Disengagement and Performance in Science Inquiry within Inq-Its, a Virtual Learning Environment

A Major Qualifying Project

Submitted to the Faculty of

Worcester Polytechnic Institute

in partial fulfillment of the requirements for the

Degree in Bachelor of Science

in

Psychological Science

By

_____

Adrian Oyola

_____

Ian Schuba

Date: 4/30/2014

Project Advisor:

_____

Professor Janice Gobert, Advisor

Abstract

Engagement in learning has increasingly become an important area of focus. This paper addresses the difficulty of accurately measuring engagement through prior methods, and presents a method of measuring engagement concretely through the use of a detector. We build off of prior work in operationalizing and detecting disengaged behaviors which are detrimental to engagement through the use of a real-time detector (Gobert, Baker & Wixon, 2015; Wixon, 2013). This detector is modified and applied to students engaging with the Inquiry-Intelligent Tutoring System, Inq-ITS (Gobert et al, 2013). Lastly, we look at relationships between disengagement and performance during science inquiry learning.

Acknowledgements

This MQP project was made possible by help and contributions from many professors and students. Special thanks to:

Prof. Janice Gobert for advising this project and assisting with revisions,

Ermal Toto and Michael Sao Pedro for helping with data collection,

Luc Paquette for helping with modifying the detector and helping with the data coding,

Yoon Jeon Kim for helping with the data interpretation and assisting with revisions,

Michael Wixon for assisting with revisions,

and the students who participated in our research.

This research was supported by grant "Empirical Research: Emerging Research: Using Automated Detectors to Examine the Relationships Between Learner Attributes and Behaviors During Inquiry in Science Microworlds", National Science Foundation award #DRL-100864 awarded to Janice Gobert and Ryan Baker.

Executive Summary

Engagement is an important part of learning, and plays a big role in determining academic success. Being able to accurately and concretely measure engagement are high priority goals due to this importance. Prior methods of engagement have not been able to satisfactorily meet these goals due to the difficulty of measuring engagement. By instead measuring *dis*engagement, we seek to overcome this difficulty. Building off of prior work in developing a detector that is able to determine disengaged behaviors unobtrusively, we believe that the detector satisfies these goals in a way that does not interfere with student learning. This detector specifically looks at a kind of disengagement known as being *Disengaged from Task Goal* (DTG), where students are engaging with a learning task in a way that was not intended by the system's designers and/or task goals. Applying this DTG detector to data collected from the virtual learning environment known as Inquiry-Intelligent Tutoring System (Inq-ITS, Gobert et al, 2013) allows us to determine instances of disengagement within learning tasks and relate disengagement with inquiry performance in the learning tasks.

Extending work done on the development of the original DTG detector, there was a need to rebuild it for application to a different version of the learning environment, Inq-ITS. A set of new features were established and "observed" via the new detector to make it compatible for our purposes. The DTG detector was found to have a high confidence in distinguishing between DTG and non-DTG behavior and performed considerably better than chance. The DTG detector was found to only have issues in determining edge cases, returning a higher amount of false positives than the prior detector.

Data was collected through Inq-ITS, which recorded logs of student behavior and activity within the learning tasks. These logs were converted into segmented clips that began when students entered an inquiry phase and ended when they left the experimental phase for that inquiry phase. A human coder also reviewed each clip with the added context of prior clips to check for any overlooked behaviors. A pretest and posttest were also administered to measure knowledge of the subject matter and to look for any improvement.

Negative correlations between exhibiting DTG behavior and performance in the learning tasks, as well as in the pretest and posttest, were found. This means that students who were DTG were found to perform worse than those who weren't. A pattern was also found in performance across phases within the different learning tasks. As students progressed through the phases of inquiry, performance was found to decrease, especially for DTG students. The implications of this are discussed further.

Ultimately we found the use of the DTG detector to be an effective method in measuring disengagement accurately and unobtrusively. The DTG detector also allows for rigorous detection of and statistical analysis about disengagement, which is an improvement from prior methods of measuring engagement and disengagement (Gobert, Baker, & Wixon, 2015). Applying the DTG detector with detectors that look at other aspects of disengagement such as carelessness may be an interesting extension for future work.

Table of Contents

List of Figures

List of Tables

Introduction

Importance of Engagement and Disengagement

Simple interaction, absent of structure and leadership is not enough for deep learning. Student engagement is a major factor in education, as engagement plays a big role in effecting social and psychological experiences and well as in helping students develop long-term academic success (Skinner & Pitzer, 2012). There are many dimensions of engagement that are effective, and all of them prove powerful at any education level. Research suggests that the reflective and collaborative properties of asynchronous, text-based online learning are well adapted to deep approaches to learning (i.e., cognitive presence). (Garrison, 2010).

Authentic learning environments can provide great alternatives to learning science from traditional approaches that tend to emphasize decontextualized facts and skills, and as such learning environments can enhance the acquisition and transfer of deep and lifelong learning (Piccoli et al, 2001). The research suggests that the use of authentic learning settings can provide strong supports for learners, if engaged. Authentic virtual tasks have the capability to motivate and encourage learner participation (Nolen, 2003).

In 1995, Csikszentmihalyi coined the term `flow' to refer to `optimal experience' events. The earliest writings on flow have suggested the level of difficulty may be important in an educational learning environment. Specifically, if the task is too difficult, the student may disengage; similarly, if the task is too easy, the student may disengage. At the extreme end of engagement, one might experience a flow state (Csikszentmihalyi, 1991). Flow is described as a state of complete absorption or engagement in an activity. The term was introduced through the study of people involved in common activities such as rock climbing, dancing, chess, etc. A `flow activity' is one in which the mind becomes effortlessly focused and engaged on an activity,

rather than falling prey to distractions. Flow is not an `all-or-nothing' state, but can be thought of as forming a continuum from no flow to maximum flow (Pearce, et. al. 2005).

But how can we design educational learning environments to optimize students' learning so that flow might be possible? We begin to address this here by addressing how to measure engagement in virtual environments. Also addressed is whether there is a pattern to students' disengagement and its relationship to learning.

Prior Methods of Measuring Engagement and Disengagement

One of the most common methods of measuring engagement is through administering surveys to participants. The National Survey of Student Engagement is one such survey that is annually distributed to different colleges and universities to collect data on the overall behaviors and practices of student populaces that provides insight into student engagement. Studies have previously used data from the NSSE to investigate relationships between overall student engagement and the use of online learning tools (Chen, Guidry, & Lambert, 2010). Modified versions of the NSSE have also been used to look at different engagement factors for both online and on campus students to determine their importance in learning (Hullinger & Robinson, 2008).

In a survey developed by the National Survey of Student Engagement (NSSE), the researchers utilized a hierarchical linear model (HLM) and multiple regressions to investigate the impact of Web-based learning technology on student engagement and self-reported learning outcomes in face-to-face and online learning environments. They separated their study into three research questions:

1. How often do college students in different types of courses use the Web and Internet technologies for course-related tasks?

2. Do individual and institutional characteristics affect the likelihood of taking online courses?

3. Does the relative amount of technology employed in a course have a relationship to student engagement, learning approaches, and student self-reported learning outcomes?

The results point to a positive relationship between Web-based learning technology use and student engagement and desirable learning outcomes. This shows the importance and effectiveness of online learning environments, which leads to our further investigation. (Chen, Guidry, & Lambert, 2010).

However, it is important to note that the NSSE is not able to *directly* assess student learning and engagement. It looks at the general practices of students to get a sense of overall student engagement at different colleges and universities, and is not effective at looking at individual engagement or engagement *during* learning tasks. Our method of measuring engagement is more rigorous and valid, and can be used in real time in Inq-ITS (Gobert et al, 2013).

The Study Process Questionnaire is a method of measuring engagement that is designed for looking at individual student engagement (Biggs et al, 2001). The SPQ scores students in terms of different levels of engagement. The general categorizations look at the surface style of engagement where the minimum effort and action required is met, deep engagement, which involves an intrinsic interest in the task and topic as well as the inter-relation with prior student knowledge. The SPQ also scores students on achieving engagement where there is a sense of competition with others and a tendency towards behaving like an archetypal model student, who is organized and uses established learning strategies. The SPQ previously has been used to look at student engagement within online courses, such as assessing the depth of online learning and

the nature of interactions (Cleveland-Innes & Garrison, 2005), and in assessing the relationship between cognitive engagement in learning within online courses (Newby & Richardson, 2006).

However, there are issues and limitations with using surveys as a method for measuring student engagement. One limitation is that administering a survey repeatedly throughout a task may be disruptive to the student and the task. Therefore, these surveys must be distributed out of the context of the task. Because of this, one will only be able to get data on the overall engagement during the task compared to before the task. This method of measuring engagement assumes that engagement levels would not change within the task and that it would stay static, contrary to the more fluid nature of engagement (Appleton et al, 2008; Fredricks et al 2004). Any data acquired this way would not be as accurate as data acquired during the task.

The use of these surveys in measuring engagement also all involve self-report from the students. This comes with several issues such as the student's perception of their own engagement may not be as accurate or precise as direct measurements of engagements. A student exploiting the system might also be hard to detect, making data collected less reliable. This is paired with the general vagueness of how engagement is defined within these surveys as compared to potential quantifiable measurements of engagement. The more qualitative methods of defining the data collected from these surveys leaves a lot of the definition of engagement up to subjective interpretation.

Important to this area of research is our goal: we are addressing and detecting disengagement during student learning in Inq-ITS (Gobert et al, 2013). In attempting to measure engagement, student disengagement has been overlooked as a part of the overall picture of student engagement (Gobert, Baker, & Wixon, 2015). Looking at and detecting student

disengagement provides a source of data that will help clarify the overall construct of engagement and could be the missing piece of the puzzle that is needed in rigorously observing and identifying engagement (Gobert et al, 2015).

Rationale

Being able to properly measure engagement is necessary due to the importance of student engagement in learning (Corno & Mandinach, 1983). A method of measuring engagement with an implementation that would be able to collect data throughout a learning task unobtrusively would be the most effective way to collect data on engagement. This would allow for detecting changes in the level of engagement throughout the task, looking at points during the task where students become more or less engaged. Observing engagement throughout the task is important as it has been observed that engagement is malleable and changes throughout tasks (Appleton et al, 2008; Fredricks et al 2004). Our work here seeks to implement such a method through automatic detection of disengaged behavior, using a previously developed detector of disengagement (Gobert, Baker & Wixon, 2015; Wixon, 2013). This is unnoticeable and unobtrusive to the student and is able to collect information on their engagement throughout the task, rather than a collective survey-based assessment after the task, as has been largely done in the past (Hullinger & Robinson, 2008; Cleveland-Inness & Garrison, 2005; Newby & Richardson, 2006).

A method like this also has to strictly define engagement in terms of what is being measured, and in doing so provides quantifiable data on engagement. Our work builds off of an established framework for detecting specific disengaged behavior that we refer to as being Disengaged from Task Goal (DTG). DTG takes the form of situations wherein students attempt to use learning tools in ways that are contrary to achieving learning goals, such as in drawing

pictures within a math plot or in running excessively large amounts of trials in a virtual experiment. Another example of behavior indicative of DTG is when students just click through tasks within seconds, in a way that indicates they are not attempting to learn and apply skills. Prior work has previously been successful at identifying disengagement in Inq-ITS (Gobert, Baker & Wixon, 2015; Wixon, 2013).

In observing engagement, paying special attention to disengagement as an area of focus to operationalize will provide a better model for observing engagement, which has not previously been accomplished. In taking account these goals in measuring engagement, we seek to explore the use of a detector developed to concretely look at moments of disengagement through observing DTG moments. As previously stated, prior research on engagement has only looked at areas such as on-task or off-task student behaviors (Carroll, 1963; Lahaderne, 1968; Karweit & Slavin, 1982) or at an overall view of engaged and disengaged behaviors (Fredericks et al., 2011). In observing DTGs we can implement a method for measuring engagement through disengagement in ways that have not previously been attempted outside of our research and in a way that provides meaningful data about when students disengage within a learning task. It is also possible to look at relationships between DTG behavior and performance in a task, determining whether students who are disengaged from the task goal also have poorer performance on learning tasks in the Inq-ITS environment. Observing patterns in DTG behavior also may be potentially used as a tool to determine effectiveness of tasks in retaining student engagement.

Method

Participants

Our participants were 157 8th Grade public school students from a middle school in Central Massachusetts. Participants were assigned a 4 digit ID for identification and confidentiality purposes.

Materials

Inq-ITS

The Inquiry Intelligent Tutoring System, which we refer to as Inq-ITS (www.slinq.org; Gobert et al, 2013), is a virtual learning environment to help students learn and hone inquiry skills using various microworlds that cover various science related topics from middle school science including earth science, life science, and physical science. Inq-ITS includes various microworlds with which students can explore and cover topics such as cellular biology and the physics of free fall. Using Inq-ITS as the environment in which we measure disengagement allows us to gather data on student interactions, as it logs all students' interactions within the environment. This will help give us information that can be used to detect student engagement and disengagement in a virtual science learning environment in real time when implemented into the system.

For our purposes, we will be using the Phase Change microworld, where students are provided a virtual environment where they can change variables in the simulation of solids, liquids and gases during the process of conducting science inquiry. Students are tasked with developing hypotheses about how changes will affect the boiling point and melting point of the substance, then run experiments that will provide them with the data they will interpret to determine if their hypotheses were scientifically accurate. In the Phase Change microworld, this involves manipulating a block of ice through a Bunsen burner, looking at how variables such as heat intensity and the amount of ice will affect what will happen. The Phase Change microworld

consists of 4 different activities looking at heat and boiling point, amount of ice and boiling point, amount of ice and melting point, and size of container and boiling point. Each activity has 4 separate subsections where students formulate a hypothesis, design and run an experiment, analyze and interpret their results, and warrant their claims. Within the microworld, student actions and behaviors are logged, allowing us to collect relevant data for analysis of their disengagement and the relationship between disengagement and learning.

DTG Detector

The Disengaged from Task Goal (DTG) detector we are using was developed to help measure disengagement unobtrusively (Wixon, 2013; Gobert, Baker & Wixon, 2015). The kind of disengagement specifically observed are behaviors that are not necessarily off-task, but that instead shows behaviors within the task that ignore the task's goals or structure. For instance, a student who would repeatedly run the same experiment much more than needed would be considered disengaged from the task goal. Another example is a student who changes the independent variable many more times than needed. A student who takes excessive pauses between actions would also be considered as DTG.

Procedure

DTG Detector Development

This DTG detector can be applied for use in an Inq-ITS session and looks at a set of defined features to determine disengagement. The DTG detector uses the overall statistics of the data collected which are considered as several features, such as the total number of actions, the average time between actions, the maximum time between actions, and the number of trials. Other features observed include features based on pauses observed, such as the number of pauses during each run, the average length of each pause, and the duration of the longest observed

pause. Along with pauses, resets are also observed, with the number of trials run without either being noted, as well as the number of trials that included a reset, the average time spent before trials that did not include a reset as well as that did include a reset, and the maximum time spent before a trial that was reset before being finished. The DTG detector also looks at features related to the time elapsed in the experimental phase, looking at the total and average time spent between each trial, as well as the standard deviation and the maximum time spent between trials. Finally, the DTG detector looks at the features related to changes students make to variables while forming their hypotheses. This includes the number of changes to the independent variable during the experiment and the total and average time spent before a variable change, as well as the standard deviation of these cases.

The DTG detector was constructed through the use of machine learning, using defined features as variables to design algorithms that would be able to establish connections between the outlined features (Gobert, Baker & Wixon, 2015). The models produced from these algorithms consist of varying conditional statements that function as rules. Specifically, the detector then follows a model to determine disengagement from task goal from these features, using a set of six rules to determine disengagement from the task goal. This model has been cross-validated and avoids the risk of confirmation bias from researchers (Gobert, Baker & Wixon, 2015; Wixon, 2013).

During the development of the DTG detector, various classification algorithms were tested for effectiveness, including Naïve Bayes and J48 decision trees. Ultimately, this detector performed the best with the use of the PART algorithm, which establishes rules through the repeated construction of a decision tree and then determining the path which ends at the optimal lead node, setting a rule with the knowledge of this path. To evaluate this model, students were

separated randomly into six groups and used the data from five of these groups to develop a detector for every combination. This detector was then tested on the sixth group of students, allowing for the cross-validation of this model and ensuring its accuracy.

To apply the DTG detector to the new data set, the DTG detector had to be rebuilt. This was necessary, as the Inq-ITS system used in the present data collection uses different features than the ones on which the original DTG detector was based. The new detector looks at a different set of features, starting with the amount of runs. It also looks at the amount of time spent overall (both the maximum time spent overall and the average time spent overall) and the maximum amount of time spent on runs. Other features include the amount of changes to hypothesis variables, which includes the total counts, the amount of changes overall, and the standard deviation. Finally, the detector also looks at the amount of independent variable changes when experimenting, which includes the amount of changes overall, the average amount of changes, and the standard deviation. With the DTG detector rebuilt, it was here applied to new data acquired from the Inq-ITS system.

To evaluate the new DTG detector, A', Kappa, precision and recall were examined as metrics which are used to determine its performance. A' is the probability that the DTG detector will properly differentiate a clip that shows DTG behavior and a clip that doesn't, with an A' of .5 meaning essentially chance, similar to a coin flip, and an A' of 1.0 meaning complete accuracy. The DTG detector was able to achieve an A' of 0.869+/-0.089, which would mean that it would be able to tell the difference between examples involving or not involving DTG behavior approximately 86.9% of the time, considerably higher than chance. Cohen's Kappa is an evaluation metric that determines if the DTG detector performs better than chance when determining if clips demonstrate DTG behavior, with a Kappa of 0 representing a chance

performance and a Kappa of 1 representing complete accuracy. The Kappa value was determined to be .319+/-0.127, meaning that the detector's performance was approximately 31.9% better than a chance performance. Taking these values into account demonstrates that this detector is able to determine DTG behaviors correctly, only being incorrect in vague or "edge" cases.

Table 1: Confusion Matrix

	True 0	True 1
Predicted 0	463	8
Predicted 1	23	7

Looking at the algorithm's performance in a confusion matrix, we can see whether the detector is classifying the presence of DTG behavior accurately (Table 1). The detector was mostly able to correctly identify the presence of DTG behavior, and could moderately distinguish between DTG and non-DTG behavior. The detector had precision of 23% and recall of 46.7%, which are similar to values determined in prior implementation of the DTG detector (Gobert, Baker, & Wixon, 2015). Compared to the old detector, the new detector has more false positives. However, it is important to consider that A' is the only metric that takes into account detector confidence and that A' went up from the old detector, which would mean that the detector has a higher certainty when it distinguishes between behavior.

Data Collection

The students who participated in the study engaged in the tasks in the Inq-ITS system as part of their science classes. They were given a pretest within the system to test their prior content knowledge about state change. Using the Phase Change microworld, the students experiment and learn about the changes of matter. The students formed hypotheses and then tested their hypotheses by running simulated experiments. After that, the students interpreted the

data, warranted their claims, and reported their results. Students were then given a posttest after they completed their work in the Phase Change microworld testing how much they learned about the content. Actions within the software were logged as students performed these tasks, including the action type, the relevant simulation variable values, and the time stamp.

In order to review this data to find DTG behaviors, it was converted to a readable format. This was done by turning each data log into segmented clips. These clips begin when students entered an inquiry phase in Inq-ITS (i.e., made a hypothesis) and end once they left the experimental phase for that hypothesis. Students would generally move through all of the inquiry phases in order, but it was also possible for students to return to the data collection phase after analyzing and interpreting their current data, which would make a clip that would begin as they started interpreting their data and would end once they moved to collect more data.

642 clips derived from 157 students were coded individually with a human coder having access to prior clips that would provide contextual information related to DTG behavior that may otherwise have been overlooked. Of these clips, 32 clips were determined to show DTG behavior, with 5 out of 27 students that exhibited DTG behavior having more than one instance of DTG behavior. This means that approximately 5% of clips involved DTG behavior, which is similar to prior data involving detector development in disengagement (Baker & de Carvalho, 2008). In order to develop an automated detector of DTG from the log files, the features in the DTG detector allow us to pinpoint particular points in which students engage in DTG behaviors.

Pre and Post Test Learning Measures

We measured the student's knowledge of the subject matter before they performed tasks in the Inq-ITS microworld and after. This allows us to see if there were any correlations between

DTG and their knowledge or skill on the subject matter. In order to measure an individual student's improvement on a task, we used a formula different from the one frequently used in education research = (post-pre)/(100-pre). (Marx and Cummings, 2007) We needed to change the equation in the case that the student's performance after completing the Inq-ITS was lower than before they participated in the system. We decided to utilize the piecewise function proposed by Marx and Cummings to solve the limitations of the previous formula:

$$c = \begin{cases} \frac{\text{post-pre}}{100 - \text{pre}} & \text{post} > \text{pre} \\ \text{drop} & \text{post} = \text{pre} = 100 \text{ or } 0 \\ 0 & \text{post} = \text{pre} \\ \frac{\text{post} - \text{pre}}{\text{pre}} & \text{post} < \text{pre} \end{cases}$$

This equation allows us to calculate normalized changes for every student under all circumstances. We are able to accurately measure and compare the progress of each student's knowledge through the questions provided in the system.

Results

From the Phase Change microworld, we looked at pretest and posttest scores, as well as performance in the heat and boiling point activity (X1063), the amount of ice and boiling point activity (X1077), the amount of ice and melting point activity (X1096), and the size of container and boiling point activity (X1097). We were also able to look at student performances over the various inquiry skills (hypothesizing [HYPO], designing controlled experiments [DCE], interpreting data [INTER], and warranting their claims [WARRANT]), as well as their overall performance in each activity. Instances of DTG behavior and the frequency of these behaviors were also recorded, allowing us to look for correlations between being disengaged from the task goal and performance at the task.

Table 2: Correlation Matrix

	DTG	Frequency	Pretest	Posttest	GainScore
DTG	1	NA	-0.17	-0.22	-0.17
Frequency	NA	1	-0.22	-0.12	0.02
Pretest	-0.17*	-0.22	1	0.75	0.09
Posttest	-0.22**	-0.12	0.75**	1	0.59
GainScore	-0.17*	0.02	0.09	0.59**	1
X1063_DCE	-0.11	-0.23	0.27**	0.25**	0.23**
X1063_HYPO	-0.19**	-0.09	0.23**	0.26**	0.17*
X1063_INTER	-0.17**	-0.37*	0.36**	0.38**	0.25**
X1063_WARRANT	-0.17**	-0.41**	0.33**	0.37**	0.27**
X1063_Overall	-0.18**	-0.36*	0.35**	0.38**	0.27**
X1077_DCE	-0.16**	-0.35*	0.17*	0.26**	0.26**
X1077_HYPO	-0.12	-0.33*	0.23**	0.26**	0.23**
X1077_INTER	-0.18**	-0.19	0.28**	0.38**	0.32**
X1077_WARRANT	-0.2**	-0.37*	0.32**	0.41**	0.33**
X1077_Overall	-0.18**	-0.33*	0.29**	0.38**	0.33**
X1096_DCE	-0.15*	0.15	0.15**	0.26**	0.27**
X1096_HYPO	-0.14*	-0.44**	0.24**	0.29**	0.22**
X1096_INTER	-0.12	-0.36*	0.35**	0.49**	0.36**
X1096_WARRANT	-0.13	-0.34*	0.39**	0.49**	0.37**
X1096_Overall	-0.15*	-0.28	0.35**	0.47**	0.37**
X1097_DCE	-0.21**	-0.21	0.18**	0.21**	0.09
X1097_HYPO	-0.18**	-0.37*	0.14	0.16*	0.19**
X1097_INTER	-0.24**	-0.34*	0.28**	0.38**	0.31**
X1097_WARRANT	-0.27**	-0.4**	0.32**	0.4**	0.32**
X1097_Overall	-0.27**	-0.41**	0.3**	0.4**	0.29**
AcrossAll	-0.23**	-0.41**	0.36**	0.46**	0.36**

*p < 0.1, **p < 0.05

In general, it can be seen that there are negative correlations between exhibiting instances of DTG behavior and performance across all tasks in all activities (Table 2). This means that as instances of disengagement increase, performance in all tasks decrease. In addition, regarding performance on the pretest and posttest, negative correlations were found with exhibiting

instances of DTG behavior, meaning that as instances of disengagement increase, performance on the pretest and posttest decrease. These findings are generally statistically significant, as indicated by p-values below 0.05 (or below 0.1 for the designing controlled experiments and hypothesizing parts of the third activity, as well as the third activity overall), with exceptions in the designing controlled experiments (DCE) part of the first activity, hypothesizing (HYPO) in the second activity, and interpretation of the data (INTER) and warranting claims (WARRANT) in the third activity. Also shown are negative correlations between frequency of DTG behavior and performance for most tasks and in the pretest and posttest. However, findings related to the frequency of DTG behavior and performance were generally not statistically significant (Table 2).

From our data, we can observe differences in performance of students who showed DTG behavior and those who didn't. The variations between each activity help show the differences in performance for DTG and non-DTG behavior exhibiting students. Looking across each activity, there is a noticeable trend that can be seen in the progression through the different subsections of the activity. Looking at each activity, performance was worse as students progressed towards the later phases of inquiry. This can be seen through the distribution of data shown in each box plot (Figures 1-4), where the distribution falls across each inquiry phase for a given activity. This trend is found in all of the activities except for the third, amount of ice and melting point, for non-DTG students. This trend is especially apparent in the students who demonstrated DTG behavior. This trend across each subsection of each activity also led to overall performances in each activity being lower, especially for students exhibiting DTG behavior.

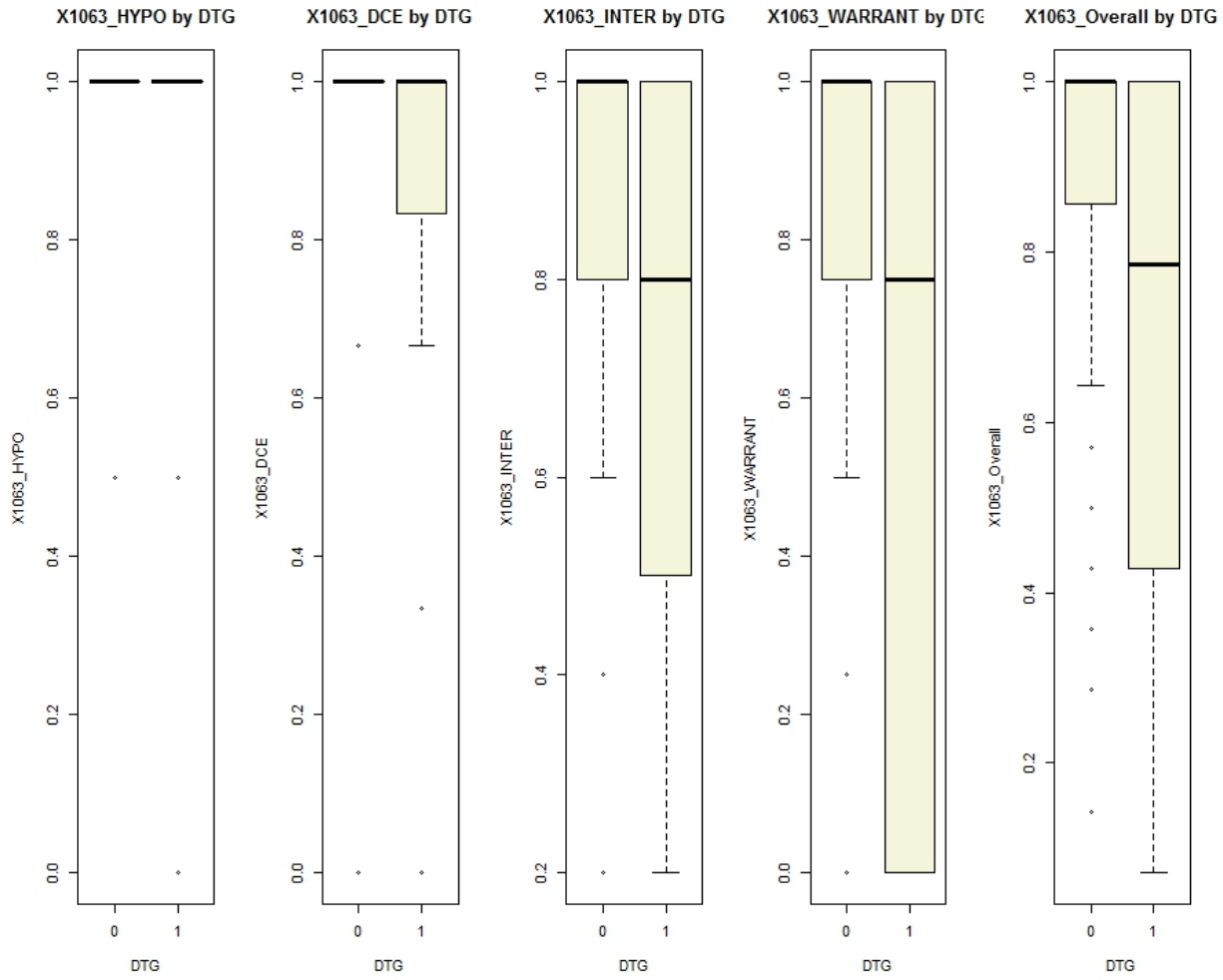

Figure 1. Performance across subsections and overall performance in the Heat and Boiling Point task by DTG

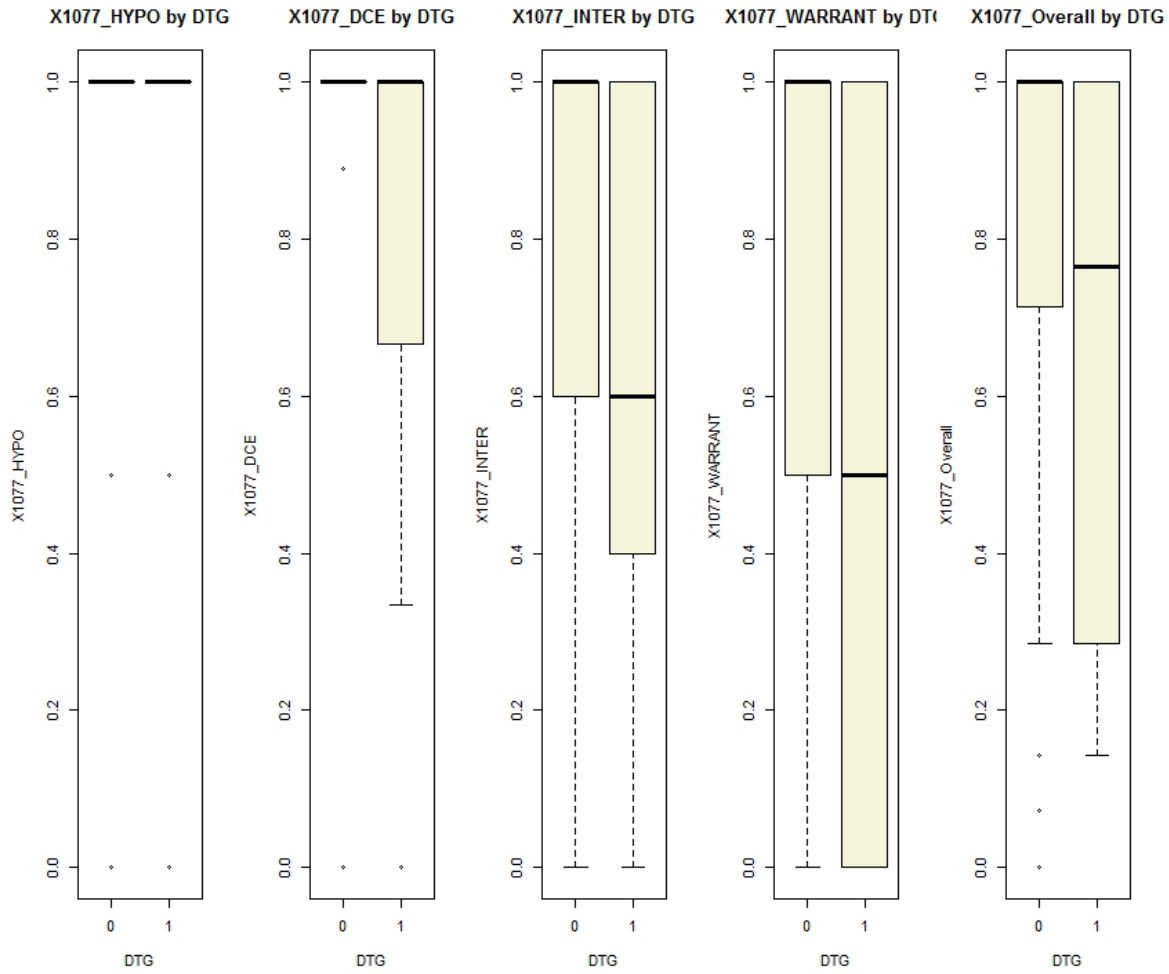

Figure 2. Performance across subsections and overall performance in the Amount of Ice and Boiling Point task by DTG

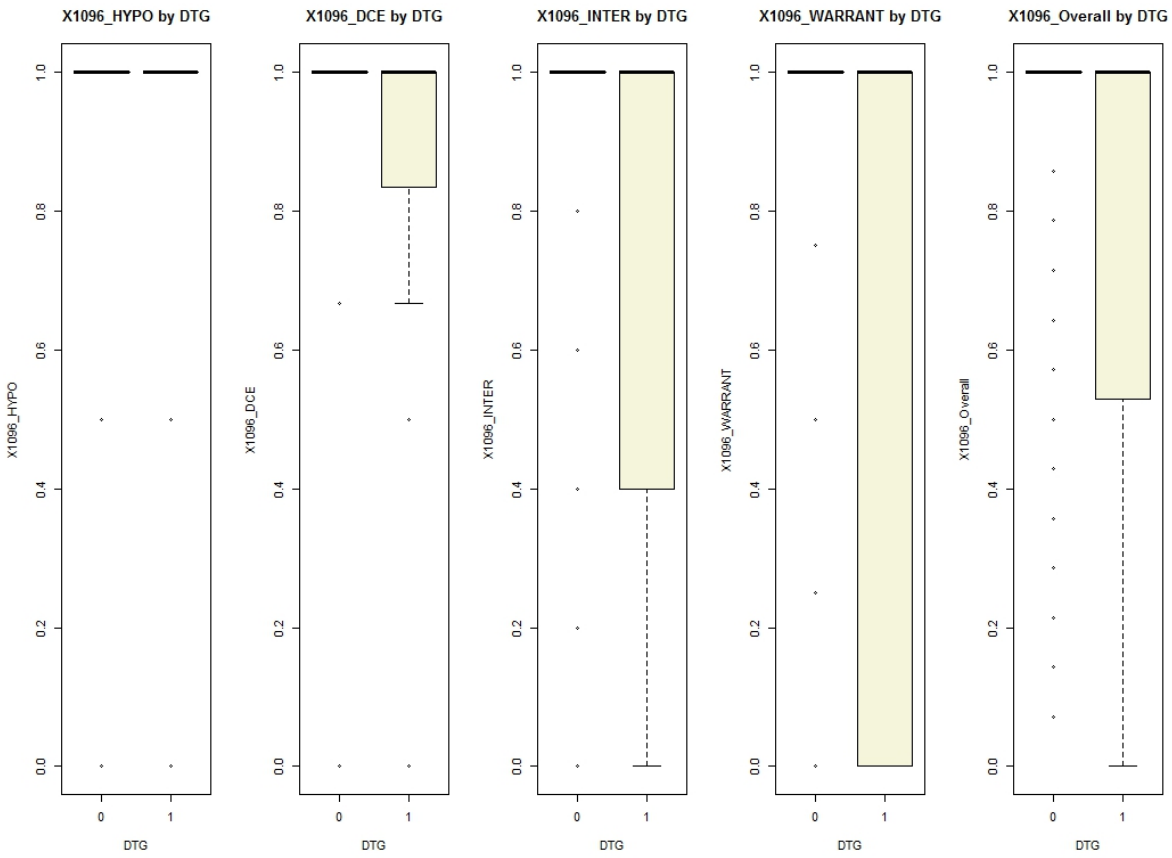

Figure 3. Performance across subsections and overall performance in the Amount of Ice and Melting Point task by DTG

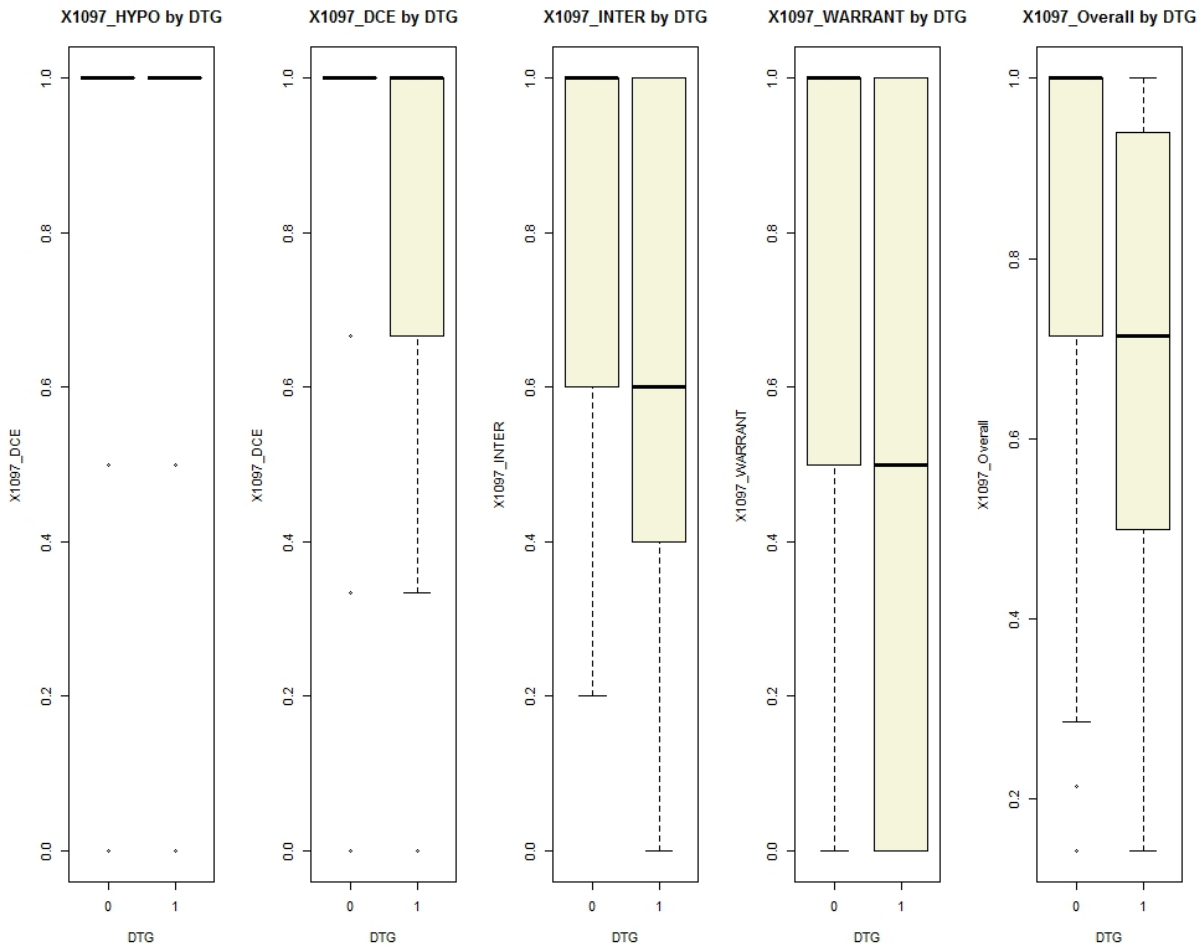

Figure 4. Performance across subsections and overall performance in the Size of Container and Boiling Point task by DTG

Another trend that can be seen in the above figures is how students who demonstrated DTG behavior had worse performance in each activity than students who didn't demonstrate DTG. This is shown in the lower distributions for DTG students (Figures 1-4) when compared to the distributions for non-DTG students. Also of note is that the data for the third activity (Figure 3), amount of ice and melting point, appears to be more erratic than the other activities, with the data for non-DTG students not increasing at all over time and the data for the DTG students having a larger amount of outliers. Despite this, it still follows the trend of performance degrading across subsections for students exhibiting DTG behavior.

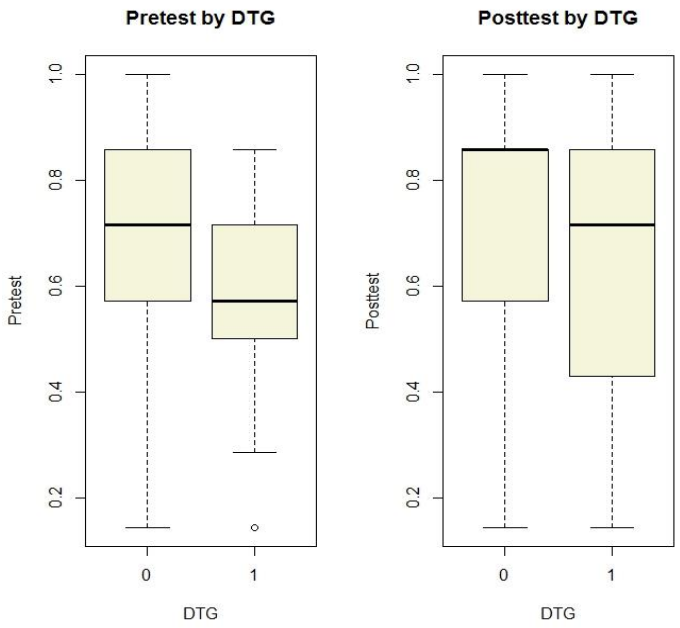

Figure 5. Performance in the pretest and performance in the posttest by DTG

Students who demonstrated DTG behavior also performed worse on both the pretest and the posttest than students who did not show DTG behavior (Figure 5). This fits with the negative correlation we found between the demonstration of DTG behavior and performance across all activities and subsections.

Table 3: Residuals for Overall Scores on Heat and Boiling Point Activity(X1063) for DTG and Pretest

Residuals					
Min	**1Q**	**Median**	**3Q**	**Max**	
-0.745745	-0.00122	0.05424	0.16514	0.28769	
Coefficients					
	Estimate Std.	**Error**	**t value**	**Pr(>|t|)**	**Signif. Level**
(Intercept)	0.61305	0.07031	8.719	1.38e-14	0
DTG	-0.06710	0.05949	-1.128	0.261455	NA
Pretest	0.38817	0.09716	3.995	0.000109	0
Residual standard error: 0.2358 on 126 degrees of freedom					
(28 observations deleted due to missing data)					
Multiple R-Squared: 0.1334			**Adjusted R-Squared:** 0.1196		
F-statistic: 9.698 on 2 and 126 DF			**p-value**: 0.0001209		

Table 4: Residuals for Overall Scores on Amount of Ice and Boiling Point(X1077) for DTG and Pretest					
Residuals					
Min	**1Q**	**Median**	**3Q**	**Max**	
-0.95902	-0.09712	0.09019	0.16535	0.28701	
Coefficients					
	Estimate Std.	**Error**	**t value**	**Pr(>|t|)**	**Signif. Level**
(Intercept)	0.61459	0.07786	7.894	1.23e-12	0
DTG	-0.04818	0.06587	-0.731	0.46583	NA
Pretest	0.34443	0.10578	3.202	0.00173	0.001
Residual standard error: 0.2611 on 126 degrees of freedom					
(28 observations deleted due to missing data)					
Multiple R-Squared: 0.08673			**Adjusted R-Squared:** 0.07224		
F-statistic: 5.983 on 2 and 126 DF			**p-value**: 0.003294		

Table 5: Residuals for Overall Scores on Amount of Ice and Melting Point(X1096) for DTG and Pretest					
Residuals					
Min	**1Q**	**Median**	**3Q**	**Max**	
-0.86959	-0.01244	0.04073	0.14709	0.25478	
Coefficients					
	Estimate Std.	**Error**	**t value**	**Pr(>|t|)**	**Signif. Level**
(Intercept)	0.64020	0.6809	9.402	3.13e-16	0
DTG	-0.05452	0.05761	-0.945	0.345773	NA
Pretest	0.37224	0.09409	3.956	0.000126	0
Residual standard error: 0.2283 on 126 degrees of freedom					
(28 observations deleted due to missing data)					
Multiple R-Squared: 0.1275			**Adjusted R-Squared:** 0.1137		
F-statistic: 9.207 on 2 and 126 DF			**p-value**: 0.0001854		

Table 6: Residuals for Overall Scores on Container Size and Boiling Point(X1097) for DTG and Pretest

Residuals					
Min	**1Q**	**Median**	**3Q**	**Max**	
-0.64312	-0.11411	0.08676	0.12918	0.34153	
Coefficients					
	Estimate Std.	**Error**	**t value**	**Pr(>|t|)**	**Signif. Level**
(Intercept)	0.65871	0.06561	10.039	<2e-16	0
DTG	-0.09266	0.05551	-1.669	0.09754	0.1
Pretest	0.29695	0.09067	3.275	0.00136	0.001
Residual standard error: 0.22 on 126 degrees of freedom					
(28 observations deleted due to missing data)					
Multiple R-Squared: 0.1121			**Adjusted R-Squared:** 0.09798		
F-statistic: 7.952 on 2 and 126 DF			**p-value**: 0.0005593		

With our data, we also performed regressions for the overall score of each activity, the student's total DTG, and pretest scores in order to look at relationships between pretest scores and performance, as well as between performance and DTG (Tables 3-6). We expected the pretest scores to correlate with the scores in each activity. We were focused on looking at the significance between the relationship of the scores and DTG, in order to determine whether they were correlated. The only significant finding was found for activity X1097 (Table 6), with a significance at the 0.1 level of alpha, supporting evidence to reject the null hypothesis that these two (DTG and X1097 overall score) are unrelated by random chance. Upon reviewing the frequency of DTG on overall scores for activity 4, we can see that the students with DTG on this activity showed the lowest median scores. This result has some potential for importance, but requires further investigation.

Discussion

In discussing the use of this DTG detector, it is important to note its successful use in this study. This DTG detector has demonstrated that it can function accurately in that it can distinguish between instances of DTG and non-DTG behavior approximately 86.9% of the time. This is a rate that is considerably above chance and helps confirm the detector's ability to correctly identify DTG behavior. This DTG detector can dependably be used for further research involving behaviors demonstrating disengagement from the task goal, and is an effective tool in looking at disengagement.

From the information gathered using the DTG detector and data collected from the Phase Change microworld, several interesting points involving DTG behaviors arose. The negative correlation between performance in each task and DTG behaviors fits with prior research on disengagement and its relationship with performance (Salamonson, Andrew, & Everett, 2009). Students who are disengaged when learning will struggle with understanding the material and will perform worse due to the less effective learning they would experience than if they were fully engaged in the material and the tasks. Those who showed DTG behavior were found to perform worse than those who didn't. This was found in the various subsections of each activity, overall performance in each activity, and also in both the pretest and the posttest. The poorer performance in the pretest may mean that there is also reason to believe that students who were already performing worse would then start to disengage and exhibit DTG behaviors, rather than performing worse because they were disengaged. Both are reasonable interpretations from these data, and a more in depth look into each will likely be fruitful.

Going back to the erratic data that came up within the third activity, a potential cause may be due to the ordering of the activities. The first activity, heat and *boiling* point, is followed

up by the amount of ice and *boiling* point activity. However, these are followed by the amount of ice and *melting* point activity, which is then followed up with the size of container and *boiling* point activity. Of the four activities, the third activity is the only one that deviates from the dependent variable of the boiling point and instead focuses on melting point. This may have thrown students off, compared to the two prior tasks, which both have boiling point as the dependent variable. A possible strategyto test whether this interpretation is accurate would be to modify the task so that it better orients students to the change in the dependent variable.

Looking at each activity separately, it was found that performance was worse as students progressed through each subsection. This trend was especially pronounced for students who exhibited DTG behaviors. A possible reason for this is that failures would compound as students progressed. If a student failed to understand earlier concepts or did not understand what to do early on, as they continued, their confusion increased due to the addition of new ideas and tasks that built off of the prior work.

The importance of scaffolding engagement is a critical consideration, especially within virtual learning environments and learning software (Reiser, 2004; Gobert et al, 2015). For DTG students specifically, when they start to engage in DTG behaviors, they start to learn less effectively, missing crucial information that could cause them to do worse on learning. If this is the case, it is important to detect disengaged behavior early on in order to effectively scaffold the student to get them back on track before it exacerbates further disengaged learning. Additionally, it is possible that by scaffolding the student to re-engage in the task, the student's learning would increase. This, of course, is an empirical question.

Implications For Future Work

This DTG detector has been shown to be a useful and operational tool in identifying disengaged behavior in Inq-ITS. It could readily be applied to future research involving engagement and disengagement as a way to measure a level of disengagement accurately and unobtrusively. Future use of the DTG detector in these research areas would also allow for a quantifiable measure of disengagement that provides more opportunity for statistical analysis while also working well with observation as supplemental data. It also provides the basis upon which students could be scaffolded in order to re-engage them in inquiry; this is one of the goals of the Inq-ITS project (Gobert et al, 2013). It is an empirical question as to whether scaffolding students who are disengaged could lead them to re-engage in learning.

The DTG detector focuses on a specific kind of disengagement, so combining it with other detectors that can cover different kinds of disengagement may be a way to get a broader view of disengagement. One other kind of disengagement that could be detected along with DTG would be carelessness. Carelessness is a kind of disengagement where the student starts to make careless errors *after* having achieved mastery rather than being due to a lack of knowledge about the subject matter. Recently, carelessness has been studied with the use of extractable data based on log files of the student learning , as well as the recent methods of student modeling and educational data mining (Gobert, Baker & Wixon, 2015).

Implementing detectors of carelessness would permit studying the relationships between carelessness and the student learning goals (Baker, Corbett & Aleven, 2008). A carelessness detector has been developed to determine carelessness through implementing an automated model that calculates the contextual probability that an error is due to carelessness (Hershkovitz et al, in press). This model is a modified version of the Bayesian Knowledge Tracing model that

does not assume that the probability of an error due to carelessness will always be the same for a specific skill. Future research into the application of both detectors and the relationship between these different levels of disengagement should be a good step towards being able to observe and measure various forms of disengagement, providing a method of unobtrusively gaining an in-depth look at learner disengagement and knowledge that could be used towards developing more engaging methods and tools in virtual learning environments.

References

Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. Psychology in the Schools, 45(5), 369-386.

Baker, R.S.J.d., Corbett, A.T., Aleven, V. (2008) Improving Contextual Models of Guessing and Slipping with a Truncated Training Set. *Proceeding of the 1st International Conference on Educational Data Mining,* 67-76.

Baker, R.S.J.d., Corbett, A.T., Gowda, S.M., Wagner, A.Z., MacLaren, B.M., Kauffman, L.R., Mitchell, A.P., Giguere S. (2010) Contextual Slip and Prediction of Student Performance After Use of an Intelligent Tutor. *Proceedings of the18th Annual Conference on User Modeling, Adaptation, and Personalization,* 52-63.

Baker, R. S., & De Carvalho, A. M. J. A. (2008, June). Labeling Student Behavior Faster and More Precisely With Text Replays. In *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 38-47).

Biggs, J., Kember, D., & Leung, D. Y. (2001). The revised two-factor study process questionnaire: R-SPQ-2F. *British Journal of Educational Psychology*,*71*(1), 133-149.

Carroll, J. (1963). A model of school learning. The Teachers College Record, 64(8), 723-723.

Chen, P. S. D., Guidry, K. R., & Lambert, A. D., (2010). Engaging online learners: The impact of Web-based learning technology on college student engagement. Computers & Education, 54(4), 1222-1232.

Cleveland-Innes, M., & Garrison, D. R. (2005). Facilitating cognitive presence in online learning: Interaction is not enough. The American Journal of Distance Education, 19(3), 133-148.

Corno, L., & Mandinach, E. B. (1983). The role of cognitive engagement in classroom learning and motivation. *Educational psychologist*, *18*(2), 88-108.

Csikszentmihalyi, M. (1991). *Flow: The Psychology of Optimal Experience*. New York: Harper Perennial.

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

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

Garrison, Randy D., Cleveland-Innes Martha. (2010). Facilitating Cognitive Presence in Online Learning: Interaction is Not Enough. American Journal of Distance Education, 19(3), 133-148.

Gobert, J., Baker, R. S., & Wixon, M. (2015). Operationalizing and Detecting Disengagement During On-Line Science Inquiry. *Educational Psychologist, 50(1),* 43-57.

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

Herrington, Jan, & Reeves, Thomas C.. (2003). Patterns of engagement in authentic online learning environments. 19(1), 59-71.

Hershkovitz, A., Gobert, J., Baker, R. S., Sao Pedro, M., & Wixon, M. (in press). Goal Orientation and Carelessness in Computer-Based Science Inquiry.

Hullinger, H., & Robinson, C. C. (2008). New benchmarks in higher education: Student engagement in online learning. Journal of Education for Business, 84(2), 101-109.

Karweit, N., Slavin, R.E. Time-On-Task: Issues of Timing, Sampling, and Definition. Journal of Experimental Psychology, 74 (6) (1982), 844-851.

Lahaderne, H.M. Attitudinal and Intellectual Correlates of Attention: A Study of Four Sixth-Grade Classrooms. Journal of Educational Psychology, 59(5) (1968), 320-324.

Discovery and Data Mining (KDD 2006), 935-940. New York, NY: ACM Press.

Marx, Jeffrey D. & Cummings, Karen. (2007). Normalized change. *American Journal of Physics*, 75(1) (2007), 87-91.

Newby, T., & Richardson, J. C. (2006). The role of students' cognitive engagement in online learning. The American Journal of Distance Education, 20(1), 23-37.

Nolen, S. B. (2003). Learning environment, motivation, and achievement in high school science. *Journal of Research in Science Teaching*, *40*(4), 347-368.

Pearce, Jon, Mary Ainley, and Steve Howard. (2005) The ebb and flow of online learning. Computers in Human Behavior, 21(5), 745-771.

Piccoli, G., Ahmad, R., & Ives, B. (2001). Web-based virtual learning environments: A research framework and a preliminary assessment of effectiveness in basic IT skills training. *MIS quarterly*, 401-426.

Reiser, B. J. (2004). Scaffolding Complex Learning: The Mechanisms of Structuring and Problematizing Student Work. *The Journal of the Learning Sciences*, 13(3), 273-304.

Salamonson, Y., Andrew, S., & Everett, B. (2009). Academic Engagement and Disengagement as Predictors of Performance in Pathophysiology Among Nursing Students. *Contemporary Nurse*, 32(1-2), 123-132.

Skinner, E. A., & Pitzer, J. R. (2012). Developmental dynamics of student engagement, coping, and everyday resilience. In *Handbook of research on student engagement* (pp. 21-44). Springer US.

Wixon, M., d Baker, R. S., Gobert, J. D., Ocumpaugh, J., & Bachmann, M. (2012). WTF? detecting students who are conducting inquiry without thinking fastidiously. In User Modeling, Adaptation, and Personalization (pp. 286-296). Springer Berlin Heidelberg.

Wixon, M. (2013). Detecting students who are conducting inquiry Without Thinking Fastidiously (WTF) in the Context of Microworld Learning Environments (Doctoral dissertation, Worcester Polytechnic Institute).