *Psychometrika, 60*(4), 523-547.

Martin, J. & VanLehn, K. (1995). Student assessment using Bayesian nets. *International Journal of Human-Computer Studies, 42*, 575-591.

Maydeu-Olivares, A. & D'Zurilla, T. J. (1995). A factor analysis of the social problem-solving inventory using polychoric correlations. *European Journal of Psychological Assessment, 11*(2), 98-107.

Muldner, K., Burleson, W., Van de Sande, B. & VanLehn, K. (2011). An analysis of students' gaming behaviors in an intelligent tutoring system: Predictors and impacts. *User Modeling and User-Adapted Interaction, 21*(1-2), 99-135.

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.

Pardos, Z. A., Baker, R. S. J. d., Gowda, S. M. & Heffernan, N. T. (2011). The sum is greater than the parts: Ensembling models of student knowledge in educational software. *SIGKDD Explorations, 13* (2), 37-44.

Pavlik, P. I., Cen, H. & Koedinger, K. R. (2009). Performance factors analysis -- A new alternative to knowledge tracing. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 531-538).

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Rodrigo, M. M. T., Baker, R. S. J. d., Agapito, J., Nabos, J., Repalam, M. C., Reyes, S. S. & San Pedro, M. C. Z. (2012). The effects of an interactive software agent on student affective dynamics while using an intelligent tutoring system. *IEEE Transactions on Affective Computing, 3* (2), 224-236.

Rodrigo, M. M. T., Baker, R. S. J. d., McLaren, B., Jayme, A. & Dy, T. (2012). Development of a workbench to address the educational data mining bottleneck. In *Proceedings of the 5th International Conference on Educational Data Mining* (pp. 152-155).

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*, 267-280.

Rowe, J., McQuiggan, S., Robison, J. & Lester, J. (2009). Off-task behavior in narrative-centered learning environments. In *Proceedings of the 14th International Conference on Artificial Intelligence and Education* (pp. 99-106).

Sabourin, J., Rowe, J., Mott, B. & Lester, J. (2011). When off-task is on-task: The affective role of off-task behavior in narrative-centered learning environments. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 534-536).

San Pedro, M. O. C., Baker, R. & Rodrigo, M. M. (2011a). Detecting carelessness through contextual estimation of slip probabilities among students using an intelligent tutor for mathematics. In *Proceedings of 15th International Conference on Artificial Intelligence in Education* (pp. 304-311).

San Pedro, M. O. C., Rodrigo, M. M. & Baker, R. S. J. d. (2011b). 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*.

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

Shute, V. J. (1995). SMART: Student modeling approach for responsive tutoring. *User Modeling and User-Adapted Interaction, 5* (1), 1-44.

Walonoski, J. A. & Heffernan, N. T. (2006). Prevention of off-task gaming behavior in intelligent tutoring systems. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 722-724).

Wixon, M., Baker, R. S. J. d., Gobert, J., Ocumpaugh, J. & Bachmann, M. (2012). WTF? Detecting students who are conducting inquiry without thinking fastidiously. In *Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization* (pp. 286-298).