Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems microworld

Janice D. Gobert, Yoon Jeon Kim, Michael Sao Pedro, Michael Kennedy, Cameron Betts

Abstract. Many national policy documents underscore the importance of 21st century skills, including critical thinking. In parallel, recent American frameworks for K-12 Science education call for the development of critical thinking skills in science, also referred to as science inquiry skills/practices. Assessment of these skills is necessary, as indicated in policy documents; however, this has posed a great challenge for assessment researchers. Recently, some science learning environments seek to assess these science skills. These systems log all students' interactions within the given system, and if fully leveraged, these logs provide rich assessments of inquiry skills. Here we describe our environment Inq-ITS (**I**nquiry **I**ntelligent **T**utoring **S**ystem), that uses Educational Data Mining to assess science inquiry skills, as described as 21st century skills. Additionally, here we describe how we measure students' skills at designing controlled experiments, a lynchpin skill of inquiry, in the context of complex systems. In doing so, our work addresses 21st century skill assessment in two ways, namely of inquiry (designing and conducting experiments), and in the context of complex systems, a key topic area of 21st century skills. We use educational data mining to develop our assessment of this skill for complex systems.

Keywords: complex systems, inquiry assessment, performance assessment, educational data mining, 21st century skills

Gobert, J.D., Kim, Y.J, Sao Pedro, M.A., Kennedy, M., & Betts, C.G. (2015). Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems microworld. Thinking Skills and Creativity, 18, 81-90.

1. Introduction

1.1. Background

Following the launching of Sputnik in October of 1957, policy makers in the United States began to question the quality of science instruction in schools, which, in turn, instantiated a call for change in all science curricula. Post-Sputnik, educators and policy makers sought that science literacy should include science content knowledge, inquiry skills, and understanding of the nature of science (Perkins, 1986). Secondly, Post-Sputnik reform efforts also called for educating the broad populace rather than the top 10% of high achieving students. Taken together, the goal was and continues to be to develop a citizenry with knowledge and skills so that they can participate fully in a democracy (Stokes, 1997).

In more recent reports, policy makers continue to emphasize the need for 21st century skills (NRC, 2011; Partnership for 21st Century Skills, 2007). In brief, 21st century skills broadly include: *cognitive knowledge/skills* (e.g., critical thinking), *interpersonal skills* (e.g., communication and teamwork skills), and *intrapersonal skills* (e.g., metacognitive/motivational, self-regulated learning (Partnership for 21st Century Skills, 2007). Twenty-first century skills predict both college grades and future employment success, and as technological advancements continue, people will be increasingly expected to think in creative and divergent ways (Lai & Viering, 2012). Lastly, 21st century skills are acknowledged as important for developing innovative thinkers (Sternberg, 2006; Sternberg & Lubart, 1991, 1995; Sawyer, 2006), necessary for a knowledge-based economy Bereiter, 2002; Resnick, 2007).

In the present work, we focus on the *cognitive components* of 21st century skills, which include: critical thinking, non-routine problem-solving, and systems-thinking. Specifically, here we assess inquiry skills, critical thinking in science, in the context of complex systems (cf., Hmelo-Silver, 2007; Jacobson & Wilensky, 2006; Yoon, 2008). In other work, we address intrapersonal skills, namely engagement (Gobert, Baker, & Wixon, 2015).

1.2. Traditional Educational Assessments

The purpose of educational assessments, broadly described, is to make inferences about students' knowledge and skills. Traditionally, as in the case of science, formal assessment is done on the basis of standardized tests, which use multiple-choice items to determine the level of proficiency a student has achieved. Items are developed using standards, for example, state content standards; these tests are criterion-referenced in that they are intended to measure students in terms of their level of mastery on grade-appropriate knowledge and skills. These tests are also norm-referenced in that they compare students relative to their peers. These tests are typically implemented using paper and pencil format and multiple-choice items (Anastasi & Urbina, 2009).

However, given the richness of critical thinking involved in science inquiry, it has been acknowledged that typical science achievement tests *do not* adequately reflect the complex science knowledge and inquiry process skills that are important components of scientific literacy or of 21st century skills (NSES, 1996; National Assessment of Educational Progress, 2004; Haertel, Lash, Javitz, & Quellmalz, 2006; Quellmalz & Haertel, 2004; Quellmalz, Kreikmeier, DeBarger, & Haertel, 2007; Clarke-Midura, et al, 2011; Leighton & Gierl, 2011). As discussed elsewhere (Gobert et al, 2013), the limitations of these tests are partly due to the simplified conceptions of the nature of science understanding at the time that the tests were designed (diCerbo & Behrens 2012; Mislevy et al, 2012). Thus, more recently, it has been widely acknowledged that multiple choice items are not suitable means to assess rich inquiry skills, and instead, tasks need to be designed to elicit data that can address what students know and how they use their knowledge, rather than elicit data that we can easily collect and analyze (Pellegrino, 2009). In doing so, one can assess both the products and processes of inquiry (Rupp et al., 2010).

In short, the problem becomes: how do we use policy documents about critical thinking in science (NRC, 2011; Partnership for 21st Century Skills, 2007) use to inform the design and development of valid, reliable assessments of rich inquiry skills? (Leighton & Gierl, 2011). Furthermore, specific to this paper, we address how to do this

type of assessment in the context of complex systems, a key topic area of 21st century thinking.

1.3. Inq-ITS (Inquiry Intelligent Tutoring System)

Our design work started with the specifications for what knowledge and skills students should possess (NGAA, 2013) in order to develop a system that could provide fine-grained assessment data on students' science inquiry skills. Our environment, Inq-ITS (http://slinq.org) is a rigorous, technology-based learning environment that assesses and scaffolds middle school students in Earth, Life, and Physical Science during learning. Our work recognizes that these environments can provide a more fertile basis upon which to develop performance-based assessments by leveraging computational techniques to analyze students' log files of their inquiry processes (Gobert et al, 2012, 2013).

Inq-ITS uses microworlds (Papert, 1980) to engage students in inquiry. Microworlds are computerized representations of real-world phenomena whose properties can be inspected and changed (Pea & Kurland, 1984; Resnick, 1997). Since microworlds share many features with real apparati (Gobert, 2005; in press), they provide greater authenticity for "doing science". In turn, microworlds afford authentic performance assessment of inquiry skills because with a microworld in Inq-ITS, students can generate a hypothesis, test it, interpret data, warrant their claims with data, and communicate findings with regard to what they discover. These inquiry tasks reflect the national frameworks for inquiry (NSES, 1996; NRC, 2011), and represent the critical thinking skills used to reason logically about scientific concepts as reflected in 21st century skills documents (Partnership for 21st Century Skills, 2007).

In terms of assessment techniques, we employ techniques that originate from Educational Data Mining (EDM henceforth; cf., Baker & Yacef, 2009; Romero & Ventura, 2010), which grew from computer science, human-computer interaction, and measurement. EDM broadly described, is a set of powerful methods for analyzing patterns in educational data. It has been used for a variety of goals: to compare the efficacy of interventions (cf., Beck & Mostow, 2008; Chi, VanLehn, &

Litman, 2010), to refine domain knowledge models (Cen, Koedinger, & Junker, 2008; Pavlik et al., 2009; Desmarais, Meshkinfam, & Gagnon, 2006), to build automated detectors of relevant constructs during student learning (Baker et al., 2008; Cetintas et al., 2010; Gobert, Baker, & Wixon, 2015; Hershkovitz, Wixon, Baker, Gobert, & Sao Pedro, 2011), and to do both formative and performance assessment (Mislevy et al., 2012; Gobert et al., 2012).

Educational data mining can be a powerful method; however in order to inform pedagogy and assessment of inquiry, data mining needs to be guided by theoretical principles about students' inquiry learning (Gobert, in press). EDM, particularly exploratory data mining, on the face of it, appears to be distinct from the top-down, forward-design processes used in the psychometric community (Mislevy et al., 2012) in which design principles are derived exclusively from theoretical principles. In fact, elsewhere, we articulate how evidence-centered design, a rigorous and detailed framework for assessment design, was used in our system (Gobert et al., 2012). Here, we argue that our approach, which is both top-down and bottom up, can lead to valid metrics for developing of assessment models. Specifically, here, top-down processes are used to guide the development of categories for hand tagging, and bottom-up processes, namely, machine learning (aka Educational Data Mining) are then used to predict hand tagging.

Here we address a key skill of inquiry, namely, designing controlled experiments, a lynch pin skill of inquiry. This skill is commonly referred to as the control for variables strategy (cf., Chen & Klahr, 1999). Of all of the skills underlying inquiry, this one is particularly difficult for students: students may gather insufficient evidence to test hypotheses (Shute & Glaser, 1990; Schauble, Glaser et al., 1991), may run only one trial (Kuhn, Schauble, Garcia-Mila, 1992) or run the same trial repeatedly (Kuhn, Schauble & Garcia-Mila, 1992; Buckley, Gobert & Horwitz, 2006). They also change too many variables at once (Glaser et al., 1992; Reimann, 1991; Tschirgi, 1980; Shute & Glaser, 1990; Kuhn, 2005; Schunn & Anderson, 1998, 1999; Harrison & Schunn, 2004; McElhaney & Linn, 2008, 2010). They also run experiments that try to achieve an outcome (i.e., make something burn as quickly as possible) or design experiments that are enjoyable to execute or watch

(White, 1993), as opposed to testing a hypothesis (Schauble, Klopfer & Raghavan, 1991; Schauble, Glaser, Duschl, Schulze & John, 1995; Njoo & de Jong, 1993).

Having successfully developed detectors for this skill for Physical science topics (Sao Pedro, Baker, Gobert, Montalvo, & Nakama, 2013; Sao Pedro et al, 2012), we conduct our assessment development in the area of Complex Systems, also referred to as Systems Thinking, a key aspect of 21st Century science knowledge (Lai & Viering, 2012). Our ecosystems microworld targets students' understanding of the ways in which organisms interact and have different functions within an ecosystems to enable survival (Sao Pedro, Gobert, & Betts, 2014). Since the ecosystems environment has multiple variables interconnected in a non-linear fashion (Yoon, 2008; Greiff, Wustenberg, & Funke, 2012), the hypothesis space increases (Klahr & Dunbar, 1988), and the understanding of the effects the independent variables on dependent variables(s) is more challenging because, as previously stated, the simple control for variables strategy (cf., Chen & Klahr, 1999), desribed above, cannot be applied in a straightforward manner. The complexity that arises here is illustrated when contrasted to the application of this skill in Physical Science topics. Specifically, in Physics phenomena (at the middle school level) there is ***one*** independent and ***one*** dependent variable (ivs and dvs) underlying the causal system. Many Life Science topics, by contrast, are inherently different from Physical Science because the former have a number of interconnected, non-linear elements that are interacting in a complex causal system (Yoon, 2008; Jacobson & Wilensky, 2006), as in topics like Ecosystems and Cell functions.

In brief, students have difficulties with complex systems because students view relationships between variables as univariate, simple, and direct (Grotzer & Perkins, 2000; Grotzer & Bell-Basca, 2003). Additionally, there are many emergent properties that are not predictable from the behavior of individual parts (Wilensky & Resnick 1999), and students favor explanations that assume central control and deterministic causality (Resnick & Wilensky, 1993), rather than thinking about the interconnectedness of multiple variables. In terms of conducting inquiry, an important implication that impacts students' difficulty is that the control of variables strategy (cf., Chen & Klahr, 1999) no longer works in it is simple form (Bachmann et al., 2010) because of

the multiple interacting independent variables, i.e., where variables Variable 1 and Variable 2 interact, changing Variable 1 and keeping all else fixed will yield different results depending on the value at which Variable 2 is fixed. This is extremely difficult for students to understand (Hmelo-Silver et al, 2007; Wilensky & Resnick, 1999; Yoon, 2008). These complexities cause a challenge to middle school students both in understanding complex systems and in conducting inquiry in complex systems (Hmelo-Silver et al, 2007); as a corollary of these, students' inquiry strategies are also difficult to measure.

In our microworld, students are said to demonstrate the skill of designing controlled experiments when they generate trials that make it possible to infer how changeable factors (e.g., seaweed, shrimp, small fish, and large fish within an Ecosystem) affect outcomes (e.g., the overall balance of the ecosystem) (Sao Pedro, Baker, Gobert, Montalvo, & Nakama, 2013). This skill relates to application of the Control of Variables Strategy (CVS; cf., Chen & Klahr, 1999), but unlike CVS, it takes into consideration *all* the experimental design setups run with the simulation, not just isolated, sequential pairs of trials (Gobert, Sao Pedro, Baker, Toto, & Montalvo, 2012; Sao Pedro, Baker, Gobert, Montalvo, & Nakama, 2013).

In this paper, we aim to demonstrate how data-mining algorithms can be developed to assess students' science inquiry skills (namely, designing and conducting experiments) in the context of complex systems. This is a well-acknowledged assessment challenge since this inquiry skill is ill-defined, i.e., there are many ways (both correct and incorrect) that students go about designing and conducting experiments (Kuhn, 2005). Specifically, we discuss the development and evaluation of a data-mined model that classifies the students who are demonstrating designing controlled experiments skill (vs. those who are not demonstrating this skill) in a simulation of a complex system.

3. Method

3.1. Participants

101 eight graders at a Central Massachusetts middle school participated in this study. The teachers used the Life Science microworld during their regular science classes after students learned about food webs. Each student had access to an individual computer to engage in the mi-

crowold. 53% of the participants were female students, and the average age of the all participants was 15.67 (SD = 1.32).

3.2. Materials

Inq-ITS (Gobert, et al., 2012, 2013) is a web-based environment in which students conduct inquiry with interactive simulations and inquiry support tools. The simulations are designed to assess inquiry in content areas aligned to middle school Physical, Life, and Earth Science as described in the NGSS standards (NGSS Lead States, 2013). Each Inq-ITS activity provides students a driving question and requires them to investigate that question using the simulation and tools (see Figure 1 for an example Ecosystems activity). Students make hypotheses, collect data by changing the simulation's variables and running trials, analyze their data, warrant their claims, and communicate their findings. A key aspect of Inq-ITS is that activities are performance-based assessments of inquiry skills. Metrics on students' skills are derived from the processes they follow while conducting inquiry and the work products (Rupp et al., 2010) they create with the support tools.

3.3. Microworld and Inquiry Scenarios

The students engaged in inquiry within Inq-ITS environment (Gobert et al., 2012, 2013) using the EcoLife simulation. The EcoLife simulation (Figure 1) is an aquatic ecosystem containing big fish, small fish, shrimp, and seaweed where students conduct inquiry about how the populations of producers, consumers, and decomposers are interrelated. The microworld consists of two inquiry scenarios. In the first, students were asked to stabilize the ecosystem. In the second, students were asked to stabilize the shrimp population (or alternatively, ensure that the shrimp population is at its highest). For each scenario, students form a hypothesis, collect data by changing the population of a selected organism (on the left side of Figure 1), analyze data by examining automatically generated data tables and population graphs (on the right side of Figure 1), and communicate findings by completing a brief lab report.

This microworld addresses the two strands of the Massachusetts Curricular Frameworks: (1) the functions of organisms and the ways in which organisms interact within an ecosystem that enable the ecosystem to survive and (2) the roles and relationships among producers,

consumers, and decomposers in the process of energy transfer in a food web.

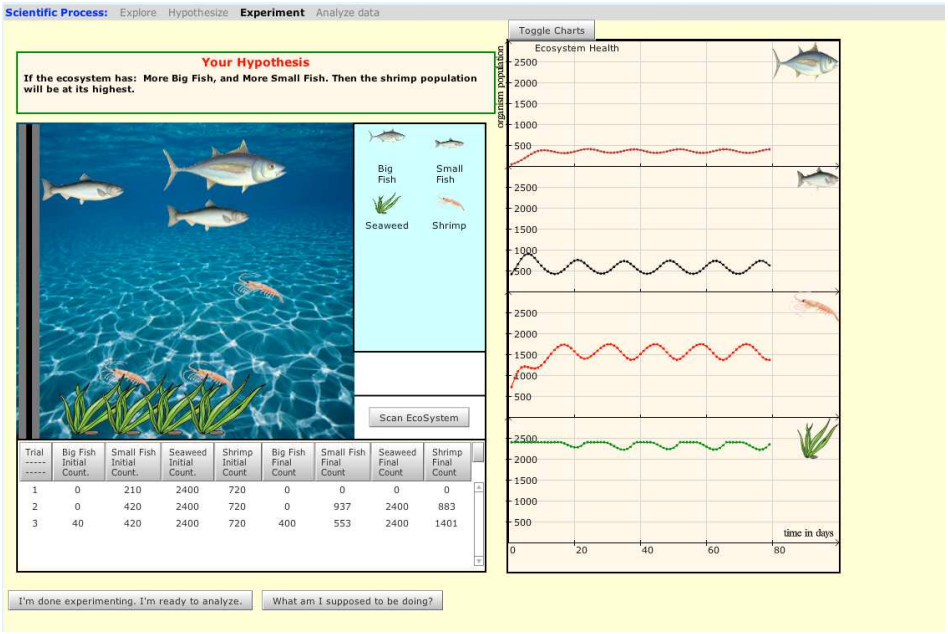

Figure 1. EcoLife design and conduct an experiment stage. Here, students add and remove organisms and scan the ecosystem to investigate how the population changes over time.

4. Data Analysis

4.1. Hand-Scored Classification

With log data from 101 students, we carried out text replay tagging (Baker, Corbett, & Wagner, 2006) to classify students who demonstrated the skill of designing controlled experiments from students who didn't demonstrate the skill. This classification yields a label variable (i.e., skill demonstration vs. no skill demonstration) that can be later used for supervised machine learning of the model. In text replay tagging, human coders are presented "pretty-printed" versions of log files

(i.e., clips), that contain textual sequences of low-level student actions,

then coders assign one or more tags (e.g., designing controlled experiments) per clip. For the EcoLife microworld, the grain-size of a clip contains all actions associated with formulating hypotheses (e.g., selecting shrimp population as independent variable) and all actions associated with designing and running experiments (e.g., increasing the population of shrimp). After producing these classifications, each student's activity sequences were summarized by creating a feature set from the data, which was later used to generate a machine-learned detector that can categorize who is demonstrating the skill of interest (Sao Pedro et al., 2013). There are several advantages of using machine-learned detectors (Sao Pedro, 2013). First, such models can capture relationships that people cannot easily specify while leveraging the human coders' ability to recognize demonstration of skill. Second, as machine learning approaches use standard methods for predicting how well models will generalize to new data (e.g., cross-validation), accuracy and generalizability of machine-learned models can be easily verified.

Two coders participated in the hand-scoring of the clips. One coder hand-scored all the clips, and a second coder coded the first 50 clips to compute inter-rater reliability. A kappa of .71 was obtained between the two coders for these first 50 clips. This kappa was considered to be adequate and commensurate with coding such data in our prior work (Sao Pedro et al., 2013). Within the corpus of tagged clips, 52.2% of students had demonstrated the skill of controlling for variables.

4.2. Feature Distillation

To build a data minded model (or detector) for designing controlled experiments that predicts the hand-coded labels of whether or not students demonstrate this skill when collecting data, we then distilled certain features from the log files to use as predictors of the detector. Initially, we identified and extracted 73 features that were based on earlier literature on students' inquiry (e.g., Buckley et al., 2006; Kuhn et al., 1992; Chen & Klahr, 1999). In our earlier work, we further refined these features by iteratively testing how varying configurations of these

features contribute to model performance, and selected 11 features that have good generalizability and construct validity based on literature review of indicators that are associated with science inquiry (See Sao Pedro et al., 2012 and Gobert et al., 2013 for detailed discussion of this process).For the current study, we also used these 11 features to build a detector. We briefly describe each feature as follows:

1. All actions count: This is a count of all low-level actions found in a clip including all actions in the hypothesize and experiment phases of inquiry. These actions include: changing variables when making hypotheses; proposing hypotheses; running, pausing or resetting the simulation; changing values of simulation variables when designing experiments; and displaying or hiding the data table and hypothesis list from the simulation interface.
2. Complete trials count: The number of trials in which the student ran the simulation to completion (i.e., without restarting the trial).
3. Total trials count: The total number of trials started within the clip, regardless of whether the student allowed the simulation to run to completion.
4. Simulation pause count: The number of times the simulation was paused.
5. Simulation variable changes count: The number of times the values of simulation variables were changed while the student was designing experiments.
6. Simulation variable changes count related to stated hypotheses: The number of times the values of simulation variables explicitly stated in hypotheses were changed.
7. Number of pairwise repeated trials: A count of the pairs of trials that had identical experimental setups. This count considers any two trials in the entire clip.
8. Number of successive repeated trials: The same as the pairwise count, except that only adjacent (successive) trials (e.g., between Trials 2 and 3, between Trials 4 and 5) are considered.
9. Number of pairwise controlled trials, with repeats: A count of the pairs of trials in which exactly one simulation variable (independent variable) had different values between trials, and all other variable values were identical (cf., Chen & Klahr, 1999). Because it is a pairwise count, any pair of trials is considered. Furthermore, if any trial is a repeat of an earlier trial, it is still considered in this count.

10. Number of successive controlled trials, with repeats: Same as the pairwise controlled trial count, except that this count only considers successive trials.
11. Number of pairwise controlled trials, ignoring repeats: Same as the pairwise controlled count previously mentioned, except that if a trial is a repeat of an earlier trial, it is not considered.

4.3. Detector Generation and Validation

Continuing with the EDM-based method used in our group (Sao Pedro et al. 2013; Gobert et al. 2013), machine-learned detectors were developed using the hand-coded clips (i.e., label variable) and the 11 features distilled from students' log data (i.e., predictor variables) within EcoLife using RapidMiner 6.3 (Mierswa, Wurst, Klinkenberg, Scholz, & Euler, 2006). We used J48 decision trees algorithm with automated pruning as method to generate the detector. J48 decision tree algorithm is an open-source implementation of the C4.5 decision tree algorithm (Quinlan, 1993), and it has been widely used to detect behaviors in technology-enhanced learning environments (e.g., Baker & de Carvalho, 2008). J48 decision trees are particularly good at reducing over-fitting (i.e., the model is fitting to noise rather than the underlying relationship) as it uses a post hoc pruning approach that reduces tree complexity (Quinlan, 1993). That is, the pruning process removes nodes of the decision tree that does not provide significant information, which yields a comprehensible decision tree without unnecessary complexity.

The J48 decision tree has two parameters that we can control: minimum number of instances per leaf (M) and the confidence threshold for pruning (C). In our previous work (Sao Pedro et al., 2013), we set these values at 2 for M and .25 for C (which are the default values for this algorithm). For the current study, we set the confidence thresh-

old at .25, and the minimum number of instances per leaf was put at 10 to yield a parsimonious tree that can is more generalizable. This setting was selected to about 5% of the data points available. To further minimize possible over-fitting, six-fold cross-validation was conducted at the student level, meaning that detectors were trained on five randomly selected groups of students and tested on a sixth group of students. By cross-validating at this level, we can increase confidence that detectors will be accurate for new groups of students. We chose this technique for the following reasons. J48 decision trees have led to successful behavior detectors in previous research (e.g. Walonoski & Heffernan, 2006; Baker & de Carvalho, 2008; Sao Pedro et al., 2013). Also, decision trees produce relatively human-interpretable models (i.e., attributes and associated rules). For example, as depicted in Figure 2, each node is essentially a feature and the value associated with it that can be used to classify which incident is demonstrating designing for controlled experiments. This model in turn can be used to assess student behavior or integrate within the existing learning environments to update student model real-time (Mislevy, Behrens, Dicerbo, & Levy, 2012).

5. Results

The confusion matrix (Table 1) captures raw agreement between the detector's prediction and the human coders' tags under stu-

dent-level cross-validation. For example, the first column of the confusion matrix ("Hand-coded Positive") indicates that among 118 hand-coded clips labeled as demonstrating designing controlled experiments skill, the machine learned detector also classifies them as the case while 7 cases were classified as negative. We used three performance metrics to evaluate the detector. Precision (.92) and recall (.94) are simply accuracy of the detector where precision indicates the ratio of correct positive predictions and recall indicates the ratio of positive cases that were captured by the model. We further calculated Cohen's Kappa (κ), a widely used metric to evaluate goodness of data-mined models (Baker & Inventado, 2014). Kappa assesses whether the detector is better than chance at identifying the correct action sequences. A Kappa of 0 indicates that the detector performs at chance, and a Kappa of 1 indicates that the detector performs perfectly. This decision tree gave a kappa of .795 indicating a high agreement between the decision tree's and human coders' classification of students who demonstrate designing controlled experiments. We should note that this value was a little bit higher than the inter-rater reliability of .71, which might indicate possible over-fitting.

Figure 2 illustrates a fragment of the decision tree generated for the detector. Because a decision tree contains attributes and associated rules, it is more interpretable than other mining approaches (Bresfelean, 2007). For example, the very first feature used to classify students who demonstrate designing for controlled experiments skill is, "Adjacent controlled with repeats" (i.e., feature # 10 from the list of the detector features). Following down the decision tree, if there is no controlled experiment (smaller than 1), then the detector is 94 out of 98 confident that the incident is not demonstrating the skill (i.e., N for no). If the incident has "Adjacent controlled with repeats" count greater than 1, then the detector uses the second feature, "Adjacent controlled with no repeats" to continue classification. The decision tree obtained for the present detector is very much aligned with our previous detectors ob-

tained using the data from Physical Science microworlds (e.g., Sao Pedro et al., 2013)

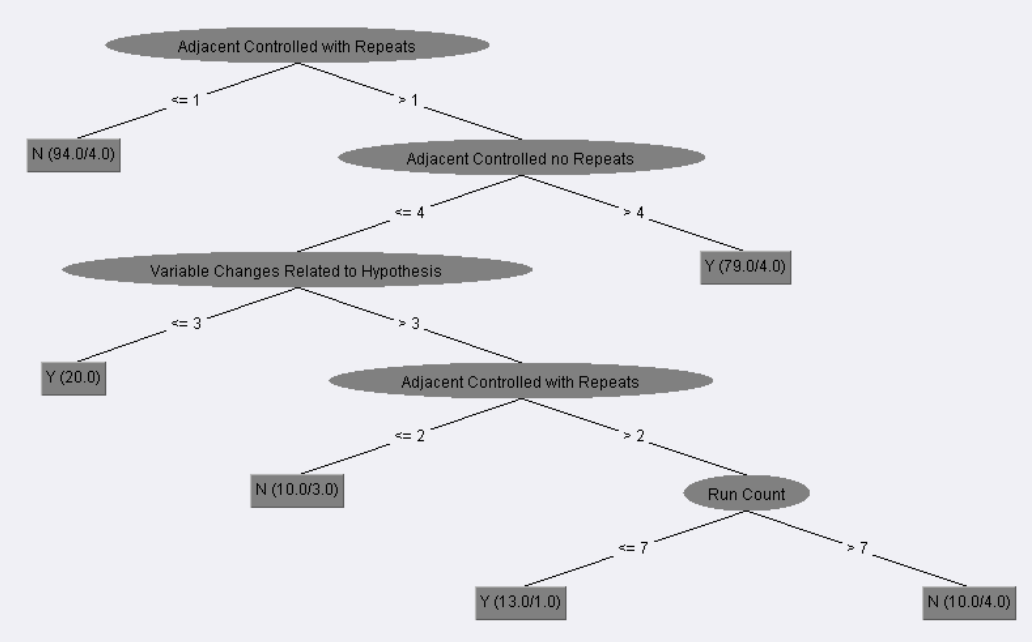

Figure 2. A fragment of the decision tree generated for the detector.

6. Discussion and Conclusions

Ill-defined science inquiry (e.g., Clarke-Midura, Dede, & Norton, 2011; Gobert, Sao Pedro, Baker, Toto, & Montalvo, 2012), such as the skill of designing and conducting experiments present many assessment challenges since traditional multiple choice items cannot be used to assess such skills (Haertel, Lash, Javitz, & Quellmalz, 2006; Quellmalz, Kreikmeier, DeBarger, & Haertel, 2007; Leighton & Gierl, 2011). In the present work, there is added difficulty in measuring students' experimental strategies during inquiry because we sought to assess these skills in a complex system, namely, Ecosystems. This adds an assessment challenge because as previously noted the simple control for variables strategy, whereby students should vary the target variable of interest and hold the remaining variables constant, cannot be applied because the independent variables interact in a complex causal system

leading to the change in a dependent or set of variables. Thus, the goal of this paper was to determine whether our EDM-based detector, used in other domains in our work (Sao Pedro et al, 2013; Gobert et al, 2012, 2013), could be successfully used to score students' skills at designing and conducting experiments when applied to logs from inquiry in a complex system microworld with multiple interacting variables.

As described in the results section, the detector's performance was quite high and indicate that the detector can be used to evaluate students' inquiry performance for the designing for controlled experiments skill in the Ecosystems activities. It can distinguish when a student designed controlled experiments in Ecosystems from when they did not 79% of time (κ = .795). This performance is on par (slightly better than) with previous metrics computed at the student-level across our three physical science topics for this skill, κ ranging from .45 to .62 across studies (Sao Pedro, Baker, & Gobert, 2012; 2013).

It is important to note that the features used for the presented detector were the same features that were used in the development of our detector for this skill in Physical Science (Sao Pedro et al, 2013). There are several explanations for this. First, given that this task, though a complex system per se, it is representative of a fairly simple system in that only 4 variables (i.e., seaweed, shrimp, small fish, and large fish populations) are interacting. Specifically, Narayanan et al., (2003) laid out five characteristics of complex systems as follows: (1) they exhibit hierarchical structures composed of subsystems and components; (2) their subsystems and components exhibit natural behaviors or engineered functions; (3) the component/subsystem behaviors causally influence other components/subsystems; (4) the propagation of the causal influences create chains of events in the operation of the overall system, and gives rise to its overall behavior and function; and (5) these chains of events extend in temporal and spatial dimensions. As such, it appears that our Ecosystem microworld, if viewed with these criteria, is at the less complex end of the spectrum. Additionally, our Ecosystems microworld can be solved using an "engineering approach", as outlined by Narayanan, and thus our features used to detect the design of controlled experiments can get us "pretty far" in detecting this skill in students because there are only 4 interacting variables. With this in mind, it is not surprising then that the same features can be used for both Physical

science topics and Ecosystems. It is an empirical question whether the same set of features would yield reliable metrics for evaluating this skill in a "more complex" complex system (say with 8 interacting variables), as outlined by Narayanan et al. (2003). Another possible explanation for these findings is that students' skills on this task are bimodal, i.e., either very buggy or very skilled and thus, the detector, as constructed, can discriminate "good" from "poor" examples of designing controlled experiments in these data.

In closing, this work contributes to the literature on performance-based assessment, and to the assessment of students' skills at designing and conducting experiments in complex systems. Taken together with our earlier work (Sao Pedro, Baker, Gobert, Montalvo, & Nakama, 2013; Sao Pedro et al, 2012) our results demonstrate the potential power of EDM for the broad scalability of our assessments across multiple science domains. Lastly, given their generalizability and power, these techniques provide a solution towards assessing this ill-defined skill in the context of complex systems, also referred to as systems thinking, as called for in reform documents on 21st century skills (NGSS, 2013; Partnership for 21st Century Skills, 2007). As previously stated, more research is needed with a other complex systems in which there are a greater number of interacting variables, etc., to address how well these techniques can validly assess students' experimentation strategies. This work represents an advance in assessment, in particular in complex systems, a here-to-fore difficult context in which to conduct inquiry assessment, given the multiple interacting variables. As such, it also represents a step towards the assessment of other inquiry skills in the context of complex systems, a necessary component of 21st Century skills (NGSS, 2013; Partnership for 21st Century Skills, 2007).

Acknowledgements

This research was funded by the National Science Foundation (NSF-REAL-1252477) and awarded to Janice Gobert (Principal Investigator) at WPI. Any opinions expressed are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Anastasi, A. & Urbina, S. (2009*). Psychological testing* (7th Ed.). Upper Saddle River, NJ: Prentice-Hall Publishers.

Baker, R. S., Corbett, A. T., & Wagner, A. Z. (2006). Human classification of low-fidelity replays of student actions. In *Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems* (No. 2002, pp. 29–36).

Baker, R. S., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction*, *18*(3), 287–314. doi: 10.1007/s11257-007-9045-6

Baker, R. S., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining* , *1* (1), 3–17.

Bereiter, C. (2002). *Education and mind in the knowledge age*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bresfelean, V. P. (2007). Analysis and predictions on students' behavior using decision trees in Weka environment. *In ITI 2007 29th International Conference on Information Technology Interfaces* (pp. 51–56). doi: 10.1109/ITI.2007.4283743

Buckley, B. C., Gobert, J. D., & Horwitz, P. (2006). Using log files to track students' model-based inquiry. In *Proceedings of the 7th international conference on learning sciences* (pp. 57–63). International Society of the Learning Sciences.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child development*, *70*(5), 1098–1120. doi: 10.1111/1467-8624.00081

Chi, M., VanLehn, K., & Litman, D. (2010). Do micro-level tutorial decisions matter: applying reinforcement learning to induce pedagogical tutorial tactics. In V. Aleven, J. Kay, & J. Mostow (Ed.), *ITS 2010, Part I, LNCS 6094* (pp. 224–234). Berlin Heidelberg: Springer-Verlag. doi: 10.1007/978-3-642-13388-6_27

Clarke-Midura, J., Dede, C., & Norton, J. (2011). *The road ahead for state assessments*. Policy Analysis for California Education and Rennie Center for Educational Research & Policy, Cambridge, MA.

Glaser, R., Schauble, L., Raghavan, K., & Zeitz, C. (1992). Scientific reasoning across different domains. In E. DeCorte, M. Linn, H. Mandl, & L. Verschaffel, *Computer-based Learning Environments and Problem-Solving* (pp. 345–371). Heidelberg, Germany: Springer-Verlag. doi: 10.1007/978-3-642-77228-3_16

Gobert, J. D. (submitted). Inq-ITS: Design decisions used for an inquiry intelligent system that both assesses and scaffolds students as they learn. To appear in Leighton, J. & Rupp, A. (Eds.) *Handbook of cognition and assessment*. New York: Wiley/Blackwell.

Gobert, J. D. (in press). Microworlds. In Gunstone, R. (Ed.) *Encyclopedia of science education*. Springer. doi: 10.1007/978-94-007-2150-0_55

Gobert, J. D. (2005). Leveraging Technology and Cognitive Theory on Visualization to Promote Students' Science Learning and Literacy. In J. Gilbert, *Visualization in science education* (pp. 73–90). Dordrecht, The Netherlands: Springer-Verlag. doi: 10.1007/1-4020-3613-2_6

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

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, *22*(4), 521–563. doi: 10.1080/10508406.2013.837391

Gobert, J. D., Baker, R. S. & Wixon, M. (2015) Operationalizing and Detecting Disengagement Within Online Science Microworlds. *Educational Psychologist*, 50(1), 43–57. doi: 10.1080/00461520.2014.999919

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

Goldstone, R. L. (2006). The complex systems see-change in education, *Journal of the Learning Sciences, 15*(1), 35–43. doi: 10.1207/s15327809jls1501_5

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving a new assessment perspective. *Applied Psychological Measurement, 36*(3), 189–213. doi: 10.1177/0146621612439620

Grotzer, T. A., & Basca, B. B. (2003) How does grasping the underlying causal structures of ecosystems impact students' understanding?, *Journal of Biological Education, 38*(1), 16–29. doi: 10.1080/00219266.2003.9655891

Haertel, G., Lash, A., Javitz, H., & Quellmalz, E. (2006). *An instructional sensitivity study of science inquiry items from three large-scale science examinations*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Hershkovitz, A., Wixon, M., Baker, R. S., Gobert, J. D., & Sao Pedro, M. A. (2011). Carelessness and goal orientation in a science microworld. In G. Biswas et al. (Eds.): AIED 2011, LNAI 6738 (pp. 462–465). Heidelberg, Germany: Springer. doi:10.1007/978-3-642-21869-9_70

Hmelo-Silver, C., & Azevedo, R. (2006). Understanding Complex Systems: Some Core Challenges. *Journal of the Learning Sciences, 15(1),* 53–61. doi: 10.1207/s15327809jls1501_7

Jacobson, M. J., Wilensky, U., Goldstone, R., Landy, D., Son, J., Lesh, R., ... & Azevedo, R. (2006, June). Complex systems in education: conceptual principles, methodologies, and implications for research in the learning sciences. In Proceedings of the 7th international conference on Learning sciences (pp. 1073–1077). International Society of the Learning Sciences. doi: 10.1207/s15327809jls1501_4

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive science*, *12*(1), 1–48. doi:10.1207/s15516709cog1201_1

Kuhn, D. (2005). *Education for thinking.* Cambridge, MA: Harvard University Press.

Kuhn, D. Schauble, L., and Garcia-Mila, M. (1992). Cross-domain development of scientific reasoning, *Cognition and Instruction, 9*(4), 285–327. doi: 10.1207/s1532690xci0904_1

Lai, E. & Vierling, M. (2012). *Assessing* 21st *Century Skills: Integrating Research Findings*. National Council on Measurement in Education, Vancouver, B.C.

McElhaney, K. W., & Linn, M. C. (2008). Impacts of students' experimentation using a dynamic visualization on their understanding of motion. In *Proceedings of the 8th international conference on International conference for the learning sciences-Volume 2* (pp. 51-58). International Society of the Learning Sciences.

McElhaney, K. W., & Linn, M. C. (2010). Helping students make controlled experiments more informative. In *Proceedings of the 9th International Conference of the Learning Sciences-Volume 1* (pp. 786-793). International Society of the Learning Sciences.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining, 4*(1), 11–48.

National Committee on Science Education Standards and Assessment. (1996). *National Science Education Standards: 1996*. Washington, D.C.: National Academy Press.

NGSS Lead States. (2013). Next Generation Science Standards: For States, By States. *Washington, DC: The National Academies Press*.

Njoo, M., & De Jong, T. (1993). Exploratory learning with a computer simulation for control theory: Learning processes and instructional support. *Journal of Research in Science Teaching*, *30*(8), 821–844. doi:10.1002/tea.3660300803

Papert, S. (1980). Computer-based microworlds as incubators for powerful ideas. In R. Taylor, *The Computer in the School: Tutor, Tool, Tutee* (pp. 203–201). New York, NY: Teacher's College Press.

Partnership for 21st Century Skills (2007). P21 Framework definitions. Retrieved from http://www.p21.org/storage/documents/P21_Framework_Definitions,pdf.

Pavlik, P., Cen, H., & Koedinger, J. (2009). Performance Factors Analysis - A New Alternative to Knowledge Tracing. *Proceedings of the 14th International Conference on Artificial Intelligence in Education, AIED 2009* (pp. 531–540). Brighton, UK: IOS Press.

Pea, R., & Kurland, D. (1984). On the cognitive effects of learning computer programming. *New Ideas in Psychology, 2*, 137–168. doi:10.1016/0732-118x(84)90018-7

Pellegrino, J. (2009). The design of an assessment system for rat to the top: A learning sciences perspective on issues of growth and measurement. Center for K-12 Assessment & Performance Management, ETS.

Perkins, D. (1986). *Knowledge as design*. Hillsdale, NJ: Erlbaum.

Quellmalz, E., & Haertel, G. (2004). Technology supports for state science assessment systems *Paper commissioned by the National Research Council Committee on Test Design for K–12 Science Achievement*. Washington, DC: National Research Council.

Quellmalz, E., Kreikemeier, P., DeBarger, A. H., & Haertel, G. (2007). *A study of the alignment of the NAEP, TIMSS, and New Standards Science Assessments with the inquiry abilities in the National Science Education Standards*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Il.

Reimann, P. (1991). Detecting functional relations in a computerized discovery environment. *Learning and instruction*, *1*(1), 45–65. doi:10.1016/0959-4752(91)90018-4

Resnick, M. (1997). *Turtles, Termintes, and Traffic Jams: Explorations in Massively Parallel Microworlds.* Cambridge, MA: MIT Press.

Resnick, M. (2008). Sowing the Seeds for a More Creative Society. *Learning & Leading with Technology*, *35*(4), 18–22. doi:10.1145/1518701.2167142

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State-of-the-Art. *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews , 40* (6), 601–618. doi:10.1109/tsmcc.2010.2053532

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *The Journal of Technology, Learning and Assessment*, *8*(4).

Sao Pedro, M. A., Baker, R. S., & Gobert, J. D. (2012). Improving construct validity yields better models of systematic inquiry, even with less information. In *User Modeling, Adaptation, and Personalization* (pp. 249-260). Springer Berlin Heidelberg. doi:10.1007/978-3-642-31454-4_21

Sao Pedro, M. A., Baker, R. S., & Gobert, J. D. (2013). What different kinds of stratification can reveal about the generalizability of

data-mined skill assessment models. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 190–194). ACM. doi:10.1145/2460296.2460334

Sao Pedro, M. A., Baker, R. S., Gobert, J. D., Montalvo, O., & Nakama, A. (2013). Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Modeling and User-Adapted Interaction*, *23*(1), 1–39. doi:10.1007/s11257-011-9101-0

Sao Pedro, M. A., Baker, R. S., & Gobert, J. (2013). Incorporating scaffolding and tutor context into bayesian knowledge tracing to predict inquiry skill acquisition. In *Proc. of the 6th International Conference on Educational Data Mining, Memphis, TN* (pp. 185–192).

Sawyer, R. K. (2006). Educating for innovation. *Thinking skills and creativity*, *1*(1), 41–48.

Schauble, L., Glaser, R., Duschl, R., Schulze, S., & John, J. (1995). Students' understanding of the objectives and procedures of experimentation in the science classroom. *Journal of the Learning Sciences*, *4*(2), 131–166. doi: 10.1207/s15327809jls0402_1

Schauble, L., Klopfer, L. E., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching*, *28*(9), 859–882. doi: 10.1002/tea.3660280910

Schunn, C. D., & Anderson, J. R. (1998). Scientific Discovery. In J. R. Anderson, *The Atomic Components of Thought* (pp. 385–428). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Schunn, C. D., & Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. *Cognitive Science*, 23(3), 337–370. doi:10.1207/s15516709cog2303_3

Shute, V. J., & Glaser, R. (1990). A large□scale evaluation of an intelligent discovery world: Smithtown. *Interactive Learning Environments*, *1*(1), 51–77. doi:10.1080/1049482900010104

Shute, V. J., Glaser, R., & Raghavan, K. (1989). Inference and discovery in an exploratory laboratory. In *Learning and Individual Differences: Advances in Theory and Research.* (pp. 279–326). New York, NY: W.H. Freeman

Sternberg, R. J. (2006). The nature of creativity. *Creativity Research Journal*, *18*(1), 87–98. doi:10.1207/s15326934crj1801_10

Sternberg, R. J., & Lubart, T. I. (1991). Creating creative minds. *Phi Delta Kappa*, 608–614.

Sternberg, R. J., & Lubart, T. I. (1995). *Defying the crowd: Cultivating creativity in a culture of conformity*. Free Press.

Stokes, D. E. (1997). Pasteur's quadrant: Basic science and technological innovation. Washington, DC: Brookings Institute.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, *51*(1), 1–10. doi:10.2307/1129583

White, B. Y. (1993). ThinkerTools: Causal models, conceptual change, and science education. *Cognition and Instruction*, *10*(1), 1–100. doi:10.1207/s1532690xci1001_1

Wilensky, U., & Resnick, M. (1999). Thinking in levels: A dynamic systems approach to making sense of the world. *Journal of Science Education and Technology*, *8*(1), 3–19.

Yoon, S. A. (2008). An evolutionary approach to harnessing complex systems thinking in the science and technology classroom. *International Journal of Science Education*, *30*(1), 1–32. doi:10.1080/09500690601101672

Table 1. Confusion matrix and performance metrics for the Ecosystem clips

	Hand-coded Positive	Hand-coded Negative
Predicted Positive	111	10
Predicted Negative	7	98
Precision = 0.92, Recall = 0.94, κ = .795		