2013-04-09

Detecting students who are conducting inquiry Without Thinking Fastidiously (WTF) in the Context of Microworld Learning Environments

Michael Wixon
Worcester Polytechnic Institute

Detecting students who are conducting inquiry Without Thinking Fastidiously (WTF) in the Context of Microworld Learning Environments

by

Michael Wixon

A Master's Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science

in

Social Science & Policy Studies

May 2013

Approved:

Dr. Janice D. Gobert, Advisor

Abstract

In recent years, there has been increased interest and research on identifying the various ways that students can deviate from expected or desired patterns while using educational software. This includes research on gaming the system, player transformation, haphazard inquiry, and failure to use key features of the learning system. Detection of these sorts of behaviors has helped researchers to better understand these behaviors, thus allowing software designers to develop interventions that can remediate them and/or reduce their negative impacts on student learning. This work addresses two types of student disengagement: carelessness and a behavior we term WTF ("Without Thinking Fastidiously") behavior. Carelessness is defined as not demonstrating a skill despite knowing it; we measured carelessness using a machine learned model. In WTF behavior, the student is interacting with the software, but their actions appear to have no relationship to the intended learning task. We discuss the detector development process, validate the detectors with human labels of the behavior, and discuss implications for understanding how and why students conduct inquiry without thinking fastidiously while learning in science inquiry microworlds. Following this work we explore the relationship between student learner characteristics and the aforementioned disengaged behaviors carelessness and WTF. Our goal was to develop a deeper understanding of which learner characteristics correlate to carelessness or WTF behavior. Our work examines three alternative methods for predicting carelessness and WTF behaviors from learner characteristics: simple correlations, k-means clustering, and decision tree rule learners.

Acknowledgements

There are a lot of people I have to thank for their help in my work. I'd like to thank my advisors Dr. Janice D. Gobert & Dr. Ryan S. J. d. Baker for their advice in analytics, educational psychology, data mining, and most importantly writing. I've come a long way in terms of writing thanks to them, and before they wince at being given responsibility for the current state of my writing I'd like to thank them for also letting me know I have a long way to go and helping me on the way there. In all honesty though, I feel that I have been spoiled in terms of advisors. You guys have always been there for me in a pinch, and the things you have taught me will be invaluable in my future studies.

After my advisors I owe a huge amount of thanks to Michael Sao Pedro. Mike's always been there to advise or commiserate, but I owe him the most for impressing upon me the importance of making finishing this thesis a priority. If your help here is any indication Mike, I think you could make a great advisor someday soon. In addition to all that Mike wrote the scripts to assess inquiry skills, without those scripts I wouldn't have been able to detect carelessness.

Many thanks to Dr. Ivon Arroyo for agreeing to read this thesis and for offering to give helpful comments for my defense, I hope you enjoy this work. I've really enjoyed working with Ivon, and hope to continue our collaborations in the future.

I had a huge amount of help from several colleagues who made this work possible. Matt Bachmann deserves a lot of credit for text replay coding 200 clips for WTF behavior, also for helping me laugh even when things were at their most stressful; he truly is a one man riot. I'd like to thank Dr. Arnon Herskovitz for his work in cluster analysis of learner characteristics with carelessness. I'd like to thank Sujith Gowda for his BKT scripts, without which the carelessness detector wouldn't have been possible. A big thanks to Dr. Joseph Beck for being willing to listen to my travails and offer a healthy dose of criticism, though I wasn't able to incorporate all his suggestions, he certainly gave me a lot to think about. Also, we should all be thanking Ermal Toto and Andy Montalvo for developing Science Assistments, now called Inq-ITS. Their contributions to Learning Sciences at WPI cannot be overstated and often go unsung. Without them I would have neither data nor a learning environment to study, so I cannot possibly overstate their importance.

I'd like to thank my girlfriend Adrian for her help in copy-editing my earlier drafts and for being willing to listen to me whine about a truly awesome job.

There are more friends than I can list to thank for their diversions, advice, and just generally being willing to listen to my rants. I can list a few key players around the lab though, so I'd like to call out Juelaila Raziuddin, Jaclyn Ocumpaugh, Mike Brigham, Adriana Joazeiro, Nate Krach, Adam Nakama, Zak Rogoff, Supreeth Gowda, Lakshmi Shankar, Cameron Betts, and Lyndon Johnson in no particular order.

Finally, I want to thank my parents Jo Ann Bennett and Dennis Wixon. I recognize how fortunate I am to have parents who spent inordinate amounts of time reading and conversing with me, time that's definitely reflected in my GRE scores. On a meta-level I'd like to thank them for modeling the pleasure and reward that comes with learning, and challenging ideas.

Table of Contents

Introduction

The following research addresses disengaged behaviors: behaviors that are not directed toward learning from a task as intended by a tutoring system's design. Given this disconnect between intended use and student behaviors, there is concern that disengaged students may learn less efficiently [1 - 3]. In this work we will be focusing on two forms of disengaged behaviors. The first behavior we have identified as "Without Thinking Fastidiously" or WTF, a form of disengaged behavior in which the student uses the learning system but engages in a manner that does not appear to be directed toward completing curriculum goals. The second disengaged behavior we shall examine is carelessness. The goals of this work are first to design automated means of detecting WTF and carelessness and second to search for correlations or other relationships between these disengaged behaviors and student learner characteristics. In so doing, our intention is to enable future research into the relationship between disengaged behaviors and academic performance. A deeper understanding of the factors that play a role in disengaged behaviors such as WTF may eventually enable us to intervene in real time in order to avoid potentially inefficient learning strategies.

Prior work along these lines has been done to identify the disengaged behavior called "gaming the system". Gaming the system is defined as "attempting to succeed in an educational environment by exploiting properties of the system rather than by learning the material and trying to use that knowledge to answer correctly" [1]. While it would seem that reducing disengaged behaviors would always be desirable, such behaviors may not be harmful to learning in all cases [1, 4]. In some situations, apparently disengaged behaviors may act as self-regulation strategies [cf. 5] by using bottom-out hints as worked examples to aid learning [6]. Designing successful interventions for disengaged behaviors requires knowing when such interventions are necessary, and moreover what cognitive processes lead to disengaged behaviors. The following work addresses the first three steps of this process, predicting and identifying the: 1) likelihood of disengagement, 2) when said disengagement is harmful, and 3) what cognitive processes may lead to disengagement.

Herein, we examine disengagement in the context of science inquiry within the Science Assistments system, now referred to as Inq-ITS (www.inq-its.org). Inq-ITS is a computer based assessment system that is designed to hone inquiry skills through the use of inquiry tasks with "microworlds" [7, 8]. Inquiry skills are the underlying critical thinking skills used when applying the scientific method to a problem [9]. By observing students' log files we can examine the range of student actions and inquiry skills amongst students, allowing for rich inferences about students' cognitive processes during inquiry. It is also important to note that disengagement has not been widely studied in the context of inquiry skill based tutors or microworlds, though similar work has been conducted within Crystal Island [10, 11], a narrative centered learning environment. As such this work makes novel contributions to this literature.

Literature Review

We draw from literature on disengaged behaviors, learner characteristics, and detector development.

Disengaged Behaviors

There are several ways a student can interact with learning tools. In many, but not all cases, disengaged behaviors have been shown to be detrimental to learning [12]. Disengaged behaviors are a broad class of behaviors including off task behavior, gaming the system, carelessness, and "without thinking fastidiously". In the context of online learning disengaged behaviors are a fairly recent field of study, the constructs of carelessness and WTF have not yet been studied to the same degree as off-task behavior and gaming the system. As such, we shall provide descriptions of extant work on disengagement, i.e., primarily off-task behavior and gaming the system.

Off-Task Behavior

John Carroll accounted for task-orientation, whether or not a student is attending to a task, in his early Model of School Learning [13]. One of his initial points was to question how effectively time was being used, "'Time' is therefore not 'elapsed time' but time during which the person is oriented to the learning task and actively engaged in learning." Several studies [14-16] supported Carroll's argument; measuring "actual" time used, rather than time scheduled, this refinement revealed improved association of time to outcome measures [15].These early models were intended to improve efficiency and time-use in the school environment and lead to more nuanced models of time on task. Even given this greater drive for subtlety when modeling task-orientation, there was still a need for greater refinement; for example: several studies have modeled a student's interaction with the "task" at hand as a simplified binary variable [17, 18], either on- or off-task. Much in the same way that models of time on task evolved from scheduled hours of work to actual hours of work, our current work in disengagement is intended to address a key problem with the on task/off task binary variable by refining from actual hours of work to actual hours of productive work. If a student is carelessly doing their work, they are considered on task, but not engaged. This need for greater detail in descriptions of a student's engagement with a task belie the need for broadening our model from "off-task" behaviors to include a variety of disengaged behaviors.

There are several ways in which a student can interact with learning tools that may not promote learning as effectively as others, or may serve purposes wholly unrelated to learning. Logged student actions provide a means of identifying and tracking the finer grained subtleties and distinctions that separate new forms of disengaged behaviors from the broad categories of on- and off-task behavior [1,2, 19]. These methods and constructs represent an advancement beyond both Carroll's time on task measure, as well as Lloyd and Loper & Lee et al.'s [17,18] on vs off-task measures of disengagement.

Unfortunately, the new subtle distinctions regarding student engagement can create semantic confusion. For example in the intelligent tutoring system (ITS) "Crystal Island", students were observed behaving in ways that were deemed unproductive for learning, e.g., exploring parts of the island unrelated or unnecessary to the curriculum's stated goal [10, 11]. This behavior was termed "off-task behavior". While it is true that students in this case were not engaging with the learning task as intended, they were actively engaging with the learning environment itself, raising questions about whether this behavior should be defined as "off-task". This category of behavior will be discussed further in the later section "Without Thinking Fastidiously (WTF).

Gaming the System

Gaming the system is defined as "attempting to succeed in an educational environment by exploiting properties of the system rather than by learning the material and trying to use that knowledge to answer correctly" [1]. One of the first identified forms of gaming the system behavior was help abuse[20], where students "click through hints", by rapidly asking for additional help without taking time to read the initial hints, which give away less of the solution strategy [20]. Another form of gaming is systematic guessing, which means systematically and rapidly trying many answers until one turns out to be correct [1]. It has been shown that gaming the system has a statistically significant negative correlation with mathematics pre- and post-test scores [1, 12]. Students who gamed the system performed poorly on both the pre- and post-tests while students who performed poorly on the pre-test and did not game might still perform well on the post-test [1]. Gaming behavior is also associated with lower learning rates [21].

While gaming the system has shown to be indicative of poorer learning in several cases, it is important to note that there are also instances in which gaming behavior does NOT reduce student learning. Gaming has also been displayed by students with either high pretest scores, or high pretest to posttest gains, specifically when they game on already well-known material [1, 4]. In addition, students who game the system by using bottom out hints as worked examples show improved learning in the form of pre to posttest gains, students' use of these hints as worked examples were evidenced by time devoted to parsing these examples. Students who viewed bottom out hints and took additional time outperformed the general population who took additional time, illustrating that bottom out hints yielded a gain over what would be expected through a traditional time on task model [5]. This finding indicates that disengagement is not always indicative of poor performance or learning.

Carelessness

Another form of disengaged behavior, which has received less attention than off task behavior, is carelessness. Newman [22] developed the "Newman interview" in which students who generally gave correct answers, but occasionally answered questions incorrectly were interviewed to determine if their incorrect responses were due to a lack of knowledge or a "careless error" [23]. An operational definition of carelessness would be a situation where a student gets a question wrong, then on a subsequent opportunity gets a question utilizing the same skill right without interim opportunities to learn the skill in question. The reverse, a situation where a student performed well on numerous occasions and then

made an error, could also be categorized as carelessness. This situation of simultaneously knowing a skill while making an error in performing it corresponds to contextual slip as described in Bayesian Knowledge Tracing. While the construct of carelessness has been of interest for some time [23], methods of automated carelessness detection are only now being developed [19, 24]. Larger scale studies of this behavior can provide greater insight to the origins and mechanisms behind careless errors [19].

Without Thinking Fastidiously (WTF)

"Without thinking fastidiously" (WTF) is a form of disengaged behavior in which a student uses the learning system but engages in a manner that does not appear to be targeted toward completing curriculum goals. The construct of WTF grew out of qualitative observations of seemingly problematic student behaviors in our log files: students would interact with the microworlds in erratic ways. The actions these students performed seemed random, but were so involved they did not appear to be accidental. An example in the context of a classroom setting would be if a student were to fill in the bubbles of a Scantron exam to make the pattern of a smiley face. In the context of inquiry focused microworlds, WTF may take the form of running an inordinately large number of identical trials, changing most of the variables suggesting a total disregard for control for variables strategy (CVS), or toggling a variable back and forth repeatedly for no discernible reason.

The construct we have identified as WTF has been termed off-task behavior by researchers at North Carolina State University [10, 11]. In a study of the ITS "Crystal Island", students were observed behaving in ways that were deemed unproductive for learning e.g., exploring parts of the island unrelated or unnecessary to the curriculum's stated goal, climbing trees, and placing bananas in toilets. We view this behavior as being different from a student who is completely disengaged from the learning environment, and therefore view placing bananas in toilets as WTF behavior. The semantic difference between Sabourin's model of these behaviors and our own is that we recognize subtle distinctions between disengaged behaviors while Sabourin generalizes WTF under the aegis of Off-Task behavior.

Learner Characteristics

Learner characteristics serve as a broad group of measures relating to achievement goal theory, self-efficacy, and several student centered attributes such as prior performance in academic contexts. Our goal was to see how carelessness and WTF relate to learner characteristics. Furthermore, it seemed reasonable that perhaps carelessness or WTF behaviors in students might be predictable using learner characteristics as student trait features.

Goal Orientation

Achievement goal theory separates mastery goal orientation from performance goal orientation. Individuals with mastery goals (also termed "learning goals") are concerned with increasing their competence. Performance goal orientation occurs when individuals are concerned with gaining favorable judgments of their competence [25]. Achievement goal theory has been refined by separating

performance goals into performance-approach and performance-avoidance goal subcategories [26]. Performance-approach is characterized by a desire to perform exceptionally well, while performance avoidance is characterized by a desire not to perform exceptionally poorly. Performance-approach goals have been shown to be associated with positive academic performance in mathematics [27] and deep learning in English [28]. Performance-avoidance goals have been shown to be associated with negative academic performance in mathematics [27], and surface learning in English [28]. Mastery goal orientation has been shown to be positively correlated with both deep and surface level processing of text passages [29], and both deep and surface learning in English [28]. Finally work-avoidance goal orientation, the goal of doing the smallest amount of work possible, has been shown to be negatively correlated with learning in general as evidenced by final grade in an undergraduate psychology course, and GPA [30].

Disruptive Behavior

A tendency towards disruptive behaviors such as cutting class, being seen as disruptive by others, and being in trouble with law enforcement has been positively correlated with problems in academic performance. Students who reported no discipline problems as sophomores scored approximately 0.2 to 0.8 standard deviation higher on senior grades than those who reported serious behavior problems when grouped by race and gender [31]. Furthermore self-identified disruptiveness was found to be correlated with poorer scores on achievement tests across subjects (Reading, Language, Math, Science, & Social Studies) after controlling for gender, race, and teacher [32].

Self-Efficacy

Self-efficacy is the measure of an individual's belief in their ability to accomplish a given task. Self-efficacy for academic achievement has been shown to positively correlate with student performance [33, 34]. Self-efficacy has also been shown to be positively predictive of performance-approach goal orientation and deeper learning, as well as negatively predictive of performance-avoidance goal orientation [28]. However, it has been hypothesized that high self-efficacy may be positively correlated with carelessness [23], such that overconfidence may lead to carelessness.

Detectors of Affect, Cognition, & Disengagement

Each of the aforementioned disengaged behaviors, i.e., Gaming, Off-Task Behavior, Carelessness, and WTF behaviors, have been detected by computational models developed through data mining [2, 4, 19, 24, 35]. These detectors are meant to approximate human judgment of students' behaviors. In practice human coders judge samples of student actions by looking for particular behaviors. Then an algorithm is applied to a set of features generated from student log files in an attempt to create a set of rules to predict student behaviors (as judged by a human coders). The use of machine-learned detectors to identify student behaviors and different affect states is a fairly recent technological and theoretical development. One notable contribution has been detectors of gaming behaviors, which have achieved Cohen's kappas [36] of 0.36 to 0.4 with human coders given text replay data [37, 38]. Secondly, automated detectors of carelessness also have been employed as a means of

detecting the probability of carelessness or contextual slip. While the metric of success is different as it compares probabilities rather than a finite number of nominal classifications (e.g. Gaming/Not-Gaming), it has achieved success in terms of correlation (r-values ranging from 0.392 to 0.605).

Hypotheses

The aforementioned detectors of disengagement have relied on student actions to detect gaming and carelessness, and while these methods have predictive power, it is the goal of this work to use these predictions to support understanding. This avenue of research has been pursued with respect to gaming the system and off-task behavior in relation to performance goals, anxiety about using the tutoring system or computers in general [39], disliking the tutoring system or computers in general [39, 40], belief that the tutoring system or computers in general are not useful or are uncaring [39, 40], tendencies toward passive-aggressiveness [39, 40], desire for control [39, 40], being self-driven [39, 40] , liking mathematics, belief in the importance of mathematics [39, 40], the utility of the tutor for learning [39], state vs trait beliefs in mathematics ability [40], frustration [40], and anxiety [39, 40]. While there is evidence to suggest that gaming the system has more to do with state variables relating to a specific lesson or subtopic in mathematics rather than trait variables relating to a particular student [39], the construct of educational self-drive appears to be negatively correlated with gaming the system [40], meaning that while it may be more that disengagement (i.e. gaming in a mathematics environment) is driven by state rather than trait variables, there is a notable exception to the contrary. This exception forms a basis for our exploration of learner characteristics as possible predictors of disengagement.

While the scarcity of correlations between trait variables and other constructs of disengagement may seem discouraging, WTF and carelessness remain distinct from other forms of disengagement. The learner characteristics we intend on tracking include a broad group of measures relating to achievement goal theory [25], self-efficacy [33, 34], self-disclipline [41], and several student-centered attributes, such as prior performance in academic contexts. Ideally, these learner characteristics may begin to explain the causes of disengaged behaviors.

Hypothesis 1: Work avoidance goals [42] may be a significant and positive predictor of WTF behavior given that WTF behaviors appear to be examples of “playing with” learning environments to avoid having to engage with the learning task. By contrast, it seems likely that performance approach, or avoidance oriented students, being concerned with performing well, would be less likely to misuse the system in ways that would not improve their assessed performance. Mastery goal orientation might also be negatively correlated with WTF behaviors, as WTF behaviors do not appear to advance students through the activity or foster deeper understanding. Mastery goal orientation relates to a motivation to gain deep understanding of the material, which suggests that the frequency of WTF behaviors would be low. Performance approach and avoidance goal orientations relate to a desire for one’s work to be great (or at least not poor); given these priorities it seems unlikely that a student’s performance would be so haphazard as to verge into WTF behaviors.

Hypothesis 2: WTF behaviors may be positively predicted by disruptive behavior. Since disruptive behaviors are often correlated with poor performance, and we expect WTF to be correlated with poor

performance, it's possible that disruptive behavior and WTF may be correlated. Additionally, WTF is characterized by students not following instructions, i.e., remaining on task in a superficial capacity while performing actions unrelated to the intended purpose of the learning task. If WTF behaviors were to manifest outside of educational software, i.e., during a lecture or other classwork, they might be categorized as disruptive to other students' learning. Investigating this involves checking the correlation between the disruptive behaviors and WTF behaviors, and specifically examining their correlation to determine whether the hypothesized correlation between WTF and a tendency towards disruptive behavior is simply due to their expected mutual relationship with poorer performance.

Hypothesis 3: Carelessness may be a behavior that overconfident students are more likely to engage in [23]. If carelessness occurs more frequently in overconfident students, as suggested by Clements findings "that mathematically competent and confident children… tend to make a greater proportion of careless errors than other children" [23], there may be a link between a self-efficacy and carelessness. We will look for this correlation between self-report measures of self-efficacy and detected instances of carelessness.

Methods

Participants

All participants were eighth graders from three separate public middle schools in Central Massachusetts.

Class A

Participants were comprised of 148 eighth grade students. More than 1/3 of the eighth graders at this school scored “needs improvement” on state standardized science testing in a small town (population 10-15 thousand) in New England, with a median household income between 50 and 60 thousand dollars.

Class B

Participants were comprised of 90 eighth grade students. Over half of the eighth graders at this school scored proficient or above on state standardized science testing in a small city (population 30-40 thousand) in New England, with a median household income between 60 and 70 thousand dollars.

Class C

Participants were comprised of 96 eighth grade students. Over half of the eighth graders at this school scored needs improvement or below on state standardized science testing in a small town (population 5-10 thousand) in New England, with a median household income between 100 and 120 thousand dollars.

Materials

Survey Data

To assess students’ learning and performance goals, the students completed the Patterns of Adaptive Learning Scales (PALS) survey [43] in class, three months prior to using the phase change microworld. In this work, we analyze data from two scales: the first scale measures Personal Achievement Goal Orientation, including learning goal orientation, the goal of developing skill or learning (5 items), performance-approach goal orientation, the goal of demonstrating competence (5 items), and performance-avoid goal orientation, the goal of avoiding demonstrating incompetence (4 items) [44, 45]. The second scale measures Academic-related Perceptions, Beliefs, and Strategies, including academic efficacy [33] (5 items), avoiding novelty [46] (5 items), disruptive behavior [47] (5 items), self-presentation of low achievement (7 items) [48], the desire to prevent peers from knowing how well the student is performing, and skepticism about the relevance of school for future success (6 items). On the PALS survey, each question is given as a 5-point Likert scale, and as is standard, our

assessment of each of the above sub-scales was based on the mean of each student's answers to the items for each sub-scale.

In addition to PALS, students also completed 3 items assessing work-avoidance goals [30, 42] Zimmerman's Self-Efficacy scales [33], and Tangney's Brief Self-Control Scale [41].

Science ASSISTments

The data set used in this work was generated by students using the Science ASSISTments learning environment [7, 49]. Science ASSISTments, now referred to as Inq-ITS, is an interactive computer simulated laboratory, in which students conduct inquiry by observing a phenomenon to be studied, forming hypotheses regarding that phenomenon, then testing their hypotheses in simulations to determine if empirical data does or does not support their hypotheses. Inq-ITS was designed to be an environment that generated performance assessments of students' scientific inquiry skills in terms of warranting claims and communicating findings [9].

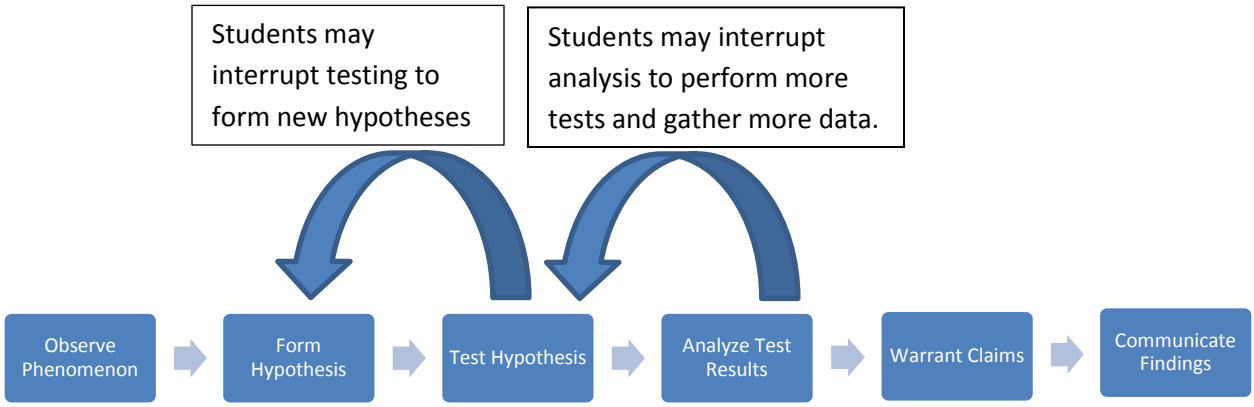

Figure 1 Inq-ITS Microworld Phases

In order to progress through the major phases of an interactive lab exercise, a student must use drop down menus to select variables and values to form their hypothesis, tests, and analysis. The particular microworld exercises that were used in these studies were those of 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 these independent variables may drive the dependent variables of 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.

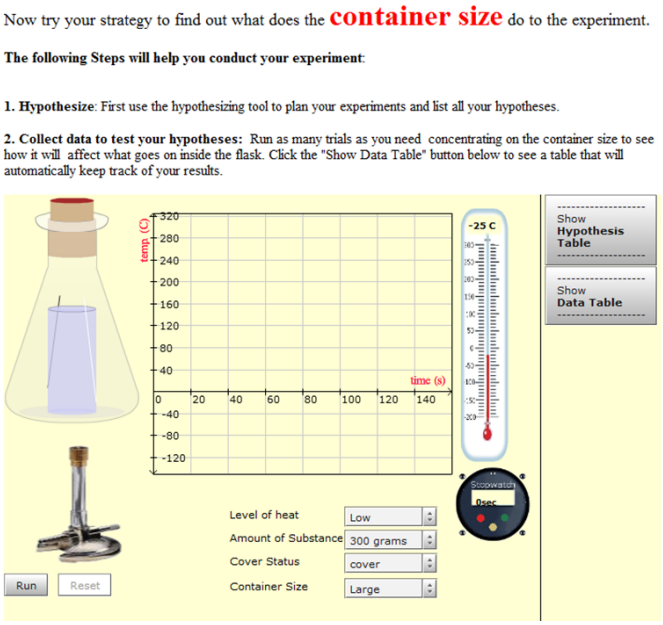

Figure 2 Phase Change Microworld

In the context of Inq-ITS, WTF occurs when students engage with the software in complicated ways that do not appear connected to the task of conducting scientific inquiry. In Inq-ITS, WTF may take the form of running an inordinately large number of identical trials, repeatedly changing most of the variables in a non-systematic fashion, or toggling a variable back and forth repeatedly for no discernible reason.

All students' fine-grained actions were logged and then analyzed at the "clip" level; a clip is a consecutive set of a student's actions describing activity in context. More information on "clips" and "clip" boundaries can be found in the Procedure section.

Procedure

Detector Development

After receiving a short introduction, all students engaged in the phase change learning activities over two class periods, about 1.5 hours in total. During this time, Inq-ITS logged all students' interactions within the learning environment as they engaged in inquiry. In the following sections, we show how we used these low-level interaction data to construct and validate machine-learned detectors of carelessness, and WTF based on existing detectors of systematic data collection behavior [50].

After this point the procedures applied to detect Carelessness (contextual slip) and WTF behaviors diverged. Carelessness is defined by having knowledge of a skill, but failing to perform that skill. In this case there are two skills being assessed with regard to carelessness: a student's skill to test their hypothesis, and a student's skill to control for variables when running an experiment.

Contextual slip refers to a part of Bayesian Knowledge Tracing (BKT) [51], a Hidden Markov Model which treats knowledge as its latent and performance as observable evidence of that hidden latent (see figure 3 below). Knowledge at various points in time is inferred from performance. As students are given more practice opportunities, their performance generally improves, allowing us to infer that the probability they know a skill increases. Figure 3 identifies the interaction between knowledge and performance as modeled in BKT. Knowledge of a skill leads to student performance, however knowing or not knowing a skill does not guarantee performance and vice versa. BKT accounts for four cases: the student is right in a practice opportunity and knows the skill, the student is wrong in a practice opportunity and does not know the skill, the student is right in a practice opportunity but does NOT know the skill, and the student is wrong in a practice opportunity but DOES know the skill. The latter two of these cases are identified as "Guess", and "Slip", respectively. In the context of a particular skill set we have identified slip as carelessness [24]: performing poorly in spite of knowing a skill.

Likewise the probability of knowledge in the future is determined by the probability of prior knowledge and performance data.

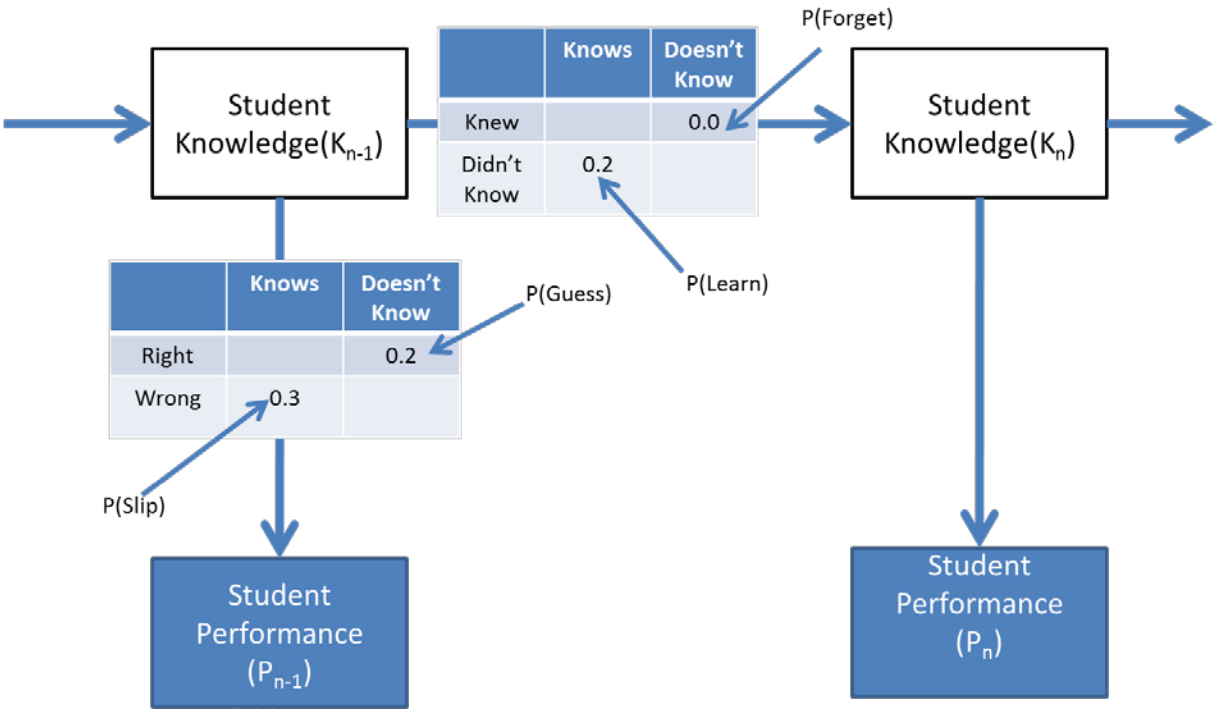

Figure 3 Bayesian Knowledge Tracing (BKT) Model

In this study BKT was used in a four step process for developing a detector of carelessness [19]. First, a BKT model was developed to predict inquiry skills [52]. Second, the best fitting parameters for the BKT model were determined by brute force grid search. Third, clips (a clip is a set of student actions which begins when a student enters the data collection phase and ends when the student leaves that phase) are tagged with a probability of carelessness based on the current probability of knowledge, and the student's performance on the two following clips. Finally, a machine learned detector is built using data only from the current or prior clips to assess carelessness.

The gold standard for WTF behaviors was achieved through text replay coding [50, 53]. Students' log data was segmented by sequences into the same type of clips as described in the last paragraph. In text replays [54], human coders are presented "pretty-printed" versions of log files (as shown in Figure 4). WTF behavior may be difficult to rationally define in log files, but behavior that is completely disconnected from the learning task can be identified by humans relatively easily. In past cases, text replays have proved effective for providing ground truth labels for behaviors of this nature [37, 38, 50]. Examples of WTF behavior in this data set include running the exact same experiment a large number of times (shown in Figure 4), toggling variable settings back and forth repeatedly, and changing large numbers of variables repeatedly. As can be seen, WTF behavior manifests in several ways, an interesting challenge for developing an automated detector of this construct.

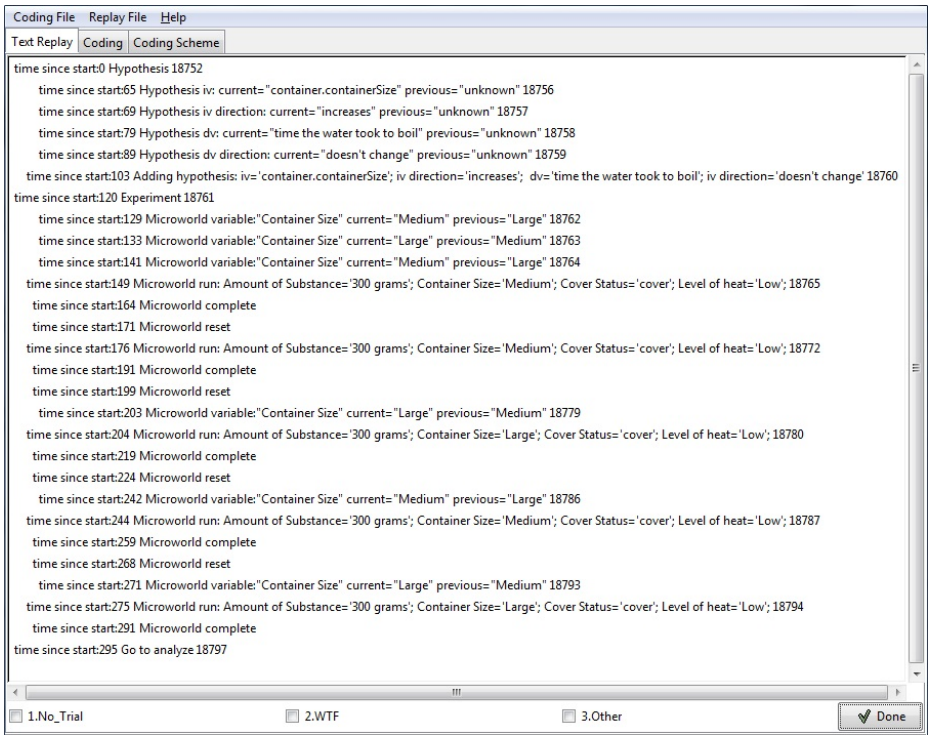

Figure 4. Text Replay Showing Student Running the Same Trial a Large Number of Times

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

Two human coders practiced coding WTF on a small subset of clips, discussing each clip. Afterwards, the two coders separately each coded the same set of 200 clips from a separate data set, not included in further analysis. The two coders achieved acceptable agreement, with Cohen's Kappa

[36] of 0.66. Afterwards, one of the coders assessed 571 clips from this data set. Since several clips could be generated per activity, a single, randomly chosen clip was tagged per student, per activity (however, not all students completed all activities, causing some student-activity pairs to be missing from the data set). This ensured all students and activities were approximately equally represented in this data set. Seventy of these clips were excluded from analysis, due to a lack of data collection actions on the student's part. Of the 501 clips remaining, 15 (3.0%) were labeled as involving WTF behavior, a proportion similar to the proportions of other disengaged behavior studied in past detector development [37]. These 15 clips were drawn from 15 (10.4% of the sample population) of the students (i.e., no student was coded as engaging in WTF behavior more than once).

After labels of ground truth were reached for both Carelessness and WTF behavior, features were extracted for the machine-learning process. In brief, the features used included the numbers of different types of actions that occurred during the clip (including the number of complete and student-interrupted trials and the number of variable changes made while designing each experiment), the timing of each action (including the average time per variable change and the maximum time the student spent studying the simulation, and many others). After this point the distilled features were fed into a data mining algorithm along with the ground truth labels in order to generate predictions of those ground truth labels.

Learner Characteristic Analyses

Once the detectors of WTF and Carelessness were completed the next step was to look for relationships between these constructs and goal orientation or learner characteristics. Constructs of goal orientation and learner characteristics were established by measurement through self-report. Generally using self-response Likert Scale surveys i.e. Patterns of Adaptive Learning Scales (PALS) [43], the Tangney Brief Self-Control Scale [41], Zimmerman's Self-Efficacy scales [33], and Harackiewicz's measures of Work Avoidance [42]. By using PALS, alongside Tangney's, Zimmerman's, and Harackiewicz's surveys we were able to achieve reasonable measures of these constructs.

In order to compare these measures against the detected constructs of WTF and Carelessness, simple bivariate correlations using SPSS [55] were conducted at the student level. These analyses were performed across the student sample group "Class A" which the detectors were built upon and two additional student sample groups "Classes B & C". Following these analyses, cluster analyses were performed on these data sets as well to see if students who typically engage in WTF or careless behaviors fit with a particular combination of other attributes determined through goal orientation and learner characteristic surveys [19].

Cluster analysis is a data mining technique by which instances (in this case individual students) are sorted into clusters based on their degree of similarity across multiple features. In this case, measures gained through learner characteristic surveys would serve as our features.

The clustering approach used here is k-means clustering [56]. K-means clustering uses the following process. First, each data point is plotted in n-dimensional space where each feature acts as a dimension. For example, a data set which includes features: age, height, and weight could be plotted in

three dimensional space, where age would be plotted on the x-axis, height on the y-axis, and weight on the z-axis. Second, a number of centroids "k" are plotted in the multi-dimensional space of the total data set. After plotting these points, "k" centroids are randomly assigned to be the centers of "k" clusters. Points are assigned to each centroid based on the ordinary Euclidian distance metric, then the means of the dimensions of each point in a cluster is calculated and the centroid is moved to sit in that new centroid as defined by all points in the cluster. Points are reassigned to their new respective centroids and the process continues until the cluster assignment then stabilizes.

Results

Carelessness (Contextual Slip)

Using an existing set of probabilities of carelessness for Class A, a machine-learned detector of carelessness was built using distilled features of student activities in the microworld. The algorithm of this detector was built using W-REPTree [56], a regression tree available from Weka through RapidMiner [57]. The model was built and evaluated with six-fold batch cross validation at the student level [58] and achieved a correlation of 0.62 and an RMSE of 0.16. The process of student-level cross validation validates whether the model is overfit for the particular group of students used as the sample. In the process of six-fold student level batch cross validation, 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" test set of students. By cross-validating at this level, we increase confidence that detectors will be accurate for new groups of students.

The data was then examined for correlations between carelessness and the aforementioned learner characteristics. There were marginally significant correlations between careless errors and disruptive behavior $r=-0.15$, $F(1,128)=3.02$, $p=0.08$ and careless errors and self-presentation of low achievements $r=-0.16$, $F(1,128)=3.45$, $p=0.07$.

Self-presentation of low achievement is a measure of students' concern that high academic performance will result in negative social repercussions from classmates. The negative correlations between self-presentation of low achievement and carelessness, and disruptive behavior and carelessness were not addressed in our initial hypotheses.

In the original publication of these results [19], the initial hypothesis was that performance goal orientations would be positively correlated with carelessness, while mastery goal orientations will be negatively correlated with carelessness. While this hypothesis was not directly confirmed, cluster analysis produced an interesting result.

Table 1 Cluster Analysis of Carelessness and Learner Characteristics for Class A

Variable	Mean (std)		
	Cluster 1	**Cluster 2**	**Cluster 3**
Learning goal orientation	4.66 (0.40)	4.38 (0.64)	2.07 (0.87)
Performance-approach goal orientation	1.69 (0.57)	3.20 (1.04)	2.40 (0.82)
Performance-avoid goal orientation	1.86 (0.72)	3.78 (0.67)	3.62 (0.68)
Academic efficacy	4.41 (0.49)	4.22 (0.55)	3.65 (1.06)
Avoiding novelty	1.96 (0.60)	2.58 (1.00)	3.02 (1.21)
Disruptive behavior	1.54 (0.68)	1.61 (0.68)	2.07 (1.01)
Self-presentation of low achievement	1.33 (0.31)	1.59 (0.60)	3.43 (1.00)
Skepticism about the relevant of school for future success	1.57 (0.49)	1.92 (0.82)	2.07 (0.87)
N	**35**	**66**	**20**
Carelessness	**0.16 (0.22)**	**0.12 (0.13)**	**0.05 (0.05)**

It was shown that when the sample was broken into three clusters based on the most prevalent commonalities between students in their survey responses, the clusters produced included a mastery (learning goal) oriented cluster, a performance oriented cluster, and a cluster with neither mastery nor performance goal orientation. Here the students with neither performance nor mastery goal orientations scored the lowest in terms of careless errors.

Following these analyses new learner characteristics were tracked including: Self-Control [41], Work Avoidance Goal Orientation [42], and Self-Efficacy [33]. In the prior data set carelessness was measured with respect to the skills Control of Variables Strategy and Test Hypothesis; correctness in terms of these skills was assigned through a detector of inquiry skills [50]. Unfortunately this detector had only been applied to data from a subset of Class B, and carelessness in the skills previously measured could not be applied to Class C or the total set of Class B.

Table 2 Carelessness & Learner Characteristic Correlations for as subset of Class B

Carelessness	PALS1 Mastery	PALS1 Performance Approach	PALS1 Performance Avoid	PALS4 Academic Efficacy	PALS4 Novelty Avoid	PALS4 Disruptive Behavior	PALS4 Self Presentation of Low Achievement	PALS4 Skeptical of School's Relevance	Tangney Brief SelfControl Scale	Work Avoid Harackiewicz	Self-Efficacy Zimmerman	WTF
Pearson Correlation	-0.294	-0.033	-0.154	0.072	0.116	0.024	0.134	0.124	-0.119	0.081	0.019	-0.151
Sig. (2-tailed)	0.045	0.825	0.302	0.632	0.438	0.871	0.370	0.407	0.424	0.588	0.900	0.311
N	47	47	47	47	47	47	47	47	47	47	47	47

Table 2 contains the correlations between carelessness and all measured learner characteristics in a subset of class B addition to WTF. Carelessness of individual clips were averaged to create a carelessness rating at the student level. There was one significant correlation: carelessness is negatively correlated with mastery goal orientation. With regard to hypotheses 3 and 4, neither Self-Efficacy nor Self-Presentation of Low Achievement significantly correlated with carelessness.

Following this, we applied k-means cluster analysis again, except in this case we had access to survey data including additional measures: Self-Control, Work Avoidance Goals, and Self-Efficacy.

Table 3 Cluster Analysis of Carelessness and Learner Characteristics for Class B

Cluster	Full Data	Cluster 0	Cluster 1	Cluster 2
Sample Size (N)	47	22	6	19
	Mean Scores			
PALS1 Mastery Goal Orientation	22.8936	23.5455	24	21.7895
PALS1 Performance Approach Goal Orientation	13.3404	10.5455	19.6667	14.5789
PALS1 Performance Avoidance Goal Orientation	11.7872	8.7273	16.1667	13.9474
PALS4 Academic Efficacy	19.1064	20.2273	19.5	17.6842
PALS4 Novelty Avoidance	12.5106	9.9091	8.8333	16.6842
PALS4 Disruptive Behavior	8.6809	7.6818	5.3333	10.8947
PALS4 Self-Presentation of Low Achievement	12.4681	9.8636	10.8333	16
PALS4 Skeptical of School Relevance	12.383	9.5	9.1667	16.7368
Tangney Brief Self-Control Scale	45.3617	49.5455	51.6667	38.5263
Work Avoidance Goal Orientation Harackiewicz	9.5106	7.4545	7.1667	12.6316
Self-Efficacy Scale Zimmerman	42	45.3636	45.8333	36.8947
Carelessness	0.280953	0.284990	0.174395	0.309929

Information about carelessness was held out of the data set when cluster analysis was applied. In order to easily display what differences were statistically significant and which ones were not the following color coding scheme was adopted: if two cells are white text on a black background then differences between them are not significantly different, if two cells are on a gray background then differences between them are marginally significant $p<0.1$, if two cells are black text on a white

background then differences between them are significantly different $p<0.05$. Finally, if only one cell in row has a background of a particular shade it has that significance in relation to all other cells in that row, e.g. cluster 1's academic efficacy is not significantly different from clusters 0 or 2 and cluster 1's carelessness is marginally significantly different from clusters 0 and 2.

Cluster 0 has the lowest rates of performance goal orientation as well as low: self-presentation of low achievement, novelty avoidance, skepticism of school's relevance, and work avoidance goals. It tends to have higher rates of: mastery goal orientation, self-control, and self-efficacy.

Cluster 1 has the highest rate of performance approach goal orientation and the lowest rate of disruptive behavior. It tends to have higher rates of: mastery goal orientation, self-control, and self-efficacy and lower rates of: novelty avoidance, self-presentation of low achievement, skepticism about school's relevance, and work avoidance goals.

Cluster 2 has the lowest rates of: mastery goal orientation, self-control, and self-efficacy. Also it had the highest rates of: novelty avoidance, disruptive behavior, self-presentation of low achievement, skepticism of school's relevance and work-avoidance.

There were several non-significant differences between the clusters, possibly due to the much smaller sample size N=47 as opposed to the prior sample size of N=121. In spite of this, cluster 1 was roughly half as careless as the other two clusters. P-values for differences between clusters in terms of carelessness were as follows: cluster 0 vs 1 $p=0.105$, cluster 0 vs 2 $p=0.761$, cluster 1 vs 2 $p=0.041$.

These findings run counter to the earlier cluster analysis wherein the least careless cluster had the lowest ratings of performance and learning goal orientations. One possible explanation for this is the small sample size, while cluster 1 has a much lower incidence of carelessness it's made up of only 6 people.

WTF

Attempts were made to fit detectors of WTF using 11 common classification algorithms, including Naïve Bayes, and J48 decision trees. The best model performance was achieved by the PART algorithm [59], an algorithm that produces rules out of C4.5 decision trees (essentially the same algorithm as J48 decision trees). The implementation of PART from WEKA [56] was run within RapidMiner 4.6 [57]. In this algorithm, a set of rules is built by repeatedly building a decision tree and making a rule out of the path leading to the best leaf node at each iteration. PART has not been frequently used in student modeling, but was used in one instance to predict student course success [60]. These models were evaluated with the aforementioned six fold student level cross-validation process [58].

The validity of our detectors was assessed using four commonly used metrics, A' [61], Kappa [36], precision [62], and recall [62]. A' is the probability that the detector will be able to distinguish a clip involving WTF behavior from a clip that does not involve WTF behavior. A' is equivalent to both the area under the ROC (receiver operating characteristic) curve which plots true positives on the y-axis and false

positives on the x-axis in signal detection theory and to W, the Wilcoxon statistic [61]. A model with an A' of 0.5 performs at chance, and a model with an A' of 1.0 performs perfectly. An appropriate statistical test for A' in data across students would be to calculate A' and standard error for each student for each model, compare using Z tests, and then aggregate across students using Stouffer's method. However, the standard error formula for A' [61] requires multiple examples from each category for each student, which is infeasible in the small samples obtained for each student in our data labeling procedure. Another possible method, ignoring student-level differences to increase example counts, biases undesirably in favor of statistical significance. Hence, statistical tests for A' are not presented in this work.

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

A' and Kappa were chosen because they compensate for successful classifications occurring by chance [63], an important consideration in data sets with unbalanced proportions of categories (such as this case, where WTF is observed 3.0% of the time). Precision and Recall give an indication of the detector's balance between two forms of error. It is worth noting that unlike Kappa, Precision, and Recall, which only look at the final label, A' takes detector confidence into account.

The detector of WTF behavior achieved good performance under 6-fold student-level cross-validation as shown in Table 4. The detector achieved a very high A' of 0.8005, signifying that it could distinguish whether or not a clip involved WTF behavior approximately 80.05% of the time. However, 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 performed 41.1% better than chance. This level of Kappa is comparable to a detector of gaming the system effectively used in interventions [37]. Kappa values in this range, combined with relatively high A' values, suggest that the detector is generally good at recognizing which behavior is more likely to be "WTF", but classifies many 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 = 41.18%, and recall = 50%. As such, it is important to use fail-soft interventions and to take detector certainty into account when selecting interventions – but there is not evidence that the detector has strong bias either in favor of or against detecting WTF behavior.

Table 4. WTF Detector Confusion Matrix

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

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

1) IF the total number of independent variable changes (feature 21) is seven or lower, AND the number of experimental trials run (feature 7) is three or lower, THEN **NOT WTF**.
2) IF the maximum time spent between an incomplete run and the action preceding it (feature 16) is 10 seconds or less, AND the total number of independent variable changes (feature 21) is eleven or less, AND the average time spent paused (feature 5) is 6 seconds or less, THEN **NOT WTF**.
3) IF the total number of independent variable changes (feature 21) is greater than one, AND the maximum time between actions (feature 3) is 441 seconds or less, AND the number of trials run without pauses or resets (feature 12) is 4 or less, THEN **NOT WTF**.
4) IF the total number of independent variable changes (feature 21) is 12 or less, THEN **WTF**.
5) IF the maximum time spent before running each experimental trial but after performing the previous action (feature 11) is greater than 1.8 seconds, THEN **NOT WTF**.
6) All remaining instances are classified as **WTF**.

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

After the detector was built the results were compared to learner characteristic and goal orientation survey responses. The data set which the detector was trained on did not have complete survey responses associated with it. This data set only included responses to PALS 1 & 4, which did not include Harakiewicz's work avoidance items [42], Tangney's Self-Control items [41], or Zimmerman's Self-Efficacy survey [33].

WTF & Learner Characteristics: Correlations

Our goal in exploring the relationship between WTF behaviors and learner characteristics is to have a better understanding of what long term characteristics may relate to WTF in much the same way that we approached carelessness and related learner characteristics.

We applied our detector of WTF behaviors to data from classes from classes B & C. All of these students worked within the phase change microworld, the same exercise students from class A had worked within for purposes of building the WTF detector. Student learner characteristic surveys were administered to all participants. The set of students used to build the WTF detector were only administered PALS [43] 1 and 4 which include Mastery Goal Orientation, Performance Approach Goal Orientation, Performance Avoidance Goal Orientation, Academic Efficacy, Novelty Avoidance, Disruptive Behavior, Self-Presentation of Low Achievement, and Skepticism About School's Relevance. Classes B & C were also administered the Tangney Brief Self-Control Scale [41], Harackiewicz Work Avoidance Scale [42], and Zimmerman's Self-Efficacy Scale [33]. The percentages of students who were detected engaging in WTF behavior out of the total sample for each class ranged from 7% for class C to 10% for class A. Each class was tested for bivariate correlations between each learner characteristic measure and WTF. In this case, probability of WTF, rather than the nominal value of WTF or Not WTF was used to create a numeric measure of WTF. WTF was aggregated at the student level rather than clip level to compare with learner characteristics which are also aggregated at the student level, to achieve this aggregation probabilities of WTF were averaged across clips to give an overall probability each student was engaging in WTF behaviors at any given time. Tables of bivariate correlations for each class are supplied in the appendix; several measures were significantly correlated with one another, and no pair of measures were significantly positively correlated with one another in one class but significantly negatively correlated with one another in another class.

WTF was the only measure that was not significantly correlated with any individual learner characteristic measure in class A, B, or C (see Appendix for tables). However, in class B a correlation between WTF and academic efficacy was marginally significant at $p=0.059$ with an effect size $r= -0.2$. This correlation was not marginally significant in either classes A or C at $p=0.543$ and $p=0.848$ respectively. As a result it seems reasonable that this correlation may be a characteristic of class B , rather than descriptive of larger populations. Subsequently, all classes were merged to see if the increased sample size would produce any significant correlations, only the old measures from PALS 1 & 4 were used here as they were the only learner characteristic measures common across classes. Once again, WTF was the only measure that was not significantly correlated with any other individual measure (see Appendix for table). In this final merged sample academic efficacy and WTF were not significantly correlated at $p=0.258$ with a smaller effect size of $r= -0.063$.

WTF & Learner Characteristics: Cluster Analysis

While no single learner characteristic was correlated with WTF behaviors, it was possible that some combination of other learner characteristics might be related to WTF behaviors. In order to investigate this, we applied k-means cluster analysis to a combined group of classes B and C using all available learner characteristic measures.

Table 5 Cluster Analysis of WTF and Learner Characteristics for Classes B & C New Measures

Cluster	Full Data	Cluster 0	Cluster 1	Cluster 2
Sample Size (N)	186	64	57	65
	Mean Scores			
PALS1 Mastery Goal Orientation	20.3925	22.0938	22.1404	17.1846
PALS1 Performance Approach Goal Orientation	12.5914	9.8438	17.7018	10.8154
PALS1 Performance Avoidance Goal Orientation	11.7312	8.8281	16.0877	10.7692
PALS4 Academic Efficacy	18.6304	20.6563	20.1339	15.3174
PALS4 Novelty Avoidance	13.25	9.3906	13.7237	16.6346
PALS4 Disruptive Behavior	8.2011	6.9219	8.0912	9.5569
PALS4 Self-Presentation of Low Achievement	11.5082	9.1406	12.5266	12.9463
PALS4 Skeptical of School Relevance	14.1538	10.4531	13.3063	18.5408
Tangney Brief Self-Control Scale	46.1808	50.5938	46.8053	41.288
Work Avoidance Goal Orientation Harackiewicz	10.3714	7.5781	10.0501	13.4035
Self-Efficacy Scale Zimmerman	41.3886	45.6563	42.5078	36.2051
WTF Probability	0.0225	0.0172	0.0150	0.0343
WTF Prediction Nominal	0.0753	0.0781	0.0351	0.1077

Information about WTF behavior was not included in the data set when cluster analysis was applied. There were several cases where measures were not significantly different from one another. Significance has been denoted in the same way as described for table 3 in the Carelessness Results section. If all cells in a row are either white or black, then differences in that particular measure are either significant or not significant with respect to one another. If a single cell in a row is different from the other two then the color of that cell denotes its relationship to all other cells in that row. For example: in terms of academic efficacy cluster 0 and 2 are significantly different from one another, but cluster 1 is not significantly different from either clusters 0 or 2, the same is true for disruptive behavior.

Students in cluster 0 had significantly the lowest: Performance Avoidance, Novelty Avoidance, Self-Presentation of Low Achievement, and Skepticism about School's Relevance. They also had the highest Self-Control and Self-Efficacy.

Students in cluster 1 had significantly the highest ratings of Performance Approach and Avoidance goal orientation.

Students in cluster 2 had significantly the lowest: Mastery Goal Orientation, Academic Efficacy, Self-Control, and Self-Efficacy. They also had the highest ratings of Novelty Avoidance, and Skepticism of School's Relevance.

Unfortunately, neither the probability of WTF, nor the categorical nominal predictions of WTF were significantly different across any of the three clusters. P-values for differences between clusters in terms of the probability of WTF were as follows: cluster 0 vs 1 p=0.784, cluster 0 vs 2 p=0.202, cluster 1 vs 2 p=0.153. P-values for differences between clusters in terms of nominal/categorical predictions of WTF were as follows: cluster 0 vs 1 p=0.315, cluster 0 vs 2 p=0.567, cluster 1 vs 2 p=0.128. The differences between clusters 1 and 2 were nearly marginally significant, suggesting that perhaps with a larger sample size a significant difference might be obtained. To increase sample size we combined classes A, B and C using only the learner characteristic measures for PALS 1 & 4 which had been administered to all classes. The hope here was that a larger sample size might increase some of the differences between clusters in terms of WTF.

Table 6 Cluster Analysis of WTF & Learner Characteristics for Classes A, B, & C Merged Data Set

Cluster	Full Data	Cluster 0	Cluster 1	Cluster 2
Sample Size (N)	330	115	127	88
	Mean Scores			
PALS1 Mastery Goal Orientation	20.6844	22.3364	22.2242	16.3032
PALS1 Performance Approach Goal Orientation	12.6281	9.1881	17.0789	10.7003
PALS1 Performance Avoidance Goal Orientation	11.8	8.6504	15.4126	10.7023
PALS4 Academic Efficacy	19.6	20.9565	20.6803	16.2682
PALS4 Novelty Avoidance	13.0906	9.7043	13.3214	17.1828
PALS4 Disruptive Behavior	8.5906	7.0261	8.3174	11.0294
PALS4 Self-Presentation of Low Achievement	11.442	9.2783	11.9088	13.5959
PALS4 Skeptical of School Relevance	13.3459	10.1739	12.2662	19.0494
WTF Probability	0.0244	0.0230	0.0211	0.0310
WTF Prediction Nominal	0.0818	0.0870	0.0551	0.1136

The new clusters (shown in table 6) can be described in much the same way, which is unsurprising since most of the data in this data set is the same. WTF is still not significantly different across any of these three clusters. P-values for differences between clusters in terms of nominal/categorical predictions of WTF were as follows: cluster 0 vs 1 p=0.335, cluster 0 vs 2 p=0.530, cluster 1 vs 2 p=0.119. P-values for differences between clusters in terms of probability based predictions of WTF were as follows: cluster 0 vs 1 p=0.828, cluster 0 vs 2 p=0.395, cluster 1 vs 2 p=0.320. In this case, significance has dropped given an increase in sample size, suggesting that lack of significance was not due to sample size. These findings are consistent with the lack of correlations of any learner characteristic with WTF behavior. They do not support a relationship between WTF and the listed learner characteristics when these characteristics are combined, in a similar way to how the correlations shown in the appendix do not indicate a relationship between WTF and any individual learner characteristic.

Additionally, the factors which seem to be related through cluster analysis also seem to be related through individual correlations. For example, in table 5, cluster 0 contains the lowest Performance Avoidance, Novelty Avoidance, Self-Presentation of Low Achievement, and Skepticism of School's Relevance. If we look at the table of Total Correlations in the appendix we can see that Performance Avoidance is positively correlated with Novelty Avoidance and Self-Presentation of Low Achievement, Novelty Avoidance is positively correlated with the aforementioned as well as Self-Presentation on Low Achievement and Skepticism of School's Relevance, and Self-Presentation of Low Achievement is positively correlated with the aforementioned as well as Skepticism of School's Relevance.

The cluster analysis findings support the correlation analysis findings, this support comes in the form of the aforementioned results and in prior paragraph, as well as the null results with regard to a relationship between WTF behavior and the measured learner characteristics.

WTF & Learner Characteristics: Decision Tree Rule Learner

The methods we have employed to find a relationship between WTF and Learner Characteristics so far have largely been conducted in order to determine what characteristic, or characteristics, relate positively or negatively with WTF behavior. However, using decision tree rule learners we can attempt to build special conditional rules that predict WTF behavior, from learner characteristics. In this case our primary goal is not to improve detection of WTF, but rather to gain a better understanding of what types of students are more likely to engage in WTF behaviors. Perhaps the relationship between learner characteristics and WTF might not be so simple as "more or less of a given learner characteristic implies a greater or lesser propensity for WTF behavior". Rather, it may be the case that moderate or divergent extremes of certain learner characteristics may imply a greater or lesser propensity for WTF behavior, much in the same way that in our original detector of WTF behaviors both extremes of high and low totals of independent variable changes in a clip implied WTF behavior, while moderate amounts suggested not WTF.

Ideally, a detector could be built for each data set and then the rules for determining WTF behavior could be compared. Unfortunately, the detector built based on the first data set under 6-fold cross validation performed non-satisfactorily (Cohen's kappa <0). Even building a detector without 6-fold cross validation, instead using the training set as the test set generates a detector that classifies all instances as "not WTF". So it was necessary to merge all data sets to produce results in the same way that data sets were merged for cluster analyses.

The merged data set of measures (PALS1 & PALS4) which included data from classes A, B, & C also generated a detector that classified all instances as "not WTF" under 6-fold cross validation, the same problem occurred outside of cross validation as well.

These results are not informative in the way that we had intended: they told us nothing about how WTF behavior relates to learner characteristics in terms of conditional rules. However, they support the finding that these learner characteristics are, at best, weakly related to WTF behavior.

Discussion

We demonstrated that WTF could be identified by human coders and that an automated detector of WTF could be successfully built and utilized. Furthermore, we investigated the possible relationship between learner characteristics and WTF behaviors by attempting to create models of WTF behaviors using student learner characteristics as features to train our detectors.

Examining the model of WTF behavior obtained provides some interesting implications about this type of behavior. Previous detectors of undesirable behavior have largely focused on identifying the specific undesirable behavior studied [37, 38, 64]. By contrast, the rules produced by the WTF detector are targeted more towards identifying what is not WTF behavior rather than identifying what is WTF behavior. Four of the six rules identify non-WTF behavior. Of the two rules identifying WTF behavior, one simply states that any behavior not captured by the first five rules can be considered WTF. As such, this model suggests that WTF behavior may be characterized by the absence of appropriate strategies and behaviors, in a student actively using the software, rather than specific undesirable behavior.

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

Overall the preponderance of results on the relationship between carelessness or WTF and learner characteristics were null results. Findings regarding carelessness do not support earlier findings; in the case of cluster analysis they appear to contradict earlier findings. Given that the new findings are based on a partial sample of Class B favor should be given to the earlier findings of Class A, which is much larger. The most significant findings regarding WTF approach marginal significance.

The lack of correlation between WTF and learner characteristics bears a resemblance to the lack of trait variables relationship to gaming the system [39]: perhaps WTF behaviors are more dependent on state rather than trait. Learner characteristics used here are entirely trait based.

It may be reasonable then to look at non-trait based signs of disengagement focusing instead on individual differences and contextual factors that may lead students to engage in WTF behavior. This behavior could be expected to emerge for several reasons, including attitudinal reasons such as not valuing the learning task [65], or immediate affective states such as confusion, frustration, and boredom [66]. A key first paper investigating this question is Sabourin et al. [11], which showed that when WTF

behavior (termed off-task behavior) emerges among students displaying different affect, it has different implications about their affect later in the task. Specifically, students who engage in this behavior when they are confused later become bored or frustrated. By contrast, students who engage in this behavior when they are frustrated often become re-engaged. These findings suggest that intelligent tutors should offer different interventions, depending on the affective context of WTF behavior, but further research is needed to determine which strategies are most appropriate and effective for specific learning situations and for learners with specific characteristics. For example, a confused student engaging in WTF behavior may need additional support in understanding how to learn from the learning environment [67]. By contrast, a student who engages in WTF behavior due to boredom or because they do not value the learning task may require intervention targeted towards demonstrating the long-term value of the task for the student's goals [68].

Automated detectors such as the one presented here have a substantial role to play in understanding the causes of WTF behavior. In specific, these detectors will make it feasible to study WTF behavior across a greater number of situations [cf. 69], helping us to better understand the factors leading to WTF behavior. By understanding the causes of WTF behavior, and how learning software should respond to it, we can take another step towards developing learning software that can effectively adapt to the full range of students' interaction choices during learning [70].

Future Work

Beyond the scope of this master's thesis, there are several research opportunities for the future. The practical application of this research will hopefully benefit the students using Inq-ITS by identifying students who engage in WTF behaviors or careless errors and in turn responding in real time to get them back on track as necessary. Furthermore, differentiating between forms of disengagement should allow for different and appropriate methods for scaffolding students. Again, disengaged behaviors do not always mean reduced learning gains [cf. 4, 35] and it has been posited that WTF-like behaviors (described as Off-Task) may serve as a self-regulation strategy for some students by allowing them breaks in study [10, 11]. Hence, modifying Inq-ITS to appropriately respond to these behaviors will depend upon understanding them.

It may be possible to use association rule or sequence mining approaches applied to these data to replicate Sabourin's finding [11] that students who performed WTF behaviors (termed Off-Task) while frustrated or confused are more likely to become re-engaged. By looking at the actions which directly precede or follow WTF behaviors along with the observed or detected affect of students we may pursue two different new research opportunities. Firstly, we may determine better whether WTF behaviors help or harm students learning strategies. Second, we may be able to use these preceding and following actions, or affect data a means of adding features to improve our WTF detector.

One learner characteristic we had intend on studying in relation to carelessness and WTF behaviors is grit as identified by Duckworth[71].Grit [71] might be negatively correlated with carelessness. If carelessness is driven by overconfidence on the part of the student, then perhaps a construct characterized by effort in spite of negative performance is also indicative of effort in spite of

positive performance. Alternately if carelessness is the product of boredom when confronted with familiar material, i.e. “high grit” students may continue to perform reliably in spite of boredom. We are currently collecting data using the short Duckworth Grit scale including simpler language intended for children [71].

The work to date has been based on a detector trained on a single data set. Further, while the kappa found in inter-rater reliability is impressive at 0.66 this is based on 2 agreed WTF positive clips, 135 agreed WTF negative clips and 2 disagreements. The construct validity of WTF should be tested by having additional coders code clips for WTF behaviors. Ideally, these coders would come from independent communities and they would code clips in new domains in environments other than Inq-ITS or SLINQ’s inquiry microworlds. This research has been a reasonable start to exploring WTF as a construct, but the WTF detector’s validity and further its utility in terms of learning and affect must be further tested.

Appendix

Class A Correlations

		PALS1 Mastery	PALS1 Performance Approach	PALS1 Performance Avoid	PALS4 Academic Efficacy	PALS4 Novelty Avoid	PALS4 Disruptive Behavior	PALS4 Self Presentation Low Achievement	PALS4 Skeptical School Relevance	WTF
PALS1 Mastery	Pearson Correlation	1	.250**	.180*	.524**	-.388**	-.380**	-.258**	-.470**	-.090
	Sig. (2-tailed)		.004	.038	.000	.000	.000	.003	.000	.302
	N	134	134	134	134	134	134	134	134	134
PALS1 Performance Approach	Pearson Correlation	.250**	1	.565**	.150	.021	.036	-.165	-.030	-.093
	Sig. (2-tailed)	.004		.000	.083	.809	.681	.057	.735	.283
	N	134	134	134	134	134	134	134	134	134
PALS1 Performance Avoid	Pearson Correlation	.180*	.565**	1	.089	-.002	-.175*	.077	-.020	.076
	Sig. (2-tailed)	.038	.000		.306	.979	.043	.374	.818	.385
	N	134	134	134	134	134	134	134	134	134
PALS4 Academic Efficacy	Pearson Correlation	.524**	.150	.089	1	-.385**	-.321**	-.219*	-.270**	-.053
	Sig. (2-tailed)	.000	.083	.306		.000	.000	.011	.002	.543
	N	134	134	134	136	136	136	136	136	136
PALS4 Novelty Avoid	Pearson Correlation	-.388**	.021	-.002	-.385**	1	.379**	.184*	.281**	.087
	Sig. (2-tailed)	.000	.809	.979	.000		.000	.032	.001	.314
	N	134	134	134	136	136	136	136	136	136
PALS4 Disruptive Behavior	Pearson Correlation	-.380**	.036	-.175*	-.321**	.379**	1	.299**	.354**	.004
	Sig. (2-tailed)	.000	.681	.043	.000	.000		.000	.000	.961
	N	134	134	134	136	136	136	136	136	136
PALS4 Self Presentation Low Achievement	Pearson Correlation	-.258**	-.165	.077	-.219*	.184*	.299**	1	.235**	.008
	Sig. (2-tailed)	.003	.057	.374	.011	.032	.000		.006	.928
	N	134	134	134	136	136	136	136	136	136
PALS4 Skeptical School Relevance	Pearson Correlation	-.470**	-.030	-.020	-.270**	.281**	.354**	.235**	1	-.035
	Sig. (2-tailed)	.000	.735	.818	.002	.001	.000	.006		.689
	N	134	134	134	136	136	136	136	136	136
WTF	Pearson Correlation	-.090	-.093	.076	-.053	.087	.004	.008	-.035	1
	Sig. (2-tailed)	.302	.283	.385	.543	.314	.961	.928	.689	
	N	134	134	134	136	136	136	136	136	144

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Class B Correlations

		PALS1 Mastery	PALS1 Performance Approach	PALS1 Performance Avoid	PALS4 Academic Efficacy	PALS4 Novelty Avoid	PALS4 Disruptive Behavior	PALS4 Self Presentation Low Achievement	PALS4 Skeptical School Relevance	Tangney Brief Self Control Scale	Work Avoidance Harackiewicz	Self Efficacy Zimmerman	WTF
PALS1 Mastery	Pearson Correlation	1	.173	.024	.141	-.032	-.080	.045	-.207	.230*	-.230*	.242*	.009
	Sig. (2-tailed)		.103	.822	.185	.761	.452	.674	.050	.029	.029	.022	.936
	N	90	90	90	90	90	90	90	90	90	90	90	90
PALS1 Performance Approach	Pearson Correlation	.173	1	.571**	.058	.257*	.087	.223*	.197	.000	.089	-.058	-.032
	Sig. (2-tailed)	.103		.000	.588	.015	.416	.035	.063	.997	.406	.584	.763
	N	90	90	90	90	90	90	90	90	90	90	90	90
PALS1 Performance Avoid	Pearson Correlation	.024	.571**	1	-.017	.240*	.109	.320**	.264*	-.132	.317**	-.177	.069
	Sig. (2-tailed)	.822	.000		.875	.023	.304	.002	.012	.216	.002	.095	.517
	N	90	90	90	90	90	90	90	90	90	90	90	90
PALS4 Academic Efficacy	Pearson Correlation	.141	.058	-.017	1	-.201	-.026	-.061	-.116	.405**	-.466**	.416**	-.200
	Sig. (2-tailed)	.185	.588	.875		.058	.811	.567	.275	.000	.000	.000	.059
	N	90	90	90	90	90	90	90	90	90	90	90	90
PALS4 Novelty Avoid	Pearson Correlation	-.032	.257*	.240*	-.201	1	.309**	.499**	.356**	-.436**	.460**	-.326**	.072
	Sig. (2-tailed)	.761	.015	.023	.058		.003	.000	.001	.000	.000	.002	.502
	N	90	90	90	90	90	90	90	90	90	90	90	90
PALS4 Disruptive Behavior	Pearson Correlation	-.080	.087	.109	-.026	.309**	1	.386**	.111	-.294**	.110	-.264*	-.032
	Sig. (2-tailed)	.452	.416	.304	.811	.003		.000	.300	.005	.300	.012	.761
	N	90	90	90	90	90	90	90	90	90	90	90	90
PALS4 Self Presentation Low Achievement	Pearson Correlation	.045	.223*	.320**	-.061	.499**	.386**	1	.415**	-.287**	.313**	-.184	-.006
	Sig. (2-tailed)	.674	.035	.002	.567	.000	.000		.000	.006	.003	.082	.953
	N	90	90	90	90	90	90	90	90	90	90	90	90
PALS4 Skeptic School Relevance	Pearson Correlation	-.207	.197	.264*	-.116	.356**	.111	.415**	1	-.272**	.449**	-.143	.069
	Sig. (2-tailed)	.050	.063	.012	.275	.001	.300	.000		.009	.000	.178	.520
	N	90	90	90	90	90	90	90	90	90	90	90	90
Tangney Brief Self Control Scale	Pearson Correlation	.230*	.000	-.132	.405**	-.436**	-.294**	-.287**	-.272**	1	-.553**	.629**	-.016
	Sig. (2-tailed)	.029	.997	.216	.000	.000	.005	.006	.009		.000	.000	.877
	N	90	90	90	90	90	90	90	90	90	90	90	90
Work Avoidance Harackiewicz	Pearson Correlation	-.230*	.089	.317**	-.466**	.460**	.110	.313**	.449**	-.553**	1	-.504**	.061
	Sig. (2-tailed)	.029	.406	.002	.000	.000	.300	.003	.000	.000		.000	.569
	N	90	90	90	90	90	90	90	90	90	90	90	90
Self Efficacy Zimmerman	Pearson Correlation	.242*	-.058	-.177	.416**	-.326**	-.264*	-.184	-.143	.629**	-.504**	1	-.146
	Sig. (2-tailed)	.022	.584	.095	.000	.002	.012	.082	.178	.000	.000		.171
	N	90	90	90	90	90	90	90	90	90	90	90	90
WTF	Pearson Correlation	.009	-.032	.069	-.200	.072	-.032	-.006	.069	-.016	.061	-.146	1
	Sig. (2-tailed)	.936	.763	.517	.059	.502	.761	.953	.520	.877	.569	.171	
	N	90	90	90	90	90	90	90	90	90	90	90	90

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Class C Correlations

		PALS1 Mastery	PALS1 Performance Approach	PALS1 Performance Avoid	PALS4 Academic Efficacy	PALS4 Novelty Avoid	PALS4 Disruptive Behavior	PALS4 Self Presentation Low Achievement	PALS4 Skeptical School Relevance	Tangney Brief Self Control Scale	Work Avoidance Harackiewicz	Self Efficacy Zimmerman	WTF
PALS1 Mastery	Pearson Correlation	1	.213*	.171	.443**	-.271**	-.146	-.063	-.446**	.290**	-.316**	.282**	.026
	Sig. (2-tailed)		.038	.095	.000	.008	.160	.549	.000	.007	.003	.009	.799
	N	96	96	96	94	94	94	93	92	87	85	85	96
PALS1 Performance Approach	Pearson Correlation	.213*	1	.591**	.153	.094	.062	.141	.009	-.075	-.052	.018	.111
	Sig. (2-tailed)	.038		.000	.140	.368	.552	.177	.931	.489	.633	.871	.280
	N	96	96	96	94	94	94	93	92	87	85	85	96
PALS1 Performance Avoid	Pearson Correlation	.171	.591**	1	.169	.212*	.061	.217*	-.052	-.151	.016	.019	-.025
	Sig. (2-tailed)	.095	.000		.103	.040	.558	.037	.625	.161	.887	.860	.807
	N	96	96	96	94	94	94	93	92	87	85	85	96
PALS4 Academic Efficacy	Pearson Correlation	.443**	.153	.169	1	-.409**	-.076	-.161	-.434**	.327**	-.370**	.446**	-.020
	Sig. (2-tailed)	.000	.140	.103		.000	.466	.124	.000	.002	.000	.000	.848
	N	94	94	94	94	94	94	93	92	87	85	85	94
PALS4 Novelty Avoid	Pearson Correlation	-.271**	.094	.212*	-.409**	1	.285**	.279**	.279**	-.319**	.430**	-.501**	-.031
	Sig. (2-tailed)	.008	.368	.040	.000		.005	.007	.007	.003	.000	.000	.769
	N	94	94	94	94	94	94	93	92	87	85	85	94
PALS4 Disruptive Behavior	Pearson Correlation	-.146	.062	.061	-.076	.285**	1	.346**	.121	-.379**	.130	-.344**	-.055
	Sig. (2-tailed)	.160	.552	.558	.466	.005		.001	.250	.000	.234	.001	.600
	N	94	94	94	94	94	94	93	92	87	85	85	94
PALS4 Self Presentation Low Achievement	Pearson Correlation	-.063	.141	.217*	-.161	.279**	.346**	1	.147	-.188	.178	-.358**	-.064
	Sig. (2-tailed)	.549	.177	.037	.124	.007	.001		.162	.082	.104	.001	.540
	N	93	93	93	93	93	93	93	92	87	85	85	93
PALS4 Skeptic School Relevance	Pearson Correlation	-.446**	.009	-.052	-.434**	.279**	.121	.147	1	-.286**	.411**	-.347**	-.032
	Sig. (2-tailed)	.000	.931	.625	.000	.007	.250	.162		.007	.000	.001	.760
	N	92	92	92	92	92	92	92	92	87	85	85	92
Tangney Brief Self Control Scale	Pearson Correlation	.290**	-.075	-.151	.327**	-.319**	-.379**	-.188	-.286**	1	-.400**	.555**	.013
	Sig. (2-tailed)	.007	.489	.161	.002	.003	.000	.082	.007		.000	.000	.902
	N	87	87	87	87	87	87	87	87	87	85	85	87
Work Avoidance Harackiewicz	Pearson Correlation	-.316**	-.052	.016	-.370**	.430**	.130	.178	.411**	-.400**	1	-.531**	.139
	Sig. (2-tailed)	.003	.633	.887	.000	.000	.234	.104	.000	.000		.000	.206
	N	85	85	85	85	85	85	85	85	85	85	85	85
Self Efficacy Zimmerman	Pearson Correlation	.282**	.018	.019	.446**	-.501**	-.344**	-.358**	-.347**	.555**	-.531**	1	.026
	Sig. (2-tailed)	.009	.871	.860	.000	.000	.001	.001	.001	.000	.000		.812
	N	85	85	85	85	85	85	85	85	85	85	85	85
WTF	Pearson Correlation	.026	.111	-.025	-.020	-.031	-.055	-.064	-.032	.013	.139	.026	1
	Sig. (2-tailed)	.799	.280	.807	.848	.769	.600	.540	.760	.902	.206	.812	
	N	96	96	96	94	94	94	93	92	87	85	85	96

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Total Correlations

		PALS1 Mastery	PALS1 Performance Approach	PALS1 Performance Avoidance	PALS4 Academic Efficacy	PALS4 Novelty Avoidance	PALS4 Disruptive Behavior	PALS4 Self Presentation of Low Achievement	PALS4 Skeptical School Relevance	WTF
PALS1 Mastery	Pearson Correlation	1	.254**	.140*	.447**	-.329**	-.246**	-.141*	-.440**	-.039
	Sig. (2-tailed)		.000	.012	.000	.000	.000	.012	.000	.484
	N	320	320	320	318	318	318	317	316	320
PALS1 Performance Approach	Pearson Correlation	.254**	1	.569**	.145**	.073	.047	.023	.015	-.018
	Sig. (2-tailed)	.000		.000	.010	.194	.407	.678	.790	.755
	N	320	320	320	318	318	318	317	316	320
PALS1 Performance Avoidance	Pearson Correlation	.140*	.569**	1	.085	.114*	-.032	.188**	.051	.042
	Sig. (2-tailed)	.012	.000		.129	.042	.564	.001	.369	.454
	N	320	320	320	318	318	318	317	316	320
PALS4 Academic Efficacy	Pearson Correlation	.447**	.145**	.085	1	-.357**	-.126*	-.154**	-.335**	-.063
	Sig. (2-tailed)	.000	.010	.129		.000	.024	.006	.000	.258
	N	318	318	318	320	320	320	319	318	320
PALS4 Novelty Avoidance	Pearson Correlation	-.329**	.073	.114*	-.357**	1	.337**	.302**	.328**	.051
	Sig. (2-tailed)	.000	.194	.042	.000		.000	.000	.000	.359
	N	318	318	318	320	320	320	319	318	320
PALS4 Disruptive Behavior	Pearson Correlation	-.246**	.047	-.032	-.126*	.337**	1	.333**	.212**	-.015
	Sig. (2-tailed)	.000	.407	.564	.024	.000		.000	.000	.796
	N	318	318	318	320	320	320	319	318	320
PALS4 Self Presentation of Low Achievement	Pearson Correlation	-.141*	.023	.188**	-.154**	.302**	.333**	1	.265**	-.018
	Sig. (2-tailed)	.012	.678	.001	.006	.000	.000		.000	.755
	N	317	317	317	319	319	319	319	318	319
PALS4 Skeptical School Relevance	Pearson Correlation	-.440**	.015	.051	-.335**	.328**	.212**	.265**	1	-.009
	Sig. (2-tailed)	.000	.790	.369	.000	.000	.000	.000		.868
	N	316	316	316	318	318	318	318	318	318
WTF	Pearson Correlation	-.039	-.018	.042	-.063	.051	-.015	-.018	-.009	1
	Sig. (2-tailed)	.484	.755	.454	.258	.359	.796	.755	.868	
	N	320	320	320	320	320	320	319	318	330

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

References

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

[2] Wixon, M., Baker, R.S.J.d., Gobert, J., Ocumpaugh, J., Bachmann, M. (2012) WTF? Detecting Students who are Conducting Inquiry Without Thinking Fastidiously. Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UMAP 2012), 286-298.

[3] Buckley, B., Gobert, J., Horwitz, P., & O'Dwyer, I. (2012) Looking inside the black box: assessments and decision-making in BioLogica. *International Journal of Learning Technology,* 5 (2), 166-190.

[4] Baker, R.S.J.d., Corbett, A., Roll, I., Koedinger, K. (2008) Developing a Generalizable Detector of When Students Game the System. User Modeling and User-Adapted Interaction, Vol. 18, No. 3, 287-314.

[5] Shih, B., Koedinger, K. R., & Scheines, R. (2008) A response time model for bottom-out hints as worked examples. In R. S. J. d. Baker, T. Barnes,& J. Beck (Eds.), Educational Data Mining 2008: 1st International Conference on Educational Data Mining, Proceedings (pp. 117–126).Montreal, Quebec, Canada.

[6] Baker, R.S.J.d., Gowda, S.M., Corbett, A.T. (2011) Automatically Detecting a Student's Preparation for Future Learning: Help Use is Key. *Proceedings of the 4th International Conference on Educational Data Mining*, 179-188.

[7] Sao Pedro, M., Baker, R.S.J.d., Gobert, J. (2012) Improving Construct Validity Yields Better Models of Systematic Inquiry, Even with Less Information. Proceedings of User Modeling and Personalization UMAP 2012, 249-260.

[8] Gobert, J. (in press) Microworlds. To appear in Richard Gunstone (Ed.) Encyclopedia of Science Education. Springer.

[9] National Research Council (1996) *National Science Education Standards.* National Academy Press, Washington, D.C.

[10] Rowe, J., McQuiggan, S., Robison, J., & Lester, J. (2009) Off-Task Behavior in Narrative Centered Learning Environments. In Proceedings of the Fourteenth International Conference on Artificial Intelligence in Education (AIED-09), Brighton, UK, pp. 99-106.

[11] 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. Lecture Notes in Computer Science, Vol. 6738, 534-536.

[12] Cocea, M., Hershkovitz, A., Baker, R.S.J.d. (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.

[13] Carroll, J. (1963) A model of school learning. Teachers College Record, 64, 723-733.

[14] Karweit, N., & Slavin, R. (1981) Measurement and Modeling Choices in Studies of Time and Learning. *American Educational Research Journal,* Vol. 18, No. 2, 157-171.

[15] Frederick, W., & Walberg, H. (1980) Learning as a Function of Time. *The Journal of Educational Research,* Vol. 73, No. 4, 183-194.

[16] Bloom, B. S. (1976) Human characteristics and school learning, New York: McGraw-Hill.

[17] Lloyd, J., & Loper, A. (1986) Measurement and Evaluation of Task-Related Learning Behaviors: Attention To Task and Metacognition. School Psychology Review, Vol. 15, No. 3, 336-345.

[18] 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.

[19] Hershkovitz, A., Baker, R.S.J.d., Gobert, J., Wixon, M. (2011) Goal Orientation and Changes of Carelessness over Consecutive Trials in Science Inquiry. Poster paper. Proceedings of the 4th International Conference on Educational Data Mining, 315-316.

[20] Aleven, V., Koedinger, K.R., (2000) Limitations of Student Control: Do Students Know when They Need Help? In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Proceedings of the 5th International Conference on Intelligent Tutoring Systems, ITS 2000* (pp. 292-303). Berlin: Springer Verlag.

[21] Walonoski, J., & Heffernan, N. Detection and Analysis of Off-Task gaming Behavior in Intelligent Tutoring Systems. (2006) In Ikeda, Ashley & Chan (Eds.). Proceedings of the Eight International Conference on Intelligent Tutoring Systems. Springer-Verlag: Berlin. Pp. 382-391.

[22] Newman, M.A. (1977) An analysis of sixth-grade pupils' errors on written mathematical tasks. In M.A. Clements & J.Foyster (Eds.), Research in mathematics education in Australia, 1977 (Volume 1, pp.239-258). Melbourne: Swinburne Press.

[23] Clements, M. A. (1982) Careless Errors Made by Sixth-Grade Children on Written Mathematical Tasks. Journal for Research in Mathematics Education, Vol. 13, No. 2, 136-144.

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

[25] Dweck, C., & Leggett, E. (1988) A Social-Cognitive Approach to Motivation and Personality. Psychological Review, Vol. 95, No. 2, 256-273.

[26] Elliot, A., & Harackiewicz, J. (1996) Approach and Avoidance Goals and Intrinsic Motivation: A Mediational Analysis. Journal of Personality and Social Psychology, Vol. 70, 461-475.

[27] Skaalvik, E. (1997) Self-Enhancing and Self-Defeating Ego Orientation: Relations With Task and Avoidance Orientation, Achievement, Self-Perceptions, and Anxiety. Journal of Educational Psychology, Vol. 89, No. 1, 71-81.

[28] Liem, A., Shun, L., & Nie, Y. (2008) The role of self-efficacy, task value, and achievement goals in predicting learning strategies, task disengagement, peer relationship, and achievement outcome. Contemporary Educational Psychology, 33, 489-512.

[29] Nolen, S. (1988) Reasons for Studying: Motivational Orientations and Study Strategies. Cognition and Instruction, Vol. 5, No. 4, 269-287.

[30] Harackiewicz, J.M., Barron, K.E., Tauer, J.M., & Elliot, A.J. (2002) Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. Journal of Educational Psychology, 94, 562-575.

[31] Myers, D., Milne, A., Baker, K., & Ginsburg, A., (1987) Student Discipline and High School Performance. Sociology of Education 60:18-33.

[32] Finn, J.D., Pannozzo, G.M., Voelkl, K.E., (1995) Disruptive and Inattentive-Withdrawn Behavior and Achievement among Fourth Graders. The Elementary School Journal Vol 95 No. 5, 421-434.

[33] Zimmerman, B., Bandura, A., & Martinez-Pons, M. (1992) Self-Motivation for Academic Attainment: The Role of Self-Efficacy Beliefs and Personal Goal Setting. American Educational Research Journal, Vol. 29, No. 3, 663-676.

[34] Bandura, A., Barbaranelli, C., Caprara, G., & Pastorelli, C. (1996) Multifaceted Impact of Self-Efficacy Beliefs on Academic Functioning. Child Development, Vol. 67, No. 3, 1206-1222.

[35] Baker, R.S., Corbett, A.T., Koedinger, K.R. (2004) Detecting Student Misuse of Intelligent Tutoring Systems. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, 531-540.

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

[37] Baker, R.S.J.d., de Carvalho, A. M. J. A. (2008) Labeling Student Behavior Faster and More Precisely with Text Replays. Proceedings of the 1st International Conference on Educational Data Mining, 38-47.

[38] Baker, R.S.J.d., 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.

[39] Baker, R.S.J.d. (2007) Is Gaming the System State-or-Trait? Educational Data Mining Through the Multi-Contextual Application of a Validated Behavioral Model. *Complete On-Line Proceedings of the Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling 2007*, 76-80.

[40] 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.

[41] Tangney, J.P., Baumeister, R.F., & Boone, A.L. (2004) High Self-Control Predicts Good Adjustment, Less Pathology, Better Grades, and Interpersonal Success. Journal of Personality, 72:2, 272-324.

[42] Harackiewicz, J.M., Barron, K.E., Carter, S.M., Lehto, A.T., Elliot, A.J. (1997) Predictors and Consequences of Achievement Goals in the College Classroom: Maintaining Interest and Making the Grade. Journal of Personality and Social Psychology, 73:6, 1284-1295.

[43] Midgley, C., Maehr, M., Hicks, L., Roeser, R., Urdan, T., Anderman, E., et al. (1997) Patterns of Adaptive Learning Survey (PALS). Ann Arbor: University of Michigan.

[44] Dweck, C. S. & Elliot, E. S. (1983) Achievement Motivation. In P. Mussen & E. M. Hetherington (Eds.), *Handbook of child psychology* (pp. 643-691). New York: Wiley.

[45] Middleton, M. J., & Midgley, C. (1997) Avoiding the demonstration of lack of ability: An unexplored aspect of goal theory. Journal of Educational Psychology, 89, 710-718.

[46] Urdan, T., Midgely, C., & Anderman, E. (1998) The Role of Classroom Structure in Students' Use of Self-Handicapping Strategies. American Educational Research Journal, Vol. 35, No. 1, 102-122.

[47] Kaplan, A., Gheen, M., & Midgley, C. (2002) Classroom goal structure and student disruptive behavior. British Journal of Educational Psychology, 72, 191–211.

[48] Midgley, C., Arunkumar, R., & Urdan, T. (1996) If I don't do well tomorrow, there's a reason: Predictors of adolescents' use of academic self-handicapping behavior. Journal of Educational Psychology, 88, 423– 434.

[49] Gobert, J., Sao Pedro, M., Baker, R.S., Toto, E., & Montalvo, O. (2012) *Leveraging educational data mining for real time performance assessment of scientific inquiry skills within microworlds,* Journal of Educational Data Mining, *Article 15, Volume 4,* 153-185.

[50] Sao Pedro, M, Baker, R.S.J.d., Montalvo, O., Nakama, A., & Gobert, J. (2010) Using text replay tagging to produce detectors of systematic experimentation behavior patterns. Proceedings of the 3rd International Conference on Educational Data Mining, 181-190.

[51] Corbett, A.T., Anderson, J.R.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction, 4, 253-278. (1995)

[52] Sao Pedro, M., Gobert, J., & Baker, R. (2012, April 15) Assessing the Learning and Transfer of Data Collection Inquiry Skills Using Educational Data Mining on Students' Log Files. Paper presented at The Annual Meeting of the American Educational Research Association. Vancouver, BC, CA: Retrieved April 15, 2012, from the AERA Online Paper Repository

[53] Gobert, J., Sao Pedro, M. Raziuddin, J., & the Science Assistments Team (2010) Studying the interaction between learner characteristics and inquiry skills in microworlds. In K. Gomez, L. Lyons, & J. Radinsky (Ed.), Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences (ICLS 2010) - Volume 2 (p. 46). Chicago, IL: International Society of the Learning Sciences.

[54] Baker, R.S.J.d., Corbett, A.T., Wagner, A.Z. (2006) Human Classification of Low-Fidelity Replays of Student Actions. *Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems*, 29-36.

[55] SPSS Inc. Released 2008. SPSS Statistics for Windows, Version 17.0. Chicago: SPSS Inc.

[56] Witten, I.H. & Frank, E. (2005) Data Mining: Practical machine learning tools and techniques with Java implementation (2nd Edition). San Fracisco, CA: Kaufman Publishers Inc.

[57] Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T. (2006) YALE: Rapid Prototyping for Complex Data Mining Tasks. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 935-940.

[58] Efron, B., Gong, G. (1983) A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. The American Statistician, 37 (1), 36-48.

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

[60] Esposito, F., Licchelli, O., & Semeraro, G. (2004) Discovering Student Models in e-learning Systems, Journal of Universal Computer Science, Vol. 10, No. 1, pp.47-57.

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

[62] J. Davis & M. Goadrich (2006) The Relationship Between Precision-Recall and ROC Curves. Proceedings of the Twenty-Third International Conference on Machine Learning (ICML'06), Pittsburgh, PA.

[63] Ben-David, A. (2008) About the Relationship between ROC Curves and Cohen's Kappa. Engineering Applications of Artificial Intelligence, 21, 874-882.

[64] Cetintas, S., Si, L., Xin, Y.P., Hord, C. (2009) Automatic Detection of Off-Task Behaviors in Intelligent Tutoring Systems with Machine Learning Techniques. IEEE Transactions on Learning Technologies, 3 (3), 228-236.

[65] Wigfield, A., and Eccles, J. S. (2000) Expectancy-value theory of achievement motivation. Contemporary Educational Psychology 25: 68–81.

[66] Baker, R.S.J.d., 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.

[67] Roll, I., Aleven, V., McLaren, B.M., Koedinger, K.R. (2007) Can help seeking be tutored? Searching for the secret sauce of metacognitive tutoring. Proceedings of the 13th International Conference on Artificial Intelligence in Education, 203-210.

[68] 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.

[69] 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.

[70] Kim, Y., & Wei, Q. (2011) The impact of user attributes and user choice in an agent-based environment. *Computers & Education,* 56 (2), 505-514.

[71] Duckworth, A., Peterson, C., Matthews, M., & Kelly, D. (2007) Grit: Perseverance and Passion for Long-Term Goals. *Journal of Personality and Social Psychology,* Vol. 92, No. 6, 1087-1101.