Operationalizing and Detecting Disengagement Within Online Science Microworlds

Janice D. Gobert

Worcester Polytechnic Institute

Ryan S. Baker

Teachers College Columbia University

Michael B. Wixon

Worcester Polytechnic Institute

Abstract

In recent years, there has been increased interest in engagement during learning. This is of particular interest in the science, technology, engineering, and mathematics (STEM) domains, in which many students struggle and where the U.S. needs skilled workers. This article lays out some issues important for framing research on this topic, and provides a review of some existing work with similar goals on engagement in science learning. Specifically, here we seek to help better concretize engagement, a fuzzy construct, by operationalizing and detecting (i.e., identifying using a computational method) *disengaged behaviors* that are antithetical to engagement. We, in turn, describe our real-time detector (i.e., machine learned model) of disengaged behavior and how it was developed. Lastly, we address our on-going research on how our detector of disengaged behavior will be used to intervene in real time in order to better support students' science inquiry learning in Inq-ITS (Inquiry-Intelligent Tutoring System; Gobert, Sao Pedro, Baker, Toto, & Montalvo, 2012; Gobert, Sao Pedro, Raziuddin, & Baker, 2013).

Keywords: science inquiry, disengagement, off-task behavior, intelligent tutoring systems, educational data mining

There has been a surge of interest and research on the topic of engagement in the last twenty years (Christenson, Reschly, & Wylie, 2012). Student engagement is an important topic for teachers, parents, and other stakeholders. Student engagement is critical to study for three reasons (Skinner & Pitzer, 2012). First, it is a necessary condition for students' learning because engagement is a critical component of long-term achievement and academic success (see, for example, Tobin & Sugai, 1999; San Pedro, Baker, Bowers, & Heffernan, 2013). Second, engagement shapes students' school experiences in school, both psychologically and socially (Skinner & Pitzer, 2012). Last, engagement plays a role in students' academic resilience, and the development of resources for coping adaptively with stressors, which in turn, may affect the development of long-term academic mindsets (Skinner & Pitzer, 2012).

In terms of research findings, engagement is associated with positive outcomes along academic, social, and emotional lines (Klem & O'Connell, 2004) and is a very good predictor of students' learning, grades, achievement test scores, retention, and graduation (Appleton, Christenson, & Furlong, 2008). Conversely, disengagement has severe consequences, particularly for students from disadvantaged backgrounds (Fredricks & McColskey, 2012). Disengaged students are less likely to graduate from high school (Fredericks, Blumenfeld, & Paris, 2004; Tobin & Sugai, 1999) and less likely to attend university (San Pedro et al., 2013).

Studying engagement and disengagement in the context of science learning is important for many reasons (Hug, Krajcik, & Marx, 2005). First, although approaches to supporting STEM learning have changed considerably over the last century, one key aspect of science education remains the same: students become disengaged and fall behind; thus, if not addressed, this is likely to continue. Second, successful learning of science skills and concepts is increasingly necessary to students’ future success both on high-stakes exams and in determining access to and success in STEM careers (Autor, Levy, & Murnane, 2003). Increasingly, engagement is thought to be critical to addressing low achievement and high dropout rates (for a review, see Fredericks et al., 2004). For example, off-task behavior as early as middle school is an excellent predictor of high school dropout rate (Tobin & Sugai, 1999), and gaming the system in middle school predicts eventual college attendance (San Pedro et al., 2013).

Difficulties with Engagement Research

Engagement is an integral aspect of learning, but is difficult to directly operationalize and observe, and due to this, it has been historically difficult to define and thus to measure (Fredricks et al., 2004). This is particularly true in science learning (Tytler & Osborne, 2012), likely due to its complexity and difficulty. The large variation in how this construct is conceptualized and measured has made it challenging to compare findings across studies (Fredricks & McColskey, 2012; Appleton et al., 2008). However, this very diversity in conceptualizing engagement has led to an acknowledgement and appreciation of the complexity of this construct (Skinner & Pitzer, 2012).

This article lays out some issues important to framing research in the area of engagement in science. As noted by Tytler and Osborne (2012), what is needed are better theoretical models that can account for student engagement (and disengagement) in science. Specifically, we seek to shed further light on how to operationalize engagement, a fuzzy construct. We do this by defining and identifying behaviors that are associated with disengagement, turning the construct "on its head" to define, operationalize, and detect (i.e., identify using a computational technique) engagement by identifying its opposite, *dis*-engagement. We provide a review of existing work on engagement with similar general goals to ours. We briefly lay out some presuppositions and operational terms, as well as describe other methodological approaches in order to contextualize our work on the development of an automatic detector of disengaged behavior (i.e., a machine learned model of disengaged behavior; Wixon, Baker, Gobert, Ocumpaugh, & Bachmann, 2012; Wixon, 2013) in our online science learning environment, Inq-ITS (Inquiry Intelligent Tutoring System; Gobert et al., 2012, 2013) We, in turn, describe our real-time detector of one form of disengaged behavior, *Disengaged from Task Goal* (DTG), and give an overview how it was developed. Lastly, we outline some of our on-going research on how our detector of disengaged behavior can be used to intervene in real time as students work in Inq-ITS in order to better support students' science inquiry learning.

Presuppositions and Terms

Engagement vs. Flow

It is important to differentiate engagement from flow (Csikszentmihalyi, 1990). We do not conceive of these necessarily as different in type but different in degree. Whereas we define engagement as being on task and aligned to the designer's goals, flow is often referred to as a very deep state of engagement that lead the learner to lose a sense of one's self. Specifically, flow is conceptualized as a state of deep absorption, as intrinsically enjoyable, as worthwhile for its own sake, and in which the individual functions at his or her fullest capacity (Shernoff, Csikszentmihalyi, Schneider, & Steele Shernoff, 2003).

School Engagement vs. Student Engagement

An important distinction is made between school engagement and student engagement. The former concerns other educational constructs (such as school bonding, belonging, and school “climate”). Here we address student engagement as the student is *oriented towards* learning that is intended by the system’s designers.

Relationship Between Engagement and Motivation

Another key issue to be underscored is the relationship between engagement and motivation. In the past, researchers generally tended to reflect motivation and engagement within a single theoretical framework and conceptualized disengagement as emerging, at least in part, from variables including student attributes and presuppositions. This is represented by research using instruments such as the PALS (Patterns of Adaptive Learning Survey; Midgley, Maehr, Hicks, Roeser, Urdan, & Anderman, et al., 1997), and the MSLQ (Motivated Strategies for Learning Questionnaire; Pintrich, Smith, Garcia, & McKeachie, 1991). For example, *goal orientation*, measured by the PALS, includes the goal of achieving mastery, the goal of avoiding failure, and the goal of avoiding work. These have been frequently hypothesized as associated with disengagement in online learning, although these relationships have not been borne out (e.g. Baker, Walonoski, Heffernan, Roll, Corbett, & Koedinger, 2008; Beal, Qu, & Lee, 2007). By contrast, relationships have been found between low grit (see, for example, Duckworth, Peterson, Matthews, & Kelly, 2007) and disengaged behaviors in online learning (Baker, Walonoski, et al., 2008). Given the known relationships between grit and student learning outcomes (Duckworth et al., 2007), it is worth studying whether disengaged behaviors mediate the relationship between lack of grit and negative learning outcomes, as positive learning strategies mediate the relationship between having learning goals or positive effort beliefs and learning (Blackwell, Trzesniewski, & Dweck, 2007). Addressing these questions effectively requires reliable and precise means of detecting when, and to what degree, students manifest disengaged behaviors. Our detector of disengagement is designed for this purpose.

Engagement as a Separate Construct from Motivation.

There are many researchers who conceptualize motivation and engagement as different constructs and posit poor motivation as the underlying reason for a given disengaged behavior. Research in this vein presupposes that engagement itself is a multi-dimensional construct. Briefly, within this perspective, 2-, 3-, and 4-component models have been proposed regarding the components of engagement. For example, Martin (2008) proposes a two-dimensional model comprising mainly cognitive and behavioral dimensions. Three-dimensional models (see, for example, Fredricks, Blumenfeld, & Paris, 2004) add an emotional component to the cognitive and the behavioral. Emotional engagement includes interest, boredom, happiness, anxiety, and other affective states. Behavioral engagement includes persistence, effort, attention, participation, involvement. Lastly, cognitive engagement includes cognitive investment in learning, meta-cognition, and self-regulated learning. Many have adopted this 3-part framework for work in this area (see Sinatra, Heddy, & Lombardi, this issue). A four-part model has also been proposed by Christenson and her colleagues (Appleton, Christenson, Kim, & Reschly, 2006; Reschly & Christenson, 2006), who added an academic component as a fourth dimension which includes time on task, credits earned, and homework completion. However, to us, these types of variables are better aligned with school engagement (as opposed to student engagement), thus, our work fits better conceptually under the three-component model, as described by Fredricks et al (2004).

Engagement is Malleable and Contextually-based

Recent theories of engagement have made a major advance by no longer conceptualizing engagement as an *attribute of the student*, but rather as a *malleable state* that is influenced by school, family, peers, tasks, and other factors (Reschly & Christenson, 2006). More specifically, and important to our persepecitve, engagement arises from the interaction of the individual with the context, task, etc. (Finn & Rock, 1997; Fredricks et al., 2004; Skinner & Pitzer, 2012). Furthermore, because the action component of student engagement with academic tasks is observable, it can be tracked at the level of individual students (Skinner et al., 2009). In our work, these manifestations of disengagement are derived from students' log files of their interactions in Inq-ITS.

Prior Methods of Measuring Engagement

With our presuppositions and terms operationalized, it is important to review prior work on the development of measures of engagement/disengagement in science. These, with their pros and cons, are briefly reviewed below.

Self-Report Surveys of Engagement

Self-report, often conducted via a pre- or post-test survey, is one of the most commonly used methods for assessing student engagement (see Fredricks & McColskey, 2012; Greene, this issue) because this method is practical, easy to administer in classroom settings, and is low cost for use with large numbers of students. Several of these types of measures have been previously validated by others, which reduces the workload of validating measures anew, and makes comparisons to others' work easier (Shea, & Bidjerano 2010; Liu, Horton, Olmanson, & Toprac, 2011). However, there are cons to this approach as well. Importantly, many of these surveys differ in terms of how they conceptualize engagement (Fredricks & McColskey, 2012). A second concern with surveys is that they are often applied out of context, either before or after an activity (Harmer & Cates, 2007). In this case surveys are measuring participants' self-report of their earlier or later engagement rather than in the context in which it is occurring. Thus, interpreting data about the relationships between engagement and specific learning tasks is problematic. Whereas methods exist for collecting self-report in real-time, these methods are often disruptive to students. Thirdly, students may not answer survey questions honestly (Appleton et al., 2006), negatively impacting the validity of the results. Fourthly, items are often worded broadly rather than to reflect engagement in targeted tasks and contexts. In sum, these methods have been criticized for being highly inferential (Appleton et al., 2006).

Field Observations & Teacher Ratings

One of the common methods for obtaining data on student engagement is to use field observations, where an observer watches students in the setting of learning, and codes engagement multi-dimensionally in real time. There is a long history of coding student off-task behavior using field observations stretching back over fifty years (Lahaderne, 1968). In the 1980s, researchers began to extend field observations of engagement to involve a wider range of behavior (e.g. Reyes & Fennema, 1981), which has since been extended to include affective as well as behavioral indices of engagement (see Fredericks et al., 2004; Olitsky, 2007; Ryu & Lombardi, this issue).

Within field observations, the data that is coded can be qualitative in nature (Papastergiou, 2009) or employ a quantitative coding method to determine whether a pre-determined category of behavior is present or absent for an individual student during a defined time interval as indicative of engagement (Annetta, Minogue, Holmes, & Cheng, 2009; Birch & Ladd, 1997). While field observations can be effective, they are time-consuming, and field coders need training. *Quantitative* field coding in particular, which uses researcher-developed categories that are used by human coders, offers an advantage in that it draws from richer data including subtle behaviors such as posture, facial expression, tone of speech, eye gaze, etc. Additionally, by employing human judgment to identify engagement, quantitative field observations have a benefit typically associated with qualitative methods: they avoid mechanistically operationalizing participant behaviors, thereby improving construct validity.

Field observations can be also used to create automated measures of engagement through data mining on log files. The most common field observation method of this type is BROMP 2.0, the Baker-Rodrigo-Ocumpaugh Monitoring Protocol (Ocumpaugh, Baker, & Rodrigo, 2012); the first version of BROMP (Baker et al., 2004a) was built off of earlier methods in field observation (Fennema et al., 1996). BROMP coders record the affective state and current engaged/disengaged behavior of each student individually, in a pre-determined order that is enforced by the Human Affect Recording Tool (HART) application (Baker et al., 2012) for the Android phone. This strict ordering avoids bias towards interesting or dramatic events in the classroom, ensuring that categories such as "engaged concentration" are accurately represented in the data. Coders have up to 20 seconds to make and verify their assessment, but record only the first affective state and behavior they identify. To build detectors of the affective states identified by the coders, field observations are synchronized with the log files of student interactions using the software, and HART synchronizes each observation to within 2 seconds of internet time, allowing researchers to accurately match each field observation window to the 20-second clip of that student's interactions that are recorded in the software's log file. The observers base their judgment of a student's affect or behavior on the student's work context, actions, utterances, facial expressions, body language, and interactions with teachers or fellow students, in line with Planalp et al.'s (1996) descriptive research on how humans generally identify affect, using multiple cues in concert for maximum accuracy rather than attempting to select individual cues.

Given that the number of potential observations per student is limited, BROMP is not an ideal method for studying the development is disengagement over relatively focused periods of time, and cannot provide the level of precision of estimate of a method that can provide continual estimation of student engagement (such as log file based methods). However, it can be used as the basis for obtaining the human ground truth measures (i.e., the accuracy of the training set's data) of engagement needed to build automated detectors of affect (Baker et al., 2004b; Baker, 2007).

Log File and Activity-based Measurements

Another potential method for measuring engagement is via the use of automated detectors, which infer engagement from student behavior in online learning. These measurements rely on identifying behaviors that are quantifiable using log files generated as students work in online learning environments. These methods may vary in both the dimensions of grain-size of the logs and in validity of the behavior's relationship to engagement. At its coarsest grain size, logs from an entire experimental condition can be compared to those of a control group to address differences in engagement levels. For example, Minner, Levy, & Century (2010), used differences in log files to show that students in a constructivist environment were more engaged compared to students in a more instructional environment. Automated detectors can be developed which identify student engagement from log files at a second-by-second level (Baker et al., 2004b; Beck, 2005), in particular from student pauses and self-regulated learning behaviors. Rather than simply comparing two conditions as in a typical randomized controlled trial, these analyses allow researchers to explore more complex or conditional ways that engagement may function (e.g. the sequences of behaviors that students perform). Later in the text, we detail an example of automated detection of student disengagement from log files, as was done in the development of our detector for DTG.

Mixed Methods

It is important to note that all the methods outlined previously are often not exclusive of one another. In several cases, a qualitative analysis may lend itself to a better-informed quantitative coding scheme. Likewise, data derived from a quantitative coding scheme or survey measures may inform the hypotheses and expectations of researchers in performing qualitative studies. For example, resource-intensive quantitative field observations may be used by a researcher who is developing log file-based models of engagement. In turn, these models can be used to identify engagement in a practical and scalable way (Pardos et al., 2013; San Pedro et al., 2014). In cases where quantitative models are shown to be inadequate or generalize poorly to different populations or settings, researchers are best served by returning to qualitative methods to identify the factors that their models are not capturing (Ryu & Lombardi, this issue).

Prior Work on Engagement in Online Science Learning Environments

We turn now to research that is most closely associated to our goal of addressing engagement and disengagement within science learning environments. We do so in order to contextualize our development work on our detector (i.e., machine learned model) of disengagement.

Studying disengaged behaviors in the context of online learning is a recent field of study. Research has shown that many students engage in haphazard and non-goal directed behaviors during inquiry and problem solving (Buckley, Gobert, Horwitz, & O'Dwyer, 2010); one possible explanation for this is disengagement. Some forms of disengaged behavior in online learning have been shown to not only have immediate effects on domain learning (see, for example, Baker et al., 2004a), but also have shown to result in lower achievement on standardized exams (Pardos et al., 2013), and even to lead to lower probability of attending college (San Pedro et al., 2013).

While some researchers refer to a single behavior pattern as "disengagement" (e.g., Beck, 2005; Cocea & Weibelzahl, 2009), work over the last several years has suggested that learners can disengage from learning in several ways. Instead of engaging deeply in online science learning, many students disengage by (a) *gaming the system* (Baker et al., 2004); (b) by engaging in *off-task behavior* (Baker, 2007) or haphazard learning (Buckley et al., 2010); (c) by becoming careless and giving wrong answers due to lack of effort rather than lack of knowledge (Hershkovitz, Baker, Gobert, Nakama, 2012); or (d) by engaging in *player transformation* (Magnussen & Misfeldt, 2004). All of these behaviors can occur within traditional learning settings as well as in online environments (Clements, 1982; Karweit & Slavin, 1981; Nelson-LeGall, 1985). Each of these is briefly addressed below.

Gaming the System

Some researchers have studied how students game the system, attempting to succeed in an educational task by systematically taking advantage of properties in the system used to complete that task, rather than by deeply thinking through the material (Baker et al., 2004a). One of the first identified forms of gaming the system in online learning was *help abuse* (Aleven et al., 2004), which occurs when students, who are capable of solving problems, exploit scaffolding help to avoid cognitive effort. Some hint systems employ progressively more direct forms of help in their hints. One common form of help abuse, namely, “clicking through hints” occurs when students rapidly ask for additional help without taking time to read the initial hints, which give away less of the solution strategy (Aleven, McLaren, Roll & Koedinger, 2004). Another common strategy is systematic guessing, where students quickly and systematically try different answers to find a solution (Baker et al., 2004a). It has been shown that gaming the system has a statistically significant negative correlation with mathematics pre- and post-tests, a finding replicated in multiple studies (Baker et al., 2004a; Cocea et al., 2009). Gaming the system is also associated with lower achievement on standardized examinations (Pardos et al., 2013), as well as lower eventual college attendance (San Pedro et al., 2013).

Off-Task Behavior

Off-task behavior is typically understood as a student disengaging completely from the learning task to participate in unrelated activity (Karweit & Slavin, 1981), for example, surfing the web for material unrelated to the learning task. Although off-task behavior has been found to have low but replicable negative correlations to learning in traditional academic settings (Frederick & Walberg, 1980; Goodman, 1990), these effects have not been thus far replicated in online learning (Baker, 2007; Cocea et al., 2009) or longer-term learning outcomes (Pardos et al., 2013).

Player Transformation

Students also may transform the learning task to a different task entirely (Magnussen & Misfeldt, 2004); this is called player transformation. This sort of behavior is characterized by re-conceptualizing a learning task as a game or other structured activity. For instance, students may choose to focus on helping each other in an online learning activity, rather than trying to succeed themselves, to get a high score in a system designed to reward helping behaviors. While player transformation is a relatively underdeveloped concept compared to gaming the system, it seems to be characterized by “play”, whereas gaming the system is characterized by “exploitation.”

Rationale

We conduct our work in the context of a science inquiry environment (Inq-ITS system; **I**nquiry **I**ntelligent **T**utoring **S**ystem; www.inq-its.org), a computer based learning environment designed to hone inquiry skills using microworlds (Gobert et al., 2012, 2013). Although it is intuitively likely that there are domain-general aspects to engagement, as well as aspects specific to science engagement, it is beyond the scope of this article to address these similarities and differences. But, since Inq-ITS logs all students' fine-grained interactions within the system as they conduct science inquiry, this has the affordance of generating data with which to develop a detector to identify disengagement. In this sense, we are studying engagement specific to science learning.

As described in the prior work on this topic addressed earlier in the article, there is now a pressing and articulated need to provide conceptual clarity and methodological rigor to identify disengagement, which to date, has not been achieved (Glanville & Wildhagan, 2007; Skinner et al., 2009). This can, in our view, be achieved by working towards establishing construct validity for student engagement. Furthermore, mixed methods (the approach taken here) may be likely very productive for this goal (Fredricks & McColskey, 2012). Lastly, rigorous, real time, domain-specific measures of engagement are needed because prior measures make it difficult to examine engagement within its specific context (Fredricks & McColskey, 2012). Our work here addresses this need for online science inquiry, making it possible, in time, to intervene when students become disengaged within our online environment, Inq-ITS. Some domain-specific methods have been developed for mathematics (Kong, Wong, & Lam, 2003) and for reading (Wigfield et al., 2008); thus, our work adds to the existing work in these domains.

Our work also builds on earlier work on behavioral engagement that tended to focus on whether a student was primarily on-task or off-task (see, for example, Lahaderne, 1968; Karweit & Slavin, 1981). Research since then has begun to consider the multiple ways that disengagement or engagement can manifest behaviorally (see, for example, Finn and Rock, 1997). However, most work on engaged and disengaged behaviors still focuses on a student's overall incidence of each behavior (see, for example, Fredericks et al., 2011). We extend this approach using educational data mining methods to produce a new measure that is fine-grained and can be applied at scale. Specifically, our method, described next, identifies indicators of a specific disengaged behavior, Disengaged from Task Goal (DTG), within Inq-ITS (Inquiry Intelligent tutoring System; Gobert et al., 2012, 2013). As has been argued elsewhere, this computational approach in which we identify what log features are critical for predicting a skill and/or online behavior can help further refine the construct under study (Sao Pedro et al., 2011).

Disengaged from Task Goal (DTG)

In addition to the forms of disengagement identified earlier (e.g., gaming the system, player transformation, and off task behavior), there are additional ways in which a student can interact with learning tools that are not focused on using the learning environment as it was intended by the instructional designer. We operationalize this as “Disengaged from Task Goal.” This type of behavior has been seen in online learning, but given a variety of names in the published literature (Sabourin, Rowe, Mott, & Lester, 2013; Buckley et al., 2010; Wixon et al., 2012). In one example, in one of the authors’ data collection sessions, students plotted points from a function in a cognitive tutor for high school mathematics instead plotted a smiley face. In another example, referred to as off-task behavior by the authors, learners chose to obtain virtual cacti and put them on top of virtual patients, rather than trying to determine why the patients were sick (Sabourin, et al., 2013). In a third example, referred to as haphazard inquiry by the authors, learners play around with a science simulation in a fashion unrelated to the stated learning goals of the simulation (Buckley et al., 2010).

In the context of online science learning, DTG may take several forms, including running an inordinately large number of identical trials, changing most of the variables repeatedly within a single trial, and toggling a variable back and forth repeatedly for no discernible reason. Later in the article we describe the features in students' log files that were identified as relevant to detecting disengagement. We label this behavior as "Disengaged from Task Goal" (DTG) rather than as off-task behavior, as the behaviors are different in nature. Off-task behavior typically involves disengaging completely from the learning task, whereas in DTG the student is engaging with the task, but in a fashion unrelated to the learning task's design goals or incentive structure. As such, it is not clear whether the two behaviors emerge for the same reasons, whether they impact learning in similar way(s), and whether they can be detected by the same automated models.

There are several steps in developing a detector such as this one. Each step will be described briefly after a description of the sample upon which our detector was built (a fuller description can be found in Wixon, 2013 and Wixon et al., 2012).

Sample and Microworld Overview

The detector developed in the work reported here was based on data produced by 144 eighth graders (generally ages 12-14), who used Inq-ITS (Inquiry Intelligent Tutoring System, Gobert et al., 2012, 2013), specifically, its Phase Change microworld, within their science classes. All students 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% oqualified 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 (Figure 1), students observe and manipulate variables in the simulation in order to conduct inquiry regarding the changes between solids, liquids, and gases. In terms of inquiry phases, 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 (NRC, 2013). In the Phase Change microworld in which students melt a block of ice in a beaker using a Bunsen burner, the independent variables that the students can change include amount of ice, flame intensity, size of beaker, and whether or not the beaker is covered. In turn, the values for the relevant dependent variables including time needed to melt the ice, time needed to boil the resulting water, the melting point of the ice, and the boiling point of the water are represented in a data table for the students.

[Insert Figure 1 about here]

Each of the students completed at least one data collection activity in the phase change environment. In this article, we focus on student actions in the hypothesizing and experimentation phases of the activity. As students conducted these tasks, their actions within the software were logged– for a total of 144,841 actions were generated. Logs included the action type, the relevant simulation variable values, and the time stamp.

Steps in Detector Development

The first step in our process of developing a data-mined detector of DTG behavior is to develop ground truth labels, using text replays (Baker et al., 2006). In text replays, human coders are presented “pretty-printed” versions of log files. Text replays have proven effective for providing ground truth labels for disengaged behaviors (Baker & de Carvalho, 2008; Baker, Mitrovic, & Matthews, 2010).

In order to create text replays, the student data was segmented into “clips,” i.e., sequences of student behavior. In this approach, a clip begins when a student enters the data collection phase and ends when the student leaves that phase of inquiry. The typical order of student actions in Inq-ITS 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 that the same student produced within the same activity so that they could detect DTG behavior that might otherwise have been missed due to lack of context. For example, a student may repeatedly switch between hypothesizing and experimentation, running the 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 DTG.

Two human coders practiced coding DTG on two sets of clips that were excluded from use in detector development. In the first set of clips, they coded together and discussed coding standards. Next, the two coders each coded a second set of 200 clips independently. The two coders achieved acceptable agreement, with Cohen’s Kappa of 0.66.

Afterwards, 571 clips were coded to develop the DTG detector. Because 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 represented approximately equally in the 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 DTG behavior, a proportion similar to the proportions of disengaged behavior studied in past detector development (Baker & de Carvalho, 2008). These 15 clips were drawn from 15 (10.4%) of the students (i.e., no student was coded as engaging in DTG behavior more than once).

Data Features

In order to develop an automated detector of DTG 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 students' use of experimentation strategies and testing the correct hypothesis within Inq-ITS (Sao Pedro et al., 2013). These are general features used to distill features of students' performance within a microworld. Given that many of these features did not appear relevant to detecting DTG behavior and using a greater number of features increases the risk of over-fitting in general (Mitchell, 1997), 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 features that were identified as relevant to detecting disengagement are briefly described in Table 1. These fit under 5 categories: (a) overall statistics for the clip, (b) features related to pauses during the run of the simulation, (c) features based on the time elapsed during experimentation, (d) features related to resetting or pausing the experimental apparatus (or the absence of this action), and (e) features involving changes to variables while forming hypotheses.

[Insert Table 1 about here]

These categories are, to us, intuitively meaningful. For example, under category (b), *pausing the simulation while it is running* can be appropriate in some situations, but doing so large numbers of times may be an indicator of DTG behavior, as the point of the simulation is to demonstrate the pattern of the phenomenon in question so stopping the simulation repeatedly while it is running is intuitively plausible an indicator of disengagement. Additionally, under category (d), *features about variable changes* are indicative of disengagement because extremely large numbers of changes would not align to any reasonable experimentation strategy during inquiry. Similarly, under category (e), *making many changes to the independent variable(s) during hypothesis formation* seems like an indicator of disengagement because the student is not acting in a systematic fashion by forming a hypothesis and then experimenting towards that hypothesis.

Detector Development

Our detector (i.e., machine learned model) of DTG in this particular context was built using a machine learning approach to determine the relationships between the features (i.e., variables) in the model rather than relying on operationalization by an expert. In machine learning, an algorithm is given access to 24 features, which is then used to construct a model associating those variables, thereby leveraging experts' ability to recognize a behavior while obviating the risk confirmation bias through researcher operationalization. This is analogous to creating a linear regression model using a set of variables: generation of the model only relies on the researcher's beliefs insofar as which variables are input as predictors. The main difference between a classification algorithm used here and a linear regression is that the resulting model of linear regression comes in the form of a linear equation, while our classification algorithms produce models composed of conditional if-then statements or "rules."

We attempted to fit detectors of DTG with machine learning using 11 common classification algorithms. A classification algorithm is a model that attempts to predict a binomial or polynomial variable (in this case, a binomial variable, whether an example of student behavior represents DTG behavior or not), using a combination of other variables. Out of those eleven algorithms, the best model performance was achieved by the PART algorithm (Frank & Witten, 1998). A full description of how PART classification models are constructed is out of the scope of this article (see Frank & Witten, 1998 for a comprehensive, several-page technical description), but the resultant model is a set of if-then rules which are considered in order. For example, the first rule is checked and provides a single answer (either DTG, or not DTG) and a confidence for that answer. If the first rule does not apply, the second rule is checked, and so on.

For these analyses, we create PART trees using the RapidMiner 4.6 data mining software (Mierswa et al., 2006); the implementation of PART used within RapidMiner was originally developed as part of the open-source data mining software

WEKA (Witten & Frank, 1999). These models were evaluated using a process of six-fold student-level cross-validation (Efron & Gong, 1983). 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.

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

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 DTG**.
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 DTG**.
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 DTG**.
4. IF the total number of independent variable changes (feature 21) is 12 or less, THEN **DTG**.

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 DTG**.
6. All remaining instances are classified as **DTG**.

As can be seen, this detector used 6 rules, determined by machine learning, to distinguish DTG behavior, which employ 8 features from the data set. Four of the rules identify the characteristics of behavior that is NOT DTG, while only two identify the characteristics of DTG behavior.

Detector Evaluation

The detector was assessed using four metrics, A' (Hanley & McNeil, 1982), Kappa, precision (Davis & Goadrich, 2006), and recall (Davis & Goadrich, 2006). A' is the probability that the detector will be able to distinguish a clip involving DTG behavior from a clip that does not involve DTG behavior. A' is equivalent to both the area under the ROC curve in signal detection theory and to W, the Wilcoxon statistic (Hanley & McNeil, 1982). 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 (Rosenthal & Rosnow, 1991). However, the standard error formula for A' (Hanley & McNeil, 1982) requires multiple examples from each category for each student, which is not feasible 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 article.

The second metric used to evaluate the detector was Cohen’s Kappa, which assesses whether the detector is better than chance at identifying which clips involve DTG behavior. A Kappa of 0 indicates that the detector performs at chance, and a Kappa of 1 indicates that the detector performs perfectly. The detector was 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 (Table 2).

[Insert Table 2 about here]

A' and Kappa were chosen because they compensate for successful classifications occurring by chance, an important consideration in data sets with unbalanced proportions of categories (such as this case, where DTG 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.

The detector of DTG behavior developed using the PART algorithm achieved good performance under 6-fold student-level cross-validation. The detector achieved a very high A' of 0.8005, signifying that it could distinguish whether or not a clip involved DTG behavior approximately 80.05% of the time. When uncertainty was not taken into account, performance was lower, though still generally acceptable. The detector achieved a Kappa value of 0.411, indicating that the detector was 41.1% better than chance. This level of Kappa is comparable to past automated detectors of other constructs effectively used in interventions (Sao Pedro et al., 2013; Baker & de Carvalho, 2008). Kappa values in this range, combined with good A' values, suggest that the detector is generally good at recognizing which behavior is more likely to be "DTG," 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 no evidence that the detector has strong bias either in favor of or against detecting DTG behavior.

What Does our Detector Reveal About Disengagement in Inq-ITS?

Examining the model of DTG behavior (described in detail in Wixon et al., 2012 and Wixon, 2013) provides some interesting implications about disengagement. Previous automated detectors of disengaged behavior have largely focused on identifying the specific undesirable behavior studied (Baker & de Carvalho, 2008; Baker, Mitrovic, & Mathews, 2010; Cetintas et al., 2009). By contrast, the rules produced by our detector are targeted more towards identifying what is not DTG behavior than identifying what *is* DTG behavior. As such, this model suggests that DTG behavior may be characterized by the absence of appropriate strategies and behaviors in a student actively using the software, as well as specific undesirable behavior.

It is also worth discussing the data feature that is most frequently employed in the model rules is the number of times the student changes a simulation variable. Though this feature is used in four of the six rules of the model (Wixon et al., 2012), there is not a clear pattern where frequently changing variables is simply either good or bad. Instead, different student actions appear to indicate DTG 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 DTG behavior. By contrast, a student who changes variables fewer times is categorized as displaying DTG 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 this time to do some meaningful during the pauses, such as study the simulation.

Discussion

Summary of Approach

In this article, we first presented a detector (i.e., machine learner model) of what we term Disengaged from Task Goal (DTG), based on data from the Phase Change microworld in Inq-ITS (Gobert et al., 2012, 2013). In this type of disengagement, the student is interacting with the software, but their actions appear to have little relationship to the intended learning task and/or the designer's goals. DTG behavior has been reported in multiple online learning environments, but has not yet been modeled or studied to the degree that it warrants.

We also presented an overview of the detector development process using human labels of the behavior and educational data mining techniques, and described how the detector was validated. Our work is a proof of concept that this behavior can be identified both by human coding of log files and by an automated detector. It is important to note that our automated detector of disengagement can be used to replicate the identification of scoring of engagement in a more practical and scalable way (see, for example, Pardos et al., 2013), thereby providing opportunities for fine-grained basic research on this construct as well as empirical studies testing the efficacy of interventions based on disengagement. Lastly, our data show that this behavior has prevalence similar to another index of disengagement, namely, gaming the system, a behavior known to be associated with poor learning outcomes (Baker et al., 2004a; Cocea et al., 2009; Pardos et al., 2013).

Value Added by the Detector

Our work addresses two main issues, which we, and others, see as pressing and imperative (Glanville & Wildhagan, 2007) for the field of research on engagement to continue to move forward. The first is the need for measures of engagement that are well aligned to the current theoretical position that engagement is highly contextualized (Fredricks et al., 2004) – a standard met by this detector that can infer a specific disengaged behavior within Inq-ITS. The second related need is to study engagement in a way that acknowledges that engagement is malleable (Appleton et al., 2008; Fredricks et al., 2004; Reschly & Christenson, 2006). Each is addressed in turn.

Precise measures of engagement. As previously stated, there has been a problem in this research area due to the lack of precision in defining and thus, in identifying engagement. Following Finn and Kasza (2009), we believe that engagement needed more clearly defined boundaries. We addressed this by operationalizing its counterpart, namely disengagement, very concretely, and in turn, developing a method using a computational technique to identify disengagement while students are engaging in online inquiry within Inq-ITS (Gobert et al., 2012, 2013). Ours is the first (to our knowledge) automated detector of this type.

Our development approach uses machine-learning techniques (i.e., educational data mining) to identify disengagement in real time within the context of learning. As such, our method is an advance over the most commonly used method, namely, self-report (Fredricks & McCloskey, 2012), in which items are often worded too broadly to reflect engagement in targeted tasks and contexts (Appleton et al., 2006) and are often administered out of context. While self-report can be obtained at a moment-to-moment level, doing so frequently is disruptive, and doing so retrospectively risks inaccuracy (see, for example, Porayska-Pomsta et al., 2013).

Automated detectors can be developed either using field observations (Ocumpaugh et al., 2012) or text replay hand-tagging of log files (as was done in the work here). Field observations are more time-intensive but more appropriate for constructs that cannot be assessed by human coders solely from log files. In our development process, resource-intensive hand labeling of log data was used as "ground truth," as opposed to an operationalized rubric, to obtain human judgments that were used, in turn, to derive fine-grained log file-based models of disengagement. This approach has the advantage of leveraging both the benefits of the activity-based, nuanced character of qualitative methods and the rigor of having a precise automated measure that can be applied at a very fine grain-size. Another advantage is that once developed, the detector can also be at scale (Pardos et al., 2013).

Studying engagement and its interactions. Another benefit of the automated detector approach is that its fine-grain size allows for an in-depth exploration of participants' behaviors, and thus, allows for greater refinement in the conclusions that may be drawn from analyses. Rather than simply comparing two conditions as in a typical randomized controlled trial, these analyses allow researchers to explore more complex or conditional ways that engagement may function (e.g. the sequences of behaviors that students perform). As such, automated detectors of disengagement can be applied with consistency across studies, making it easier to compare findings across studies, as well as provide data about the relationship between disengagement and learning (see, for example, Cocea et al., 2009; Pardos et al., 2013), and the precursors of disengagement (see, for example, Baker, D’Mello, Rodrigo, & Graesser, 2010).

Another important application for automated detectors of disengagement such as DTG behavior is to study the individual differences and situational factors leading students to disengage from learning. By measuring various types of disengaged behavior separately, we can better understand the factors leading to the emergence of different types of disengaged behavior, for example a student’s choice to misuse a learning simulation rather than simply going off-task. We will also be able to study how different types of disengaged behavior impact learning differently (see, for example, Cocea et al., 2009; Pardos et al., 2013). For example, DTG behavior could be expected to emerge for several reasons, including attitudinal reasons such as not valuing the learning task, or affective states such as confusion, frustration, and boredom. Previous research has shown that affect is associated with differences in future disengaged behavior (Baker et al., 2010). Regarding off-task behavior, Sabourin et al. (2011) found that students who go off task when they are confused later can become bored or frustrated; by contrast, students who go off task when they are frustrated often become re-engaged later in the task. These findings suggest that intelligent tutors should offer different interventions, depending on the affective context of disengaged 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 who is DTG may need additional support in understanding how to learn from the learning environment. By contrast, a student who is DTG 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 (Pekrun, 2006). By applying automated detectors, it will become feasible to study this behavior across a greater number of situations (Baker et al., 2009), helping us to better understand the factors leading to DTG behavior. By understanding the causes of DTG 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 across the full range of inquiry activities offered in Inq-ITS.

Our detector as a means to scaffold students' engagement. Our detector can identify disengagement moment-to-moment as students use Inq-ITS; this is consistent with a conceptualization of engagement as malleable (Appleton et al., 2008; Fredricks et al., 2004; Reschly & Christenson, 2006). Automated detectors are a potentially important resource for intervening when students become disengaged because teachers can often have negative reactions to students' disengagement. Specifically, several studies have shown that teachers often withdraw their support from disengaged students, which in turn, exacerbates student disengagement (Skinner & Pitzer, 2012). Baker et al. (2006) and Arroyo et al. (2007) have shown that for gaming the system, automated interventions based on detectors can be an effective method for reducing gaming and improving learning.

Because we now have a valid and reliable method of identifying disengagement for Inq-ITS, we can develop automated interventions targeted to get students back on track in real time directly via our pedagogical agent, Rex, a cartoon dinosaur who currently provides scaffolds to students in real time on their inquiry skills (Sao Pedro, 2013). Rex can prompt students to re-engage in meaningful academic activities with playful feedback to get students back on track before critical knowledge in STEM is missed. The advantage of doing this via a pedagogical agent is that it reacts objectively, not judgmentally, and without other students in the class knowing that the student needed intervention. In doing so we can provide empirical data about how malleable engagement is because interventions to remediate this behavior and get students back on track could be tested. In this way, our disengagement detector provides “value added” to both the field of engagement (Fredricks et al., 2004), and in future, to personalized, adaptive learning environments. Additionally, this would add to the growing set of strategies for engagement intervention (Christenson et al., 2012).

All told, research along these lines will support the field in developing and testing the next-generation theory about engagement, and its relationship to other constructs, such as motivation and learning, as well as allowing researchers to develop interventions that target very specific kinds of disengaged behavior (Martin, 2007; 2008).

References

Aleven, V., McLaren, B., Roll, I., & Koedinger, K.R. (2006). Toward Meta-cognitive Tutoring: A Model of Help-Seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education, 16*, 101-130.

Annetta, L. A., Minogue, J., Holmes, S. Y., & Cheng, M. T. (2009). Investigating the impact of video games on high school students’ engagement and learning about genetics. *Computers & Education*, *53*(1), 74-85. http://dx.doi.org/10.1016/j.compedu.2008.12.020

Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools, 45*, 369-386. http://dx.doi.org/10.1002/pits.20303

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology, 44*, 427-445. http://dx.doi.org/10.1016/j.jsp.2006.04.002

Arroyo, I., Ferguson, K., Johns, J., Dragon, T., Meheranian, H., Fisher, D., ... & Woolf, B. P. (2007, June). Repairing disengagement with non-invasive interventions. In *AIED* (Vol. 2007, pp. 195-202).

Autor, D., Levy, F., & Murnane, R.J. (2003). The Skill Content of Recent Technological Change: An Empirical Exploration. *The Quarterly Journal of Economics*, 118 (4), 1279-1333. http://dx.doi.org/10.1162/003355303322552801

Baker, R. S. (2007) Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems. *Proceedings of ACM CHI 2007: Computer-Human Interaction*, 1059-1068.

Baker, R. S., 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* (No. 2002, pp. 29-36).

Baker, R. S., & 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." *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 475-482.

Baker, R.S., Corbett, A.T., & Koedinger, K.R. (2004b). Detecting Student Misuse of Intelligent Tutoring Systems. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, 531-540. http://dx.doi.org/10.1007/978-3-540-30139-4_50

Baker, R. S., Corbett, A., Koedinger, K., & Wagner, A. (2004a). Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game the System." *Proceedings of ACM CHI 2004: Computer-Human Interaction*, 383-390.

Baker, R. S., D'Mello, S.K., Rodrigo, M.M.T., & Graesser, A.C. (2010). Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners'

Cognitive-Affective States during Interactions with Three Different Computer-Based Learning Environments. *International Journal of Human-Computer Studies*, 68 (4), 223-241. http://dx.doi.org/10.1016/j.ijhcs.2009.12.003

Baker, R. S., Mitrovic, A., & Mathews, M. (2010). Detecting Gaming the System in Constraint-Based Tutors. Proceedings of the *18th Annual Conference on User Modeling, Adaptation, and Personalization*, 267-278. http://dx.doi.org/10.1007/978-3-642-13470-8_25

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. (2008). Mathematics motivation and achievement as predictors of high school students' guessing and help-seeking with instructional software. *Journal of Computer Assisted Learning*, *24*(6), 507-514. http://dx.doi.org/10.1111/j.1365-2729.2008.00288.x

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

Birch, S., & Ladd, G. (1997) The teacher-child relationship and children’s early school adjustment. *Journal of School Psychology, 35*, 61-79. http://dx.doi.org/10.1016/S0022-4405(96)00029-5

Blackwell, L., Trzesniewski, K., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, 78, 246–263. http://dx.doi.org/10.1111/j.1467-8624.2007.00995.x

Boesch, E. E. (1976). *Psychopathologie des altags [Everyday psychopathology]*. Bern, Switzerland: Huber.

Bradstadter, J. (1998). Action perspectives on human development. In. W. Damon (Series Ed.) & R. M. Lerner (Vol. Ed.) *Handbook of child psychology: Vol. 1. Theoretical models of human development* (pp. 807-863). New York: Wiley

Brophy, J. (2004). *Motivating students to learn* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Buckley, B.C., Gobert, J., Horwitz, P. & O'Dwyer, L. (2010). Looking inside the black box: Assessments and decision-making in BioLogica. *International Journal of Learning Technologies, 5(2)*, 166 - 190. http://dx.doi.org/10.1504/IJLT.2010.034548

Cetintas, S., Si, L., Xin, Y. P., & Hord, C. (2009). Automatic detection of off-task behaviors in intelligent tutoring systems with machine learning techniques. *Learning Technologies, IEEE Transactions on*, *3*(3), 228-236. http://dx.doi.org/10.1109/TLT.2009.44

Chapman, M. (1984). Intentional action as a paradigm for developmental psychology: A symposium. *Human Development, 27*, 113-114. http://dx.doi.org/10.1159/000272908

Christenson, S., Reschly, A. L., & Wylie, C. (2012). *Handbook of research on student engagement*. New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-2018-7

Cocea, M., Hershkovitz, A., & Baker, R.S. (2009). The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate? *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 507-514.

Connell, J.P., & Wellborn, J. G. (1991). Competence, autonomy, and relatedness: A motivational analysis of self-system processes. In M. R. Gunnar & L. A. Sroufe (Eds.), *Self-processes and development: Minnesota symposium on child psychology* (Vol. 23, pp. 43-77). Chicago: University of Chicago Press.

Csíkszentmihályi, M. (1990). Flow: the psychology of optimal experience. New York, NY: Harper & Row.

Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233-240). ACM.

Duckworth, A.L., Peterson, C., Matthews, M.D., & Kelly, D.R. (2007). Grit: Perseverance and Passion for Long-Term Goals. *Journal of Personality and Social Psychology, 92* (6), 1087-1101. http://dx.doi.org/10.1037/0022-3514.92.6.1087

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, *37*(1), 36-48.

Finn, J.D., & Kasza, K. A. (2009). Disengagement from School. In *Engaging young people in learning: Why does it matter and what can we do?* (pp. 4-35). Wellington, New Zealand: New Zealand Council for Educational Research.

Finn, J. D., & Rock, D. A. (1997). Academic success among students at risk for school failure. *Journal of applied psychology*, *82*(2), 221. http://dx.doi.org/10.1037/0021-9010.82.2.221

Frank, E., & Witten, I. H. (1998). Generating Accurate Rule Sets Without Global Optimization. Proceedings of the *Fifteenth International Conference on Machine Learning*, 144–151

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. (2004). School engagement: Potential of the concept: State of the evidence. *Review of Educational Research, 74,* 59-119. http://dx.doi.org/10.3102/00346543074001059

Fredricks, J., McColskey, W., Meli, J., Mordica, J., Montrosse, B., & Mooney, K. (2011). Measuring student engagement in upper elementary through high school: A description of 21 instruments. *Issues & Answers Report, REL*, *98*, 098.

Frederick, W., & Walberg, H. (1980). Learning as a Function of Time. *The Journal of Educational Research*, Vol. 73, No. 4, 183-194. http://dx.doi.org/10.1080/00220671.1980.10885233

Glanville, L., & Wildhagen, T. (2007). The measurement of school engagement: Assessing dimensionality and measurement in variance across race and ethnicity. *Educational and Psychological Measurement*, *6*, 1019-1041. http://dx.doi.org/10.1177/0013164406299126

Gobert, J. D., Sao Pedro, M. A., Baker, R. S., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *JEDM-Journal of Educational Data Mining*, *4*(1), 111-143.

Gobert, J., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics for science inquiry using educational data mining. *Journal of the Learning Sciences, 22(4),* 521-563. http://dx.doi.org/10.1080/10508406.2013.837391

Greene, B. A. (in press). Measuring cognitive engagement with self-report scales: Reflections from over twenty years of research. *Educational Psychologist*.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36. http://dx.doi.org/10.1148/radiology.143.1.7063747

Harmer, Andrea J., and Ward Mitchell Cates. (2007). Designing for learner engagement in middle school science: Technology, inquiry, and the hierarchies of engagement. *Computers in the Schools 24.1-2*, 105-124. http://dx.doi.org/10.1300/J025v24n01_08

Hershkovitz, A., Baker, R.S.J.d., Gobert, J., & Nakama, A. (2012). A Data-driven Path Model of Student Attributes, Affect, and Engagement in a Computer-based Science Inquiry Microworld. In *Proceedings of the International Conference on the Learning Sciences*.

Hug, B., Krajcik, J. S., & Marx, R. W. (2005). Using innovative learning technologies to promote learning and engagement in an urban science classroom. *Urban Education*, *40*(4), 446-472. http://dx.doi.org/10.1177/0042085905276409

Karweit, N.L., & Slavin, R.E. (1981). Measurement and modeling choices in studies of time and learning. *American Educational Research Journal, 18*, 157-171. http://dx.doi.org/10.3102/00028312018002157

Klem, A. M., & Connell, J. P. (2004). Relationships matter: Linking teacher support to student engagement and achievement. *Journal of School Health, 74(7)*, 262-273. http://dx.doi.org/10.1111/j.1746-1561.2004.tb08283.x

Kong, Q., Wong, N., & Lam, C. (2003). Student engagement in mathematics: Development of instrument and validation of a construct. *Mathematics Education Research Journal*, *54*, 4-21. http://dx.doi.org/10.1007/BF03217366

Lahaderne, H. M. (1968). Attitudinal and intellectual correlates of attention: a study of four sixth-grade classrooms. *Journal of Educational Psychology*, *59*(5), 320. http://dx.doi.org/10.1037/h0026223

Liu, M., Horton, L., Olmanson, J., & Toprac, P. (2011). A study of learning and motivation in a new media enriched environment for middle school science. *Educational Technology Research and Development*, *59*(2), 249-265. http://dx.doi.org/10.1007/s11423-011-9192-7

Magnussen, R., & Misfeldt, M. (2004). Player transformation of educational multiplayer games. *Proceedings of Other Players*, Copenhagan, Denmark.

Martin, A. J. (2008). Enhancing student motivation and engagement: The effects of a multidimensional intervention. *Contemporary Educational Psychology*, *33*(2), 239-269. http://dx.doi.org/10.1016/j.cedpsych.2006.11.003

McCaslin, M. M., & Good, T. L. (1996). *Listening in classrooms*. New York: HarperCollins.

Miceli, M., & Castelfranchi, C. (2000). Nature and mechanisms of loss of motivation. *Review of General Psychology*, *4(3)*, 238-263. http://dx.doi.org/10.1037/1089-2680.4.3.238

Midgley, C., Maehr, M. L., Hicks, L., Roeser, R., Urdan, T., Anderman, E., ... & Middleton, M. (1997). Patterns of adaptive learning survey (PALS). *Ann Arbor, MI: The University of Michigan*.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006, August). Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 935-940). ACM. http://dx.doi.org/10.1145/1150402.1150531

Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching*, *47*(4), 474-496. http://dx.doi.org/10.1002/tea.20347

Mitchell, T. (1997). *Machine Learning*. New York, NY: McGraw Hill, 1997.

National Research Council (2013). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, D.C.: National Academies Press.

Nelson-LeGall, S. (1985). Help-Seeking Behavior in Learning. *Review of Research in Education, 12*, 55-90.

Newmann, F., Wehlage, G. G., & Lamborn, S. D. (1992). The significance and sources of student engagement. In F. Newmann (Ed.), *Student engagement and achievement in American secondary schools* (pp. 11-39). New York: Teachers College Press.

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.

Olitsky, S. (2007) Promoting student engagement in science: Interaction rituals and the pursuit of a community of practice. Journal of Research in Science Teaching 44, 33-56. http://dx.doi.org/10.1002/tea.20128

Papastergiou, M. (2009). Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation.*Computers & Education*, *52*(1), 1-12. http://dx.doi.org/10.1016/j.compedu.2008.06.004

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. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge*, 117-124. http://dx.doi.org/10.1145/2460296.2460320

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, *18*(4), 315-341. http://dx.doi.org/10.1007/s10648-006-9029-9

Pintrich, P. R., Smith, D. A. F., García, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)*. Ann Arbor: University of Michigan, National Center for Research to Improve Postsecondary Teaching and Learning.

Planalp, S. (1996). Varieties of cues to emotion in naturally occurring situations. *Cognition & Emotion*, *10*(2), 137-154. http://dx.doi.org/10.1080/026999396380303

Porayska-Pomsta, K., Mavrikis, M., D'Mello, S., Conati, C., & Baker, R. S. (2013). Knowledge elicitation methods for affect modelling in education.*International Journal of Artificial Intelligence in Education*, *22*(3), 107-140.

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

Reschly, A., & Christenson, S. L. (2006). Promoting school completion. In G. Bear, & K. Minke (Eds.), *Children's needs III: Understanding and addressing the developmental needs of children*. Bethesda, MD: National Association of School Psychologists.

Reyes, L., & Fennema, E. (1981). Classroom Processes Observer Manual. Buffalo, NY: U.S. Department of Education, National Center for Educational Statistics, ERIC Document Reproduction Service No. ED224793.

Rosenthal, R. & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

Sabourin, J. L., Rowe, J. P., Mott, B. W., & Lester, J. C. (2013). Considering alternate futures to classify off-task behavior as emotion self-regulation: A supervised learning approach. *JEDM-Journal of Educational Data Mining*, *5*(1), 9-38.

Sabourin, J., Rowe, J. P., Mott, B. W., & Lester, J. C. (2011, January). When off-task is on-task: The affective role of off-task behavior in narrative-centered learning environments. In *Artificial Intelligence in Education* (pp. 534-536). Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-21869-9_93

San Pedro, M. O. Z., Baker, R. S., Bowers, A. J., & Heffernan, N. T. (2013). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In *Proceedings of the 6th international conference on educational data mining* (pp. 177-184).

San Pedro, M.O.Z., Baker, R.S.J.d., & Mercedes, M.M.T. (2014). Carelessness and Affect in an Intelligent Tutoring System for Mathematics. *International Journal of Artificial Intelligence in Education, 24*, 189-210. http://dx.doi.org/10.1007/s40593-014-0015-y

Sao Pedro, M.A. (2013). *Real-time Assessment, Prediction, and Scaffolding of Middle School Students' Data Collection Skills within Physical Science Simulations*. Social Science and Policy Studies: Learning Sciences and Technologies Program Ph.D. Dissertation. Worcester Polytechnic Institute Technical Report etd-042513-062949.

Sao Pedro, M.A., Baker, R.S.J.d., Gobert, J., Montalvo, O., & Nakama, A. (2011). Leveraging Machine-Learned Detectors of Systematic Inquiry Behavior to Estimate and Predict Transfer of Inquiry Skill. *User Modeling and User-Adapted Interaction.* http://dx.doi.org/10.1007/s11257-011-9101-0

Shea, P., & Bidjerano, T. (2010). Learning presence: Towards a theory of self-efficacy, self-regulation, and the development of a communities of inquiry in online and blended learning environments. *Computers & Education*, *55*(4), 1721-1731. http://dx.doi.org/10.1016/j.compedu.2010.07.017

Shernoff, D., Csikszentmihalyi, M., Schneider, B., & Steele Shernoff, E. (2003). Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly, 18(2)*, 158-176. http://dx.doi.org/10.1521/scpq.18.2.158.21860

Skinner, E. A., Kindermann, T. A., Connell, J. P., & Wellborn, J. G. (2009a). Engagement as an organizational construct in the dynamics of motivational development. In K. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 223-245). Mahwah, NJ: Erlbaum

Skinner, E. A., Kindermann, T. A., & Furrer, C. (2009). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and Psychological Measurement*, *69*, 493-525. http://dx.doi.org/10.1177/0013164408323233

Skinner, E. A., Marchand, G., Furrer, C., & Kindermann, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic. *Journal of Educational Psychology, 100(4)*, 765-781. http://dx.doi.org/10.1037/a0012840

Skinner, E. A., & Pitzer, J. R. (2012). Developmental dynamics of student engagement, coping, and everyday resilience. In *Handbook of research on student engagement* (pp. 21-44). Springer US. http://dx.doi.org/10.1007/978-1-4614-2018-7_2

Spires, H. A., Rowe, J. P., Mott, B. W., & Lester, J. C. (2011). Problem solving and game-based learning: Effects of middle grade students' hypothesis testing strategies on learning outcomes. *Journal of Educational Computing Research*,*44*(4), 453-472. http://dx.doi.org/10.2190/EC.44.4.e

Tobin, T. J., & Sugai, G. M. (1999). Using sixth-grade school records to predict school violence, chronic discipline problems, and high school outcomes. *Journal of*

Emotional and Behavioral Disorders, *7*(1), 40-53. http://dx.doi.org/10.1177/106342669900700105

Tytler, R., & Osborne, J. (2012). Student attitudes and aspirations towards science. In *Second international handbook of science education* (pp. 597-625). Springer Netherlands. http://dx.doi.org/10.1007/978-1-4020-9041-7_41

Volpe, R. J., DiPerna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing students in classroom settings: A review of seven coding schemes. *School Psychology Reviews, 34(4*), 454-474.

Wigfield, A., Guthrie, J. T., Perencevich, K. C., Taboada, A., Klauda, S. L., McRae, A., & Barbosa, P. (2008). Role of reading engagement in mediating effects of reading comprehension instruction on reading outcomes. *Psychology in the Schools*, *45*(5), 432-445. http://dx.doi.org/10.1002/pits.20307

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wixon, M. (2013). *Detecting students who are conducting inquiry Without Thinking Fastidiously (WTF) in the Context of Microworld Learning Environments*. Master Thesis, WPI, Worcester, MA.

Wixon, M., Baker, R. S., Gobert, J. D., Ocumpaugh, J., & Bachmann, M. (2012). WTF? detecting students who are conducting inquiry without thinking fastidiously. In *User Modeling, Adaptation, and Personalization* (pp. 286-296). Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-31454-4_24

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 awarded to Janice Gobert and Ryan Baker. We thank Michael Sao Pedro for assistance in data distillation and for helpful comments and suggestions; we thank Arnon Hershkovitz and Adam Nakama for helpful comments and suggestions in the development of the detector, and thank Yoon Jeon Kim for helpful comments on the final draft of the manuscript.

Table 1

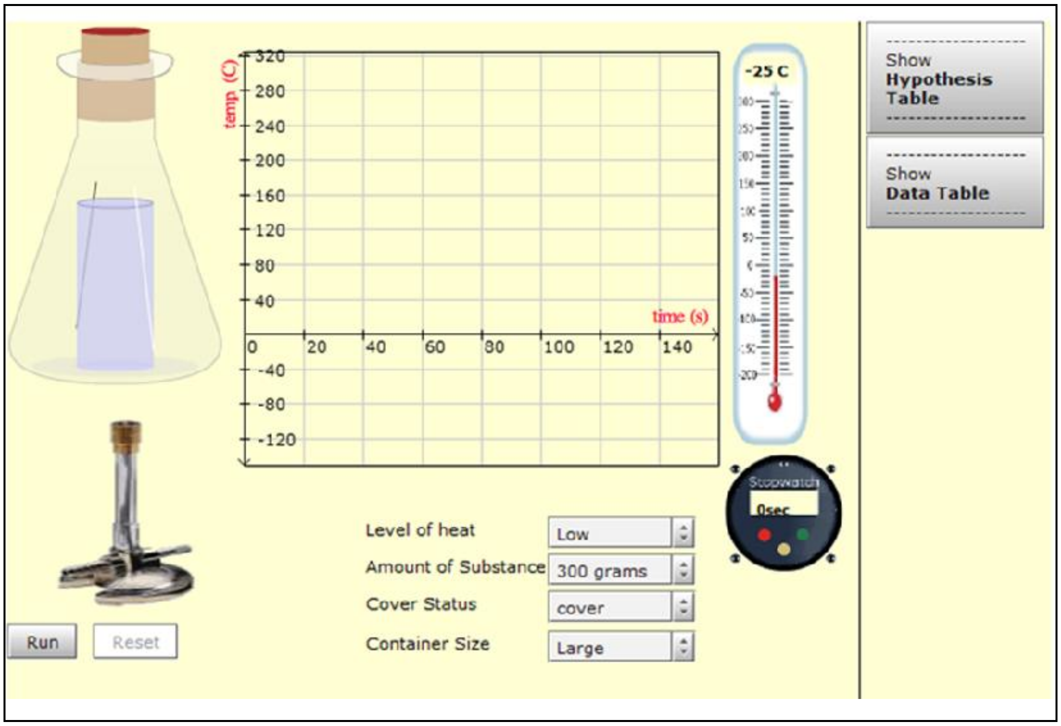
Features Used in Detector Development

(a) Overall Statistics	(b) Features Based on Pauses	(c) Features Based on Time Elapsed During Experiment Phase	(d) Features Related to Resetting or Pausing Experimental Apparatus	(e) Features Involving Changes to Variables While Forming Hypotheses
Total # of actions	Number of pauses to simulation during runs	Total time spent before running each experimental trial (but after performing the previous action)	Number of experimental trials run without either pauses or resets	Number of changes to independent variable(s) during the experiment phase
Average time between actions	Average duration of student-initiated pauses of the simulation	Average time spent by the student before running each experimental trial (but after performing the previous action)	Average time spent by the student before running each experimental trial which was completed without being reset (but after performing the previous action)	Period of time elapsed before the student changed a variable for: (a) the sum total of time elapsed in all these periods, (b) the mean time elapsed across these periods, and (c) the standard deviation of time elapsed across these periods
Average time between actions	Duration of the longest single pause	Standard deviation of time spent before running each trial (but after performing the previous action)	Number of trials where the system was reset	
Maximum time between actions		Maximum time spent before running each experimental trial (but after performing the previous action)	Average time spent before running each experimental trial that were reset (but after performing the previous action)	
Number of experimental trials			Maximum time spent before running an experimental trial that was reset before completion (but after performing the previous action)	

Table 2

Disengaged from Task Goal Detector Confusion Matrix

	Clips Coded as DTG by Humans	Clips Coded as NOT DTG by Humans
Detector Predicted DTG	7	10 (false positives)
Detector Predicted NOT DTG	8 (false negatives)	476



Level of heat Low

Amount of Substance 300 grams

Cover Status cover

Container Size Large

Figure 1. A screen shot of the Phase Change microworld (early version) in Inq-ITS.