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

Janice D. Gobert a,*, Yoon Jeon Kim a,1, Michael A. Sao Pedro a,1, Michael Kennedy b,2, Cameron G. Betts c,3

a Social Sciences & Policy Studies Department, WPI, 100 Institute Rd., Worcester, MA 01609, USA
b WPI, 100 Institute Rd., Worcester, MA 01609, USA
c Apprendis, LLC, USA

ARTICLE INFO

Article history:
Received 4 October 2014
Received in revised form 25 April 2015
Accepted 30 April 2015
Available online xxx

Keywords:
Complex systems
Inquiry assessment
Performance assessment
Educational data mining
21st century skills

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 logged 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 (inquiry intelligent tutoring system), 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.

© 2015 Elsevier Ltd. All rights reserved.

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

* Corresponding author. Tel.: +1 508 831 5619; fax: +1 508 831 5296.
E-mail addresses: jgobert@wpi.edu (J.D. Gobert), yjkim@wpi.edu (Y.J. Kim), mikesp@wpi.edu (M.A. Sao Pedro), kennedym3@gmail.com (M. Kennedy), cam@apprendis.com (C.G. Betts).
1 Tel.: +1 508 831 6618.
2 Tel.: +1 774 249 5327.
3 Tel.: +1 617 538 6868.

http://dx.doi.org/10.1016/j.tsc.2015.04.008
1871-1871/© 2015 Elsevier Ltd. All rights reserved.

Please cite this article in press as: Gobert, J. D., et al. Using educational data mining to assess students’ skills at designing and conducting experiments within a complex systems microworld. Thinking Skills and Creativity (2015), http://dx.doi.org/10.1016/j.tsc.2015.04.008
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 (National Research Council, 2010; 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 & Vierling, 2012). Lastly, 21st century skills are acknowledged as important for developing innovative thinkers (Sawyer, 2006; Sternberg, 2006; Sternberg & Lubart, 1991, 1995), necessary for a knowledge-based economy (Bereiter, 2002; Resnick, 2008).

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 & Azevedo, 2006; Jacobson et al., 2006; Yoon, 2008). In other work, we address intrapersonal skills, namely engagement (Gobert, Baker, & Wixson, 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 (Clarke-Midura, Dede, & Norton, 2011; Haertel, Lash, Javitz, & Quellmalz, 2006; Leighton & Gierl, 2011; National Committee on Science Education Standards and Assessment, 1996; Quellmalz & Haertel, 2004; Quellmalz, Kreikmeier, DeBarger, & Haertel, 2007). As discussed elsewhere (Gobert, Sao Pedro, Raziuddin, & Baker, 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, Behrens, Dicerbo, & Levy, 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, Gushta, Mislevy, & Shaffer, 2010).

In short, the problem becomes: how do we use policy documents about critical thinking in science (National Research Council, 2010; 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 base upon which to develop performance-based assessments by leveraging computational techniques to analyze students’ log files of their inquiry processes (Gobert & 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.

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; Desmarais, Meshkinaf, & Gagnon, 2006; Pavlik, Cen, & Koedinger, 2009), to build automated detectors of relevant constructs during student learning (Baker & Wixon, 2015; Baker, Corbett, Roll, & Koedinger, 2008; Hershkovitz, Wixon, Baker, & Sao Pedro, 2011), and to do both formative and performance assessment (Gobert et al., 2012; Mislevy 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, 2015a,b 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 the 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 (Glaser, Schauble, Raghavan, & Zeitz, 1992; Shute & Glaser, 1990), may run only one trial (Kuhn, Schauble, & Garcia-Mila, 1992) or run the same trial repeatedly (Buckley, Gobert, & Horwitz, 2006; Kuhn et al., 1992). They also change too many variables at once (Glaser et al., 1992; Harrison & Schunn, 2004; Kuhn, 2005; McElhaney & Linn, 2008, 2010; Reimann, 1991; Schunn & Anderson, 1998, 1999; Shute & Glaser, 1990; Tschirgi, 1980). 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, Klopfier, & 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, 2012; Sao Pedro, Baker, Montalvo, & Nakama, 2013b), 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 & Vierling, 2012). Our ecosystems microworld targets students’ understanding of the ways in which organisms interact and have different functions within an ecosystem to enable survival (Sao Pedro, Gobert, & Betts, 2014). Since the ecosystems environment has multiple variables interconnected in a non-linear fashion (Greiff, Wustenberg, & Funke, 2012; Yoon, 2008), the hypothesis space increases (Klahr & Dunbar, 1988), and the understanding of the effects the independent variables on dependent variables is more challenging because, as previously stated, the simple control for variables strategy (cf., Chen & Klahr, 1999), described 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 (Jacobson et al., 2006; Yoon, 2008), 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 & Basca, 2003; Grotzer & Perkins, 2000). 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, Gobert, & Beck, 2010) because of the multiple interacting independent variables, i.e., where variables Variables 1 and 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 & Azevedo, 2006; 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 & Azevedo, 2006); 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 et al., 2013b). 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 et al., 2012; Sao Pedro et al., 2013b).

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.

2. Method

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

2.2. 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 Fig. 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.

2.3. 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 (Fig. 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 Fig. 1), analyze data by examining automatically generated data tables and population graphs (on the right side of Fig. 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.

3. Data analysis

3.1. 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 did not 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, Baker, & Gobert, 2013a, 2013c). There are several advantages of using machine-learned detectors (Sao Pedro et al., 2013a, 2013c). 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., 2013a, 2013c). Within the corpus of tagged clips, 52.2% of students had demonstrated the skill of controlling for variables.

3.2. 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; Chen & Klahr, 1999; Kuhn et al., 1992). 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 Gobert et al., 2013; Sao Pedro et al., 2012 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.

Please cite this article in press as: Gobert, J. D., et al. Using educational data mining to assess students’ skills at designing and conducting experiments within a complex systems microworld. Thinking Skills and Creativity (2015), http://dx.doi.org/10.1016/j.tsc.2015.04.008
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.

3.3. Detector generation and validation

Continuing with the EDM-based method used in our group (Gobert et al., 2013; Sao Pedro et al., 2013a, 2013c), 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, 1996), 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, 1996). 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., 2013a, 2013c), 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 threshold 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. Baker & De Carvalho, 2008; Sao Pedro et al., 2013a, 2013c; Walonoski & Heffernan, 2006). Also, decision trees produce relatively human-interpretable models (i.e., attributes and associated rules). For example, as depicted in Fig. 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 et al., 2012).

4. Results

The confusion matrix (Table 1) captures raw agreement between the detector’s prediction and the human coders’ tags under student-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 (0.92) and recall (0.94) 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 ($\kappa$), 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.

Fig. 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 obtained using the data from Physical Science microworlds (e.g., Sao Pedro et al., 2013a, 2013c)

5. Discussion and conclusions

Ill-defined science inquiry (e.g., Clarke-Midura et al., 2011; Sao Pedro et al., 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 et al., 2006; Quellmalz et al., 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 set of variables. Thus, the goal of this paper was to determine whether our EDM-based detector, used in other domains in our work (Gobert et al., 2012, 2013; Sao Pedro et al., 2013a, 2013c), 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 ($\kappa = .795$). 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, $\kappa$ ranging from .45 to .62 across studies (Sao Pedro et al., 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., 2013a, 2013c). 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 (2007) 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

Table 1: Confusion matrix and performance metrics for the ecosystem clips.

<table>
<thead>
<tr>
<th></th>
<th>Hand-coded positive</th>
<th>Hand-coded negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted positive</td>
<td>111</td>
<td>10</td>
</tr>
<tr>
<td>Predicted negative</td>
<td>7</td>
<td>98</td>
</tr>
<tr>
<td>Precision = 0.92, Recall = 0.94, $\kappa = .795$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please cite this article in press as: Gobert, J. D., et al. Using educational data mining to assess students’ skills at designing and conducting experiments within a complex systems microworld. Thinking Skills and Creativity (2015), http://dx.doi.org/10.1016/j.tsc.2015.04.008
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 (2007). 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 et al., 2012; Sao Pedro et al., 2013b) 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 toward 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 Lead States, 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 toward the assessment of other inquiry skills in the context of complex systems, a necessary component of 21st Century skills (NGSS Lead States, 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


Please cite this article in press as: Gobert, J. D., et al. Using educational data mining to assess students’ skills at designing and conducting experiments within a complex systems microworld. Thinking Skills and Creativity (2015), http://dx.doi.org/10.1016/j.tsc.2015.04.008


Please cite this article in press as: Gobert, J. D., et al. Using educational data mining to assess students’ skills at designing and conducting experiments within a complex systems worldview. *Thinking Skills and Creativity* (2015). http://dx.doi.org/10.1016/j.tsc.2015.04.008


Please cite this article in press as: Gobert, J. D., et al. Using educational data mining to assess students’ skills at designing and conducting experiments within a complex systems microworld. Thinking Skills and Creativity (2015), http://dx.doi.org/10.1016/j.tsc.2015.04.008