CHAPTER 12 – Methods for Assessing Inquiry: Machine-learned and Theoretical

Michelle LaMar[1], Ryan S. Baker[2], and Samuel Greiff[3]
Educational Testing Service[1], University of Pennsylvania[2], University of Luxembourg[3]

Introduction

Inquiry skills are critical to virtually any kind of problem solving and particularly to the practice of science. However, these skills cannot be easily assessed using traditional static assessment items, in which it is difficult to follow (and capture) the actual inquiry process. Given a sufficiently interactive task, the challenge becomes how to identify and score the inquiry skills applied to the task. This chapter examines three different methods for assessing inquiry skills using the process data generated by computerized interactive tasks. The first two methods identify inquiry strategy-use based on specific features in the data. The first uses theoretically defined features, whereas the second leverages machine learning to identify and combine relevant features. The third method uses generative process models to compare student actions to probabilistic agents implementing targeted inquiry strategies. The advantages and drawbacks of each method are discussed along with assessment contexts that favor particular approaches. Implications for detection and scoring of inquiry within the Generalized Intelligent Framework for Tutoring (GIFT) are addressed.

In a world that increasingly uses technology in the workplace, at home, and as a medium for commerce and communication, the ability to think scientifically and solve technical problems is increasingly useful in modern life. Both scientific inquiry processes and problem-solving skills have been identified as key 21st century skills due to their pivotal role for success in contemporary societies (Dede, 2010). As with many 21st century skills, these competencies are inherently interactive, involving multi-step processes that must adapt to information gathered and results generated. In K–12 education, the need to teach authentic science practices has been a major impetus for the move from static textbooks and cookbook labs to more interactive science instruction (e.g., Next Generation Science Standards [NGSS; NGSS Lead States, 2013]) that frequently use intelligent tutoring systems (ITSs), computer simulations, and games. With this shift in instructional emphasis comes a need to also assess these skills using interactive problem-solving tasks. ITSs in particular depend upon real-time skill diagnosis and assessment to allow for customized instruction and guidance.

Scoring interactive performances, however, is much more complicated than scoring responses chosen from a small, fixed list of options such as in the multiple-choice item formats favored by standardized testing. While performance assessment is not new, evaluating performance has traditionally been conducted by human raters who call upon complex experience to interpret and judge the amount of skill and understanding displayed in a particular performance. Machine scoring has made major progress, mostly in the form of scoring written text (Shermis, 2014; Liu et al., 2016), but more recently there has been progress in scoring competencies of complex problem solving or science inquiry with interactive tasks (Gobert, Sao Pedro, Baker, Toto & Montalvo, 2012; Greiff, Wüstenberg & Avvisati, 2015). In this chapter, we discuss methods for automatically identifying and scoring inquiry strategies based on logged interactions within computerized simulations.

Science Inquiry

Science inquiry and problem solving have frequently been characterized in terms of two major phases: hypothesis generation and hypothesis testing (Popper, 1959; Klahr, 2002). While many aspects of practicing

science could be demonstrated using interactive science tasks, the focus of this chapter is on inquiry strategies that might be used in the hypothesis-testing phase of scientific investigations. Successful strategies in scientific hypothesis testing largely overlap with strategies that are useful for complex problem solving (Jonassen, 2007; Klahr & Dunbar, 1988), thus making the detection and assessment of these inquiry strategies applicable well beyond traditional science instruction.

The definition of science practices in NGSS (NGSS Lead States, 2013) includes eight science practices, of which hypothesis testing is mostly covered by two practices: "planning and carrying out investigations" and "analyzing and interpreting data". Key skills listed in these practices include identifying, controlling, and measuring relevant variables, deciding what and how much data to collect under what range of conditions, conducting systematic analysis of data, recognizing evidence that contradicts the hypothesis, and evaluating the strength of conclusions that can be made from a set of data. Different theoretical frameworks break these skills down in different ways with different emphases (for example, see Wieman, 2015). However, the key elements of investigation remain: how to collect data and how to interpret data.

Assessment and Scoring Challenges

To assess skills in science inquiry and problem solving, it is necessary to have the examinees engage in the process of science inquiry or problem solving. The design of appropriately interactive tasks to allow for the demonstration of these skills comprises the first assessment challenge. Traditional test item design favors testing one skill at a time; however, the interdependency between the inquiry practices and the time commitment required for an interactive task can make it more feasible to assess multiple skills simultaneously within a larger investigation. The impulse to limit the task, ensuring that we understand what we are measuring, pushes design toward more scaffolded tasks with a fixed number of choices with limited options. The desire to measure more authentic inquiry including interdependent skills and content knowledge, on the other hand, pushes the design to open, exploratory, scenarios in which there are many choices and many ways to go wrong. Depending upon the purpose of the assessment, a more scaffolded versus a more open design are considered but in practice the choice is often made based on the limitations of the scoring.

A second assessment challenge is scoring performances in inquiry investigation. The easiest scoring method is based on the outcome of the investigation alone. If the examinee is able to draw the correct conclusions, we can infer that they probably carried out a successful investigation. There are a number of problems with the outcome-only metric, however. While correct answers undoubtedly correlate with correct practices, it is also possible to stumble upon the correct answers by chance or, in more scaffolded tasks, to try every possible variable combination without taking the time to develop a systematic data collection plan. Furthermore, an incorrect answer might stem from a failure of any number of skills or content knowledge, especially in the more open-ended task designs. Thus, an incorrect answer could easily mask good overall inquiry skills.

As an alternative to outcome-only scoring, a record of the actual steps taken within an investigative task can be analyzed for evidence of science-inquiry skills. This approach is particularly compatible with computerized interactive tasks, which allow for a wide variety of experimental setups along with the collection of data about actions taken within the task. Scoring this collection of process data, however, is far from straightforward. For each decision the student is allowed, the number of paths through the task grows exponentially. Thus, scoring on the basis of a set of "correct" paths quickly becomes impractical. The methods presented here involve analysis of actions taken within an interactive task to detect instances of good science practice and provide scoring data for generalized problem-solving tasks. This approach requires defining not only what constitutes good inquiry, but also what the application of good inquiry skills looks like in specific task contexts. Further, detecting an instance of inquiry strategy use is not, in itself, a score.

Further analysis or a broader model might be needed to connect the identified behavior to valid inferences about the student's problem-solving ability.

In the following sections, we describe three different methods for identifying and scoring good inquiry practice from a record of interactions within a complex task. The first two methods identify inquiry strategy-use based on specific features in the data. The first uses theoretically defined features, whereas the second leverages machine learning to identify and combine relevant features. The third method creates probabilistic models of decision making based on different inquiry strategies and compares student actions to model predictions to identify strategy use.

Methods for Identifying Good Inquiry Strategy

Theoretically Defined Performance Indicators

A first approach toward identifying inquiry skills in computer-simulated task environments is based on theoretically developed behavioral indicators that are directly derived from the underlying definition of science inquiry and its theoretical components. In the theoretical approach, a definition of science inquiry is the starting point (Kuhn, 2012; Zimmerman, 2007). Along with this definition, carefully drafted tasks aimed at measuring science inquiry need to be developed. These tasks need to follow the theoretical rationale and incorporate those aspects of science inquiry that are considered crucial in the definition. Equipped with an adequate task environment, specific behaviors (or behavioral patterns) that can occur within the task environment are identified as indicating either high or low levels of science-inquiry skills. Thus, it is important that the task space is designed in a way that it allows for different behaviors that are indicative of the underlying theoretical conception and that represent high versus low proficiency levels of this conception.

A straightforward example of an important concept in the field of science inquiry is the principle of isolated variation (sometimes referred to as the vary-one-thing-at-a-time strategy [VOTAT]; Tschirgi, 1980) in which students demonstrate their ability to comprehend, use, and argue along the lines of causal relations within scientific phenomena (Kuhn, Black, Keselman & Kaplan, 2000). Because the VOTAT strategy is effective in isolating causal relationships and reducing the influence of confounding variables, it is considered relevant across a number of domains and has been identified as an important strategy in the field of complex problem solving and science inquiry (Wüstenberg, Greiff & Funke, 2012). Interestingly, Kuhn (2012) highlights that even adults often suffer from an insufficient understanding of the principle of isolated variation.

Figure 1 displays an example of a task environment taken from the field of complex problem solving, though it could easily be used as a science-inquiry task with a couple of small adaptations. This type of task, often referred to as MicroDYN-type of task (Greiff, Wüstenberg & Funke, 2012), features a small simulated system with multiple inputs or controls and multiple outputs or effects. The goal set for the student is to map the relationship between the input and output variables. The example in Figure 1 is usually used in a secondary education context and students are asked to determine how different components of a windmill affect both the noise produced by the windmill and the costs associated with operating it. The problem environment presented here is not fully open because students have a small number of actions they can perform in the environment. However, there is no explicit guidance of student approaches, which allows expression of individual differences between students. There are several strategies to solve this task, but students who are familiar with and proficient in applying and using the VOTAT-strategy usually perform better on these tasks. This is no coincidence because the task environment was developed on the basis of a theoretical understanding of complex problem solving and the types of inquiry skills that are needed to get from the initial state to the goal state. The theory, which highlights the principle of isolated variation as an

important aspect of both science inquiry and complex problem solving, ensured the development of a task in which VOTAT behavior would be both productive and distinguishable from non-VOTAT behavior. Scoring for the task would be based on whether the examinees apply the principle of isolated variation during the unguided exploration phase. The final step is to define an indicator of the targeted behavior (here VOTAT) from the information stored in computer-generated log files. A theory-driven approach can be used along with other inquiry principles to do this. The common umbrella would be to develop the task design, targeted behaviors, and behavioral indicators in such a way that they elicit evidence of understanding and competency of the theoretically defined underlying principles.

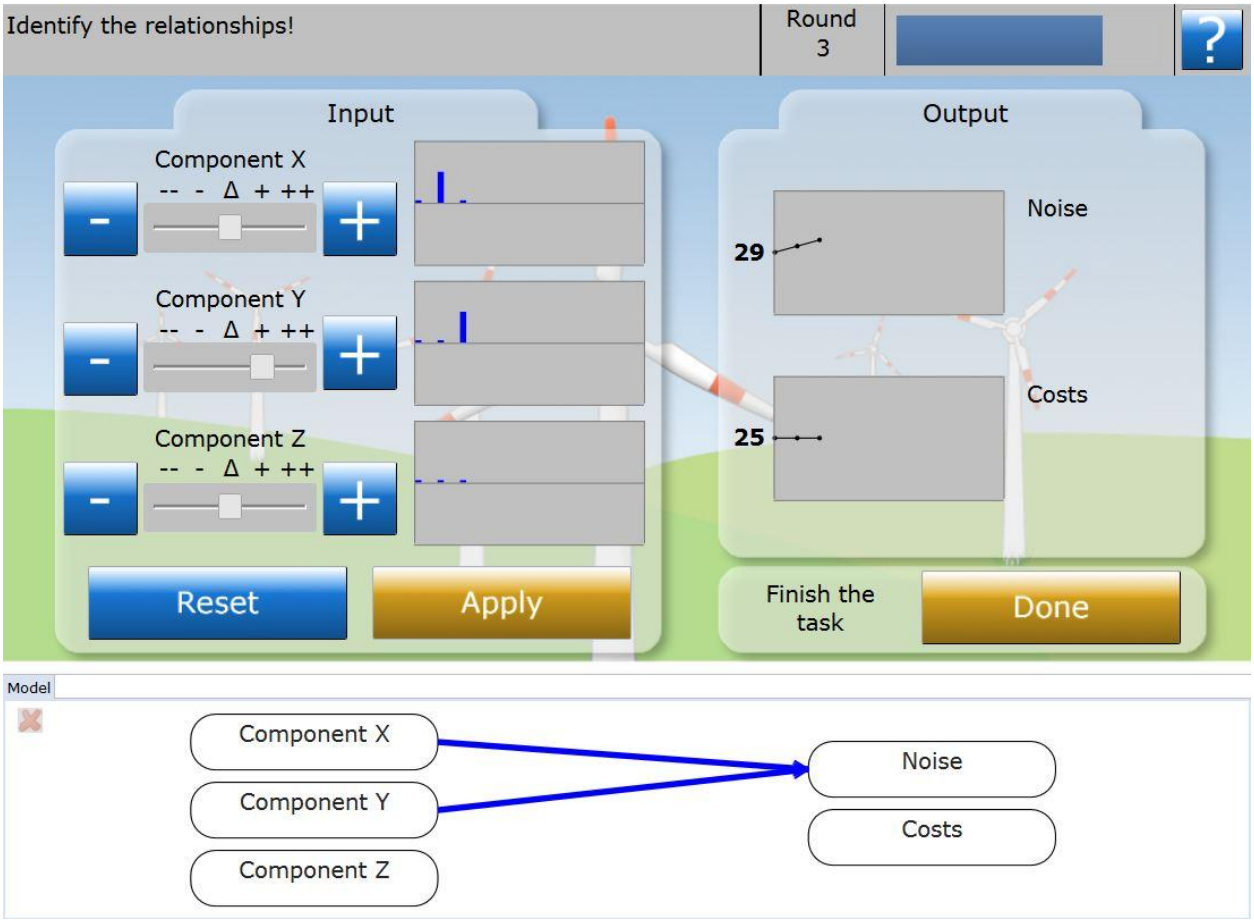

Figure 1. Complex problem-solving environment within the MicroDYN approach. Students are asked to discover the relations between the inputs on the left side and the outputs on the right side during a phase of unguided exploration.

Quantifying Use of the Principle of Isolated Variation (VOTAT-Strategy)

The operationalization of science inquiry is still a matter of some theoretical debate. Even with a profound theoretical understanding, there are many ways isolated variation as an overarching principle might be scored in different (or even the same) task environments. Depending on the specific scoring, results might vary widely. In the following, some empirical examples are presented, but it is important to note that these indicators have been theoretically defined based on the specific assumptions and goals of this assessment and manifest within the scoring practices limited by those assumptions and goals.

Vollmeyer, Burns, and Holyoak (1996) reported on experimental studies in which they investigate the impact of the use of isolated variation in a problem-solving environment called Biology-Lab, which includes four independent and four dependent variables. They differentiate between students who initially use or adopt isolated variation over the course of problem exploration from those who use other, less efficient strategies. Students were allowed four learning rounds, which consisted of six experimental trials per round. A round was coded as using VOTAT if, in at least four out of the six trials, only one input variable was varied while the others were set to zero. The general pattern of results indicated that use of an efficient strategy (here VOTAT) as compared to other strategies is associated with better learning outcomes and with a higher level of rule acquisition as an overall performance indicator. Thus, a theoretically defined process indicator of isolated variation was shown to relate to performance indicators within a problem-solving environment, thereby confirming the initial theoretical assumption.

In a similar vein, Greiff, Wüstenberg, and Avvisati (2015) analyzed log file data from the 2012 assessment of problem solving in one of the most important educational large-scale assessments, the Programme for International Student Assessment (PISA; OECD, 2014). In a large sample of 15-year-old students, Greiff et al. scored whether students employed the principle of isolated variation for all of the three input variables during the unguided exploration phase (i.e., consistent use of VOTAT) or whether they did so either not at all or inconsistently (i.e., not for all of the input variables). They reported that the strategic use of isolated variation in a MicroDYN-like task (see Figure 1) was related to both overall task performance and the overall problem-solving score in PISA. That is, the specific strategic behavior within one task was related not only to performance in this task but beyond that one task to overall problem-solving performance (and thus to general problem-solving and inquiry performance).

In explorative analyses, Greiff et al. (2015) employed an additional and more detailed scoring in which student strategy was scored along the number of input variables for which the principle of isolated variation was applied. Thus, scores ranged from 0 to 3 for the task with 3 input variables. In doing so, the authors (tentatively) reported progressive proficiency levels among students. These proficiency levels went beyond merely distinguishing between those students who used and those who did not use isolated variation as in the first analyses. Of note, since two different ways of scoring were used in the same sample, slightly different insights were gained depending on the specific way the principle of isolated variation was operationalized.

The two studies mentioned here serve as examples that consistently show the relation between the use and application of the principle of isolated variation as an important conceptual aspect of inquiry skills, its operationalization through behavior-based indicators, and external criteria that serve as validity evidence. However, there are both pros and cons of this theoretically motivated approach, which are briefly discussed in the next section.

Pros and Cons

An important advantage of the above-described approach of employing theoretically defined indicators of inquiry skills is the direct connection to theories of the human mind in general and theories on inquiry and problem solving in particular. Because it is necessary to engage in elaboration of the underlying theoretical foundation before any indicator can be clearly defined, all indicators are easily interpretable. In contrast to the machine-learning approaches that are described in the next section, theoretically defined performance indicators are always embedded in some kind of broader framework that helps put the specific indicators into perspective and give them meaning.

The theoretical approach is associated with drawbacks as well. One of them certainly is that in many cases a specific behavior (or, even more so, a non-behavior) cannot be directly mapped to an underlying theoretical defined construct. For instance, it is quite straightforward to claim – against the backdrop of the above-

mentioned empirical studies – that there is some causal connection between the use of VOTAT and performance in problem-solving environments. However, what about students who did not employ this principle? There might be several reasons for their lack of adequate strategic behavior including a lack of understanding, low level of motivation and task engagement, or issues with understanding the instructions of the task. Thus, it is often difficult to establish an isomorphic mapping between a theoretical concept and a specific behavioral indicator. In addition to this, complex strategies that involve a number of variables and indicators are difficult to detect in a purely theoretical approach because it requires a clear understanding of the specific underlying mechanisms and how they manifest as specific behaviors. Thus, the theoretical approach is often somewhat limited to a narrow set of theoretical aspects with the complex interplay of several behaviors being omitted. This last shortcoming is exemplified by the fact that the large body of literature on science inquiry mostly revolves around the rather straightforward and easily to define and detect principle of isolated variation (Kuhn, 2012). Overall, theory is the cornerstone for any sound and scientifically valid understanding. However, this type of confirmatory approach needs to be complemented by more data-driven and exploratory methods such as machine-learned inquiry detection.

Machine-Learned Inquiry Detection

A second approach toward identifying effective inquiry strategies is to use machine learning, often referred to in the domain of education as either educational data mining or learning analytics (Baker & Siemens, 2014). This research area refers to a broad range of methods that leverage the potential of analyzing thousands of possible relationships among variables in an automated fashion. Educational data-mining methods are particularly useful either when there is relatively little known about the domain being analyzed (in which case "unsupervised" methods are used that do not privilege specific variables for analysis) or when the construct to be modeled is known but it is thought that the best prediction or inference of it will involve combinations of variables that are more complex than a human could reasonably identify (in which "supervised" methods are used that attempt to discover the best combination of variables that identifies a known variable). Supervised methods, such as classification, are used to categorize specific cases into one of a small number of known categories on the basis of "features" of the data for that case. Machine-learning approaches have the advantage that the relationship between the features and the categories do not need to be known in advance; the set of features and their interactions are selected by the algorithm based on the relationships found in the data. In this section, we describe two successful uses of supervised methods to identify student inquiry skill. In one of these, humans identified successful inquiry in a limited data set and then machine learning was used to replicate their judgments at scale. In the second, successful inquiry was identified as student success at solving a puzzle that required inquiry and machine learning was used to identify patterns of behavior that led to that successful performance.

Replicating Human Judgment

In this first example, humans identified inquiry within a limited data set, and then machine learning was used to replicate that judgment. This work was predicated on the assumption that expert researchers in inquiry could recognize appropriate inquiry when they saw it, but that transforming that comprehension into simple rules is challenging and may lead to overly precise rules that either exclude some acceptable strategic behavior or treat some inappropriate strategies as appropriate. For instance, the VOTAT rule does not clarify how to treat a case where a student runs an experiment, changes two variables, runs an experiment, changes one back to the original variable, and then runs another experiment. The student has an unconfounded set of experiments, though perhaps a less efficient one, but did not use VOTAT between every pair of trials. At the same time, we would not want to credit a student who ran hundreds of trials and through exhaustion managed to hit every possible set of parameters in the simulation. Machine learning can develop rules that handle these special cases that align to expert intuition.

We studied the possibility of identifying appropriate inquiry strategies in a way that goes beyond simple rules such as VOTAT and can recognize a broader range of appropriate strategies in the context of an online learning system named *Inq-ITS* (formerly called *Science ASSISTments*), an ITS for scientific inquiry (Gobert, 2015), shown in Figure 2. Within Inq-ITS, students make hypotheses, manipulate simulations and collect data, and then interpret the results of their experiments in terms of their original hypotheses. As shown in Figure 2, students typically manipulate simulations by changing the values of a small set of parameters, where the choices are categorical (i.e., 3 choices per variable) rather than continuous.

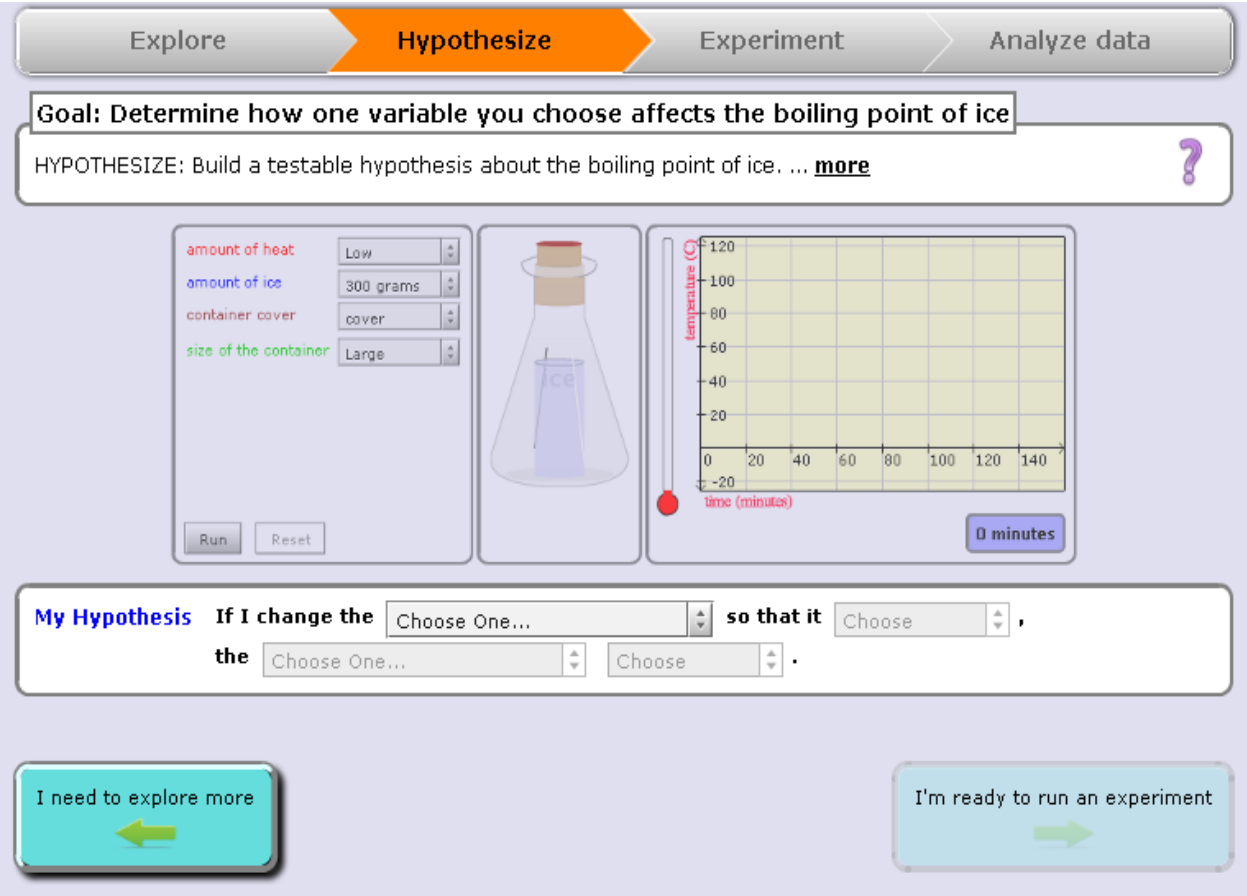

Figure 2. The version of Inq-ITS providing the data discussed in this article (more recent versions can be seen at http://www.inqits.com). Students make hypotheses, manipulate simulations and collect data, and then interpret the results of their experiments in terms of their original hypotheses.

To develop models of student inquiry within Inq-ITS, student behavior logged within Inq-ITS is transformed into a set of visualizations called "text replays". Originally proposed in Baker, Corbett & Wagner (2006), text replays are "pretty-printed" representations of student behavior over time, designed to be feasible for domain experts to read and use to identify behaviors or strategies of interest. For example, Figure 3, drawn from Gobert et al. (2012), shows text replays for Inq-ITS that were examined to identify whether the student is designing controlled experiments, whether the student is testing the stated hypothesis, and other aspects of student inquiry behavior. Within Inq-ITS, a single text replay corresponded to the actions a student made between creating hypotheses and interpreting their data in light of those hypotheses for a specific simulation. As Figure 3 shows, the human coders are able to see the relative time of student actions, what hypotheses the student made, what variables were manipulated between runs of the simulation to conduct experiments, and when (and how many times) the student ran the simulation. Not shown (but

also included in the text replays) are additional behaviors such as pausing the simulation, re-viewing the list of hypotheses, or opening the data table.

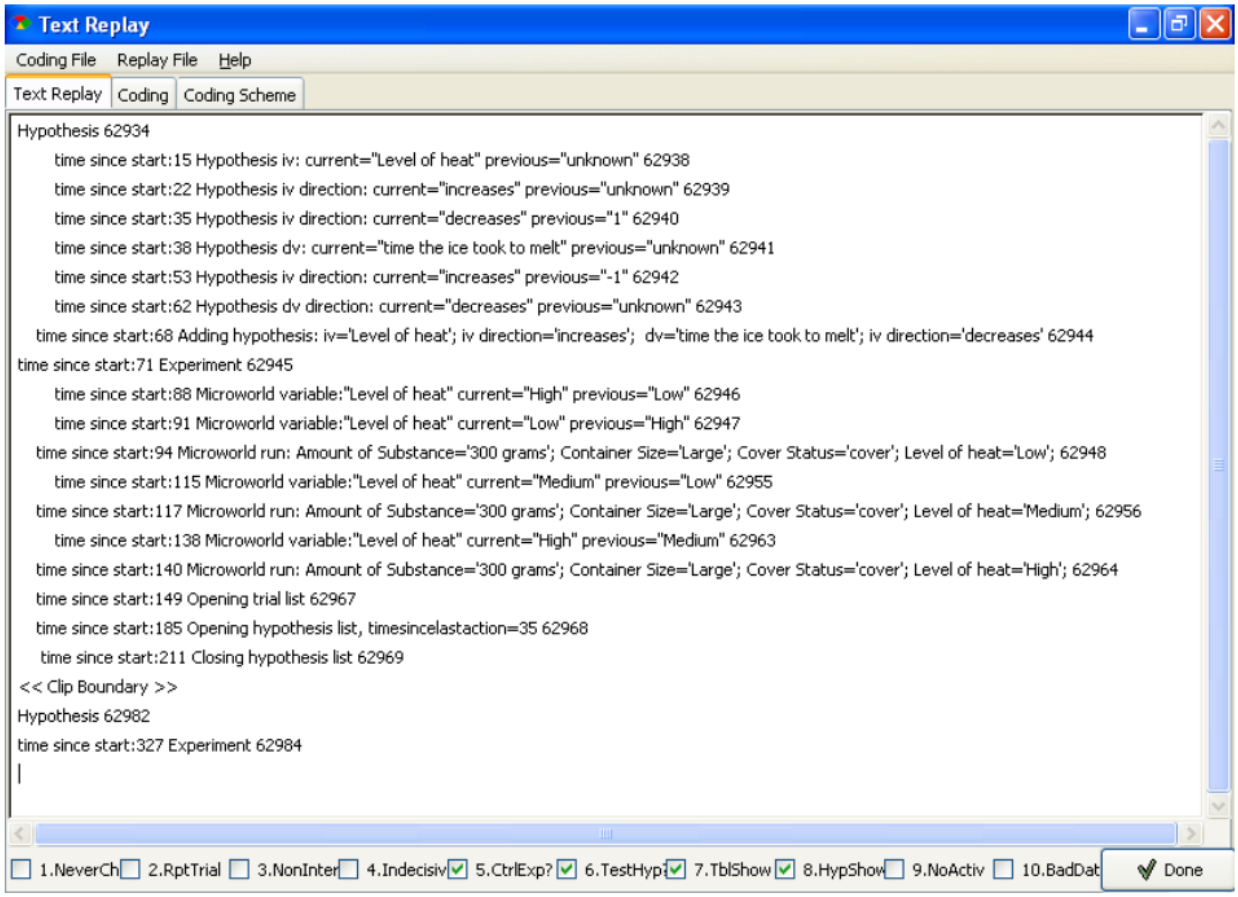

Figure 3. A text replay (readable depiction for data labeling) of student behavior in Inq-ITS.

The text replay infrastructure automatically samples which instances of student behavior will be coded by humans, in this case, stratifying the sample across students and across simulations (Sao Pedro et al., 2013a). Multiple coders label a sample of student behavior with reference to the constructs of interest and are checked for inter-rater reliability (San Pedro et al., 2013a). In this case, inter-rater reliability was good for labeling designing controlled experiments (Cohen's Kappa = 0.69) and perfect for testing the stated hypothesis (Cohen's Kappa = 1.00). A total of 571 sequences of student behavior were coded.

Next, a range of features of the student's interaction with the system was extracted from the raw data, including aspects of behavior such as how many times the student changed variables, the time between variable changes, repeated trials, and the degree of change between runs of the simulation. Features were filtered based on a domain expert's perception of which features would be most useful (Sao Pedro et al., 2012). The combined data set, including both features (as predictors) and labels made within text replays (as predicted variables), was then used as input to data-mining software. Multiple algorithms were tested to determine which made the best inference of the labels provided by the human coders from the predictor features. The algorithms were tested on data from entirely new students (Sao Pedro et al., 2013a), across the full range of contexts of use of the algorithm within the system (Sao Pedro et al., 2013a). The algorithms were also tested for validity within entirely different scientific domains. In this second validation, a model developed for a physical science simulation with simple relationships between the variables was validated

to work correctly when used in a biological science simulation, which contained more complex relationships (Sao Pedro et al., 2014). In the case of Inq-ITS, the algorithm that performed best was a relatively conservative decision tree algorithm, an unsurprising outcome given the relatively small data set. Overall, the model was able to distinguish appropriate science inquiry behavior from inappropriate science-inquiry behavior in a sequence of experimental trials 85% of the time.

These models functioned at the level of identifying whether a student demonstrated appropriate scientific-inquiry behavior in a single use of a simulation. They were also aggregated into broader inferences on student science-inquiry skill across simulations (Sao Pedro et al., 2013b, 2014). The broader models can then predict whether the student will be able to demonstrate the inquiry skill successfully in new situations. The model chosen for this was Bayesian knowledge tracing (BKT), a commonly used model for tracking student knowledge as it is changing (Corbett & Anderson, 1995). BKT takes a set of cases, each of which indicates whether a student is successfully demonstrating a skill, and aggregates the evidence over time into a running estimate of whether the student knows the skill and is likely to be able to demonstrate it in the future. The outputs from the inquiry behavior models were input into a BKT model, resulting in a model that could predict correctness on future simulations, as well as on future paper-based assessments of student inquiry skill (Sao Pedro et al., 2013b). This model was tested on new students and also across simulations (i.e., performance on one simulation was used to predict performance on another simulation – e.g., Sao Pedro et al., 2014). The model was able to predict future correctness for new students, in the same simulation, 74% of the time, and in different simulations 75% of the time.

Classification Based on Overall Inquiry Success

Our second example of inquiry modeling through machine learning involves a scenario in which it is not feasible to directly identify good and bad inquiry. Instead, we use external evidence that inquiry was successful as the basis for applying supervised machine learning. Specifically, in this example, we obtained data from students using a complex inquiry-learning environment where they had to answer a driving research question by collecting and interpreting data. In this case, our assumption is that the correctness of a student's answer to the research question would be highly correlated with productive inquiry behavior. Thus, we can use the scores from the students' submitted answer for labels, as the human-coded labels were used in the previous example, and analyze the data of student interaction with the environment to determine which student behaviors were associated with correct final answers.

This approach is used in work to model inquiry in virtual performance assessments (VPAs), which teach scientific-inquiry skills by presenting students with an authentic science problem in a virtual, simulated context (Clarke-Midura, Dede & Norton, 2011), shown in Figure 4. VPA provides students with the opportunity to interact with different scenarios in an immersive, three-dimensional (3D) environment. Students are asked to solve a scientific problem in each scenario by navigating an avatar through the virtual environment, making observations and gathering data. The avatar can collect several forms of evidence, conduct tests on it in a laboratory environment, talk to non-player characters, and read information kiosks.

Figure 4. VPAs, where students collect data, run experiments, talk to non-player characters, and read information resources in a rich, 3D virtual environment.

In this example, discussed in greater detail in Baker, Clarke-Midura, and Ocumpaugh (2016), we obtained data from a pair of VPA scenarios, referred to as the frog scenario and the bee scenario. In the frog scenario, students are asked to determine the cause of a frog mutation, specifically why the frog has six legs. The virtual world contains four farms with frogs and other evidence and where non-player characters provide competing opinions about the problem, a research kiosk where students can read about possible causes of the mutation, and a laboratory that contains dead, mutated frogs and lab water to use as comparisons. The student can collect tadpoles, frogs, and water samples from the farms, which they can then bring to the laboratory to run DNA and blood tests on the frogs and water-quality tests on the water samples. Possible explanations for why frogs are sick include parasites (the correct answer), pesticides, pollution, genetic mutation, and space aliens. Once students think they have collected enough information, they make a final claim for what they think is causing the frog mutation and support it with evidence. In the bee scenario, students are immersed in a similar environment, but in this context they are asked to determine what is killing off a local bee population (in this case, it is a genetic mutation).

Success in VPAs is identified by whether the student provides a correct final claim for why the frogs or bees are sick, and also by whether the student can correctly explain the chain of reasoning behind their claim. While it is possible to use rational or hybrid approaches to predict whether the student's behavior will be correct (e.g., Clarke-Midura & Yudelson, 2013); in this case, we engineered an automated model to predict whether the student behavior would be correct.

Our first step, as with Inq-ITS, was to distill a set of 48 semantically meaningful features (selected through structured brainstorming conducted by domain experts and machine-learning researchers) from the log files that contained data on all of the students' interactions with the environment. The features included behaviors such as moving from one virtual region to another, picking up or inspecting different objects, running laboratory tests on objects, reading informational pages at the research kiosks, and talking to non-player characters, as well as features regarding how long the student spent and details of the actions, such as how many different tests the students ran.

As in Inq-ITS, multiple algorithms were tested as potential mappings between the predictor features and classification labels. Algorithms were tested on data from entirely new students (Baker et al., 2016) to establish model validity. The models developed for each scenario were also validated to function effectively for the other scenario to increase confidence that these models would generalize to additional scenarios developed in the future for VPA. The algorithm that performed best was a relatively conservative decision rules algorithm (moderately more conservative than even the decision tree used in Inq-ITS). Overall, the model was able to distinguish appropriate science-inquiry behavior in a sequence of experimental trials 79% of the time. One surprising findings was that the time spent reading the various information pages was more predictive of the student eventually obtaining the correct answer than their specific exploratory or

experimentation behaviors in the system. Although surprising, this finding does not indicate that the virtual environment was not useful, but instead suggests that students may need declarative or expository information to best make sense of their activity within the virtual environment.

Pros and Cons

One of the biggest virtues of the machine-learning approach is that it can capture inquiry in complex settings where we don't know, a priori, exactly what good inquiry looks like. This is in contrast to the theoretically motivated approach discussed earlier. Many of the attempts to assess inquiry by more theoretical means require simplification of the inquiry task to allow inquiry skills to be recognized by clearly defined rules. While the theory-driven approach is well grounded, it risks incentivizing teachers and curriculum designers to create artificial tasks that teach students artificial specialized rules that cannot be used in many real-world inquiry tasks. To the extent that we want students to be able to use inquiry in murky, complicated, real-world situations, we need to create virtual environments where they learn inquiry strategies that are robust to situations and tasks that are not entirely straightforward.

Relatedly, this type of approach may be able to better distinguish inquiry strategies that are effective, even if not flawless, within simplified and rationally describable assessment tasks as well. Because the machine-learning approach is fundamentally data-driven, naturally occurring variations on successful strategies are more likely to be identified than in the case when content experts are required to define successful behaviors in advance. This allows machine learning to create models and assessments of student inquiry that apply more broadly and can be more flexible than purely theoretical approaches.

However, this flexibility carries with it limitations as well. Whereas a rule like VOTAT can be applied quickly across a broad range of learning systems, machine-learning models typically must be re-validated for different learning tasks, as was done above in both the Inq-ITS and VPA examples. In addition, machine-learned models are sufficiently complex that they can be harder to interpret by outsiders and do not always map back clearly to a theoretical understanding of the constructs. This can lead to questions of legitimacy and validity because it can be hard to prove that the models are not selecting for features that co-occur with successful inquiry within the given data set but are not actually relevant to good inquiry practice. Such possible spurious correlations could lead to models that validate within the original context but break down when used with a different population of students or systems. Additionally, machine learning can be more expensive than other approaches, both to create and validate models, and justify beyond the team that develops them.

Model-Driven Inquiry Detection

The previous two approaches focus on identifying specific feature markers in process data that indicate stronger or weaker application of inquiry or problem-solving skills. These methods stem from a scoring-oriented focus in which the motivating question might be: how would an expert rater recognize better applications of inquiry skills based on the data traces? In the model-driven approach, the focus shifts to the moment-by-moment decision making by the student. The motivating question here is quite different: how would we instruct an artificial agent to simulate the behavior of successful inquiry?

The model-driven approach attempts to create one or more generative models of within-task inquiry behavior based upon parameters that represent the latent traits we wish to make inferences about. The central feature of a model-driven approach is a mathematical model that does the following:

- links the latent-traits to be measured to the observable responses of the student,

- is based on a theory of how the latent-traits produce the responses, and

- predicts probabilities of specific responses as a function of both the person's latent traits and the response context (item or problem state).

Given such a model and sufficient performance data, the latent traits embedded in the model can be estimated using likelihood maximization methods. Given multiple competing models, each can be separately fit to the data and the inquiry performance can be classified based on the best fitting model. In this section, the approach is illustrated by a set of science-inquiry detectors formulated as Markov decision process (MDP) models that were developed to identify different inquiry strategies students applied to a chemistry simulation task.

Generative Models of Ideal Inquiry Behavior

The first step in the model-driven approach is to define models for the inquiry behaviors we wish to identify or evaluate. A single model of ideal behavior can be used as an expert model against which the student actions would be compared. In this case, we would estimate a single latent trait, which represents the student's ability to implement proper inquiry. Alternatively, multiple patterns of behavior can be modeled to provide more diagnostic information. Modeling both correct and incorrect strategies enables inference of not only how well the student is performing inquiry but also what misconceptions they might have or which strategies they might need to learn.

A probabilistic generative model will predict the probability of a student taking particular actions as a function of the current state of the simulation and the student latent traits. The model can be formulated to express the utility of each action as a function of the state of the experiment and the probability of choosing more useful actions as a function of the student ability. Because the action utility is dependent upon the current state of the simulation, the value of an action can change as students interact with the simulated experiment. For example, in a simulated experiment, we can imagine that a student has a choice of collecting more data under the current conditions, changing one of the experimental conditions, or analyzing the collected data. As the experiment progresses, collecting more data will become less useful and analyzing the data will become more useful. The utility of changing experimental conditions will depend upon the hypothesis being tested and the data that have already been collected.

One popular utility-based model is the MDP, which describes goal-driven behavior in a complex and possibly stochastic environment (Puterman, 1994). An MDP model relies on a definition of rewards, *R*, for achieving particular states along with costs for taking particular actions and a transition matrix *T*, which specifies the probability of transitioning from state *s* to *s'* given a particular action *a*. The reward structure *R* includes both a definition of the goal (the state which yields a high reward) and an encoding of motivation in the relative magnitudes of the goal reward and the cost of actions required to achieve the goal. The transition matrix *T* encodes beliefs about the problem space, in particular, giving the likely results of actions. MDP models are frequently used in the field of artificial intelligence for reinforcement learning (Barto, Sutton & Watkins, 1989) and have recently been used as psychometric models for estimating beliefs and ability from actions in complex tasks (Rafferty, LaMar & Griffiths, 2015; LaMar, under review).

MDPs for Inquiry Strategy Detection in the Concentration Simulation

This model-driven approach to inquiry assessment has been used with an experimental simulation-based assessment to infer inquiry skills based on student interactions with the simulation. The assessment involves mixing solvents and solutes and uses an embedded PhET simulation (Perkins et al., 2006), as shown in Figure 5. In a series of seven inquiry questions, students are asked about the relationship between amounts

of solute and resulting concentration in different mixtures. They are prompted to use the simulation to collect data and then are asked to respond and explain their response. Early cognitive labs showed that as middle school aged children interacted with the simulation they would run sequences of trials that corresponded to hypothesis-testing strategies. However, their strategies did not necessarily conform to the VOTAT-type strategy that the assessment designers had expected, nor did they confine themselves to a single strategy implementation. Instead a variety of strategies and strategy switching was commonly observed.

To identify the inquiry strategies used in particular interaction records while remaining resilient to occasional off-strategy behavior, MDP models were developed to embody both expected VOTAT strategies and additional strategies discovered in the initial data collection. Unlike the data-mining techniques described in the previous section, the generative-modeling approach requires all models to be defined in advance. The process of developing the candidate models, however, often include a combination of theoretical content knowledge and empirical discovery of student behaviors. For this study, information about student thought processes was gathered using cognitive lab protocols, in which students were asked to interact with the simulation and then explain their process to an interviewer. The combination of student-reported strategy use and the observed behavior of those students was analyzed in light of theory of student inquiry-skill acquisition to formulate preliminary strategy models. The models were then refined by using them to generate simulated student behavior and comparing the simulated actions to those taken by the original students and expert judgment of acceptable variation.

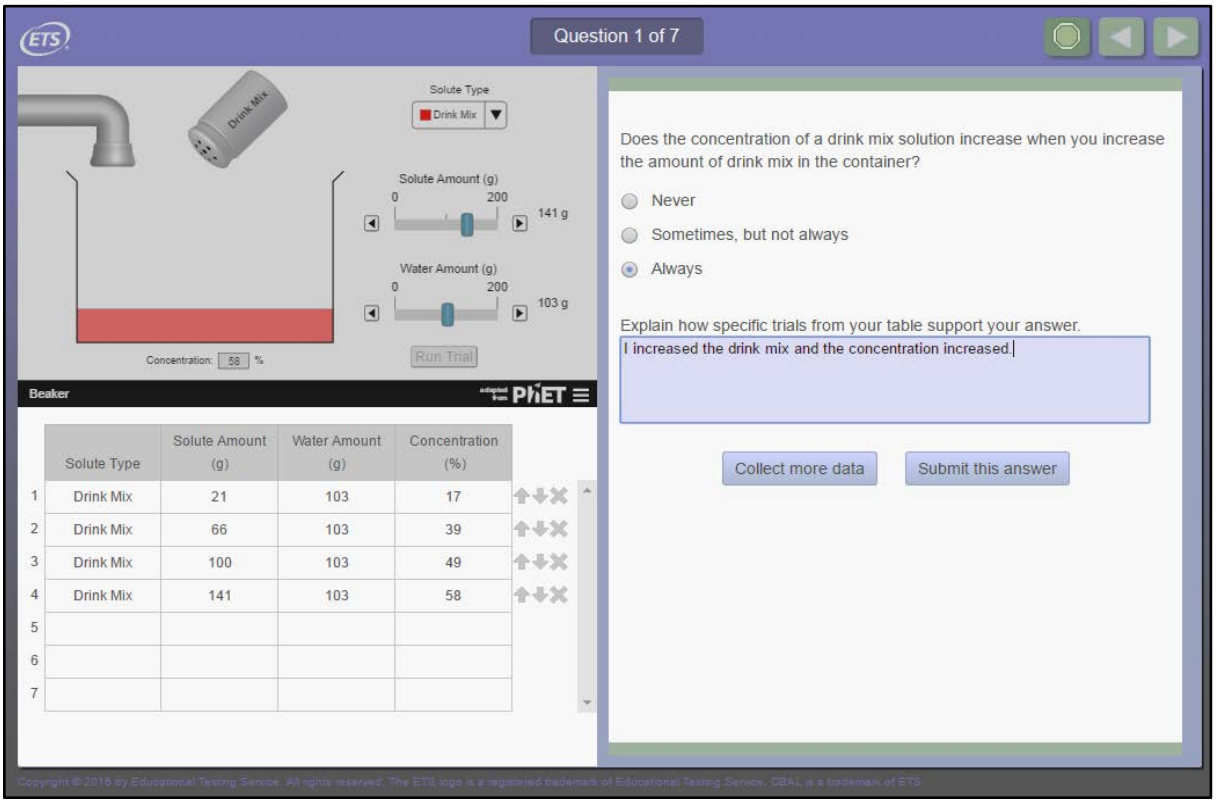

Figure 5. The second screen of item 1 for the concentration simulation.

To model inquiry behavior with an MDP, we needed to define the goal of the inquiry behavior, the sets of relevant actions and state variables, as well as the transition probabilities between states based on different actions. A single "ideal" inquiry model can be produced, which would allow comparison of student actions

with the expert model. This would result in a more score-oriented analysis where the student's inquiry skills can be estimated directly, assuming the expert model is the only valid strategy. Such an approach can be useful when the problem is fairly constrained such that there is a single correct goal with a preferred strategy for accomplishing the goal. The model would consider attempts to meet other goals as off-task behavior, resulting in a low estimate inquiry skill.

A more diagnostic approach can be taken by specifying different inquiry strategies as different MDP models. The models can then be fit to different sub-sequences of action data to estimate which strategies were being used at different parts of the inquiry process. Inquiry models can be developed for VOTAT strategies, directed search strategies, and other less productive approaches. As the goals and behaviors are quite specific, multiple productive and unproductive strategies can be modeled to provide both an overall inquiry score and diagnostic information that could be useful to adapt instruction.

Example Application of MDP Detectors

Based on data collected from 150 adult participants in an Amazon Mechanical Turk pilot study, three different inquiry strategies were identified and coded as MDPs for student interactions with question 1 (shown in Figure 5) and question 7, which was identical. Two of the strategies are different instantiations of the VOTAT strategy, one in which water is held constant and the solute is gradually increased (Increase Solute Strategy [ISS]) and the other in which water is held constant and the solute is gradually decreased (Decrease Solute Strategy [DSS]). For both ISS and DSS, the goal is to gather enough data to be able to answer the question. Because implementers of these strategies understand the importance of control-of-variables, data are considered to be a trial in which the amount of water is unchanged from the previous trial, but the solute amount is greater.

The third strategy involves a directed search in which students seek a saturated solution (Find Saturation Strategy [FSS]). The goal of students implementing FSS is to test whether the solution will indeed saturate. Students who implement this strategy are assumed to have a fair amount of content knowledge because they know that saturation is possible and they further know what conditions are likely to cause saturation. The typical behavior for this strategy involves setting the solute to a high level and dramatically decreasing the amount of water added until saturation is detected, often multiple times.

Note that all three of these strategies are productive inquiry strategies. While some non-productive behavior was observed based on student misconceptions, insufficient examples of those behaviors have been collected to formulate and test an appropriate MDP model.

To implement the identified inquiry strategies as MDPs, each strategy's goals were translated into a reward structure. Similarly, their beliefs were translated into action sets, state space variables, and transition functions. Table 1 shows the goals, rewards, and beliefs for ISS and FSS. The DSS strategy was similar to ISS with only data counted that decrease the solute rather than increase it.

Once the MDP generative models were built for the three different strategies, the models were fit to the record of trials run in the simulation to identify which strategies were most likely being used. For each student record for a particular item, the sequence of trials ran may contain zero, one, or more instances of a strategy implementation. To enable detection of strategy implementations at any point in the sequence and of any length, the trial sequences are split into all possible sub-sequences above the minimum length of 3 trials. Each of the inquiry strategy models is then fit to each candidate sequence, giving a likelihood that that particular sequence was generated by a student attempting to implement the inquiry strategy. Final strategy sequence labels were determined by maximizing likelihood over the entire record.

Table 1. Model components for the ISS and FSS models.

	Increase (Decrease) Solute Strategy	Find Saturation Strategy
Goal	Gather enough data to determine how increasing solute affects concentration at one water level.	To determine if this solution will saturate.
Rewards	Increase with: • More data • Range covered	Increase with: • Finding saturation • More data
Beliefs	Data = Two trials with an increase (decrease) in solute but constant water Water amount unimportant as long as it doesn't change	Data = Trial with high solute and low water Saturation = More than one concentration result of the same value

Using this method, log records from the first and last questions (1 and 7) were analyzed and sequences of implemented inquiry strategies were identified. For example, some students implemented multiple ISS strategies in a row (altering the amount of water in between), giving detected patterns of ISS-ISS-ISS, while others showed strategy switching such as ISS-FSS. Overall, the students who used more than one strategy implementation scored better on the following content question, indicating that more sophisticated inquiry patterns correlate with more successful conclusions.

Pros and Cons

There are a number of advantages of the model-based approach. As a theory-driven approach, the results are easily interpretable and can fit into existing frameworks for assessing science practice. Implemented as discrete single-strategy detectors, as demonstrated by the MDP detectors, multiple different inquiry strategies can be identified, including both productive strategies and those based on misconceptions, making the approach ideal for formative assessment and tutoring scenarios. The probabilistic nature of the models, meanwhile, allow for detection of strategy behavior even when the implementation of the strategy is imperfect. These factors make this approach useful in complex, open-ended inquiry tasks.

On the other hand, such complexity comes at a cost. The theory-driven modeling requires that theory exists to explain very low-level actions. Such fine-grained theory is often lacking in the current science literature. The formulation of MDP models also requires an understanding of how utility-theoretic state-space models work, making this method potentially difficult to access for science educators. Furthermore, the models that are developed are context-specific. While models for generalizable strategies, such as VOTAT, might be structurally similar in different tasks, each model needs to be customized to the actions and variables available in specific tasks.

Implications for GIFT and Tutoring

ITSs frequently include the affordance for simulation-based inquiry and problem solving (Murray, 2003). Principled assessment of student skills based on their actions within such tasks are critical to produce the relevant feedback and guidance that one expects from an advanced ITS.

For any of these approaches to be useful in a computerized tutoring environment, the identification and assessment of inquiry skills must be available in the moment, not merely in post-processing. MDP and BKT are not currently available in the GIFT framework, but their inclusion would be relatively straightforward.

In particular, future work to create flexible, lightweight versions of MDP algorithms could make them more accessible to tutoring systems for online processing. There is already support for the inclusion of automated detectors based on Rapid Miner into GIFT trainee models, making this type of algorithm readily usable within the GIFT framework. Theoretically defined behavioral indicators, meanwhile, can be easily coded into the logic of the task programming because their definition goes hand in hand with the task design.

Conclusions

Computers allow students to interact with more complex science and problem-solving scenarios and educators call for teaching and assessment to include more realistic science and engineering tasks. As a consequence, identifying and assessing inquiry strategies has become both pertinent and possible. This chapter has outlined three different approaches to assessing the science practice of hypothesis testing: a theory-driven approach based on carefully crafted assessment tasks and corresponding performance indicators, a more theoretical approach based on using machine learning to identify successful or appropriate inquiry behavior, and an approach that builds off theoretical understanding but uses generative process models to recognize a wider range of behavior.

Each of the approaches have their advantages and limitations, with the most appropriate approach likely dependent upon the type of task presented to the students and the types of skills intended to be taught and assessed. The theory-driven behavioral indicators are likely the best choice for high-stakes assessment because the behaviors detected are well understood and scoring is clean and defensible. The machine-learning approach provides a method for assessing inquiry in wide open environments and situations in which good inquiry cannot be cleanly defined. The model-based approach is something of a compromise between the pure theoretically defined behavioral indicators, which rely on consistently predictable behavior patterns, and the machine-learning inquiry detectors, which can discover patterns of behavior that experts might not predict. The MDP inquiry detectors are based in theory, although the theory can be developed iteratively with qualitative analyses of existing log files. Their probabilistically framed model allows strategy behaviors to be detected even when the strategy is imperfectly implemented.

References

Baker, R., Clarke-Midura, J., Ocumpaugh, J. (2016) Toward General Models of Effective Science Inquiry in Virtual Performance Assessments. Journal of Computer Assisted Learning, 32 (3), 267–280.

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

Baker, R., Siemens, G. (2014) Educational data mining and learning analytics. In Sawyer, K. (Ed.) *Cambridge Handbook of the Learning Sciences: 2nd Edition*, pp. 253–274.

Barto, A. G., Sutton, R. S. & Watkins, C. J. C. H. (1989). Learning and Sequential Decision Making. In *LEARNING AND COMPUTATIONAL NEUROSCIENCE* (pp. 539–602). MIT Press.

Clarke-Midura, J., Dede, C. & Norton, J. (2011). Next generation assessments for measuring complex learning in science. *The Road Ahead for State Assessments*, 27–40.

Clarke-Midura, J. & Yudelson, M. (2013) Towards Identifying Students' Reasoning using Machine Learning. *Proceedings of the 16th International Conference on Artificial Intelligence and Education*, 704–707.

Corbett, A.T., Anderson, J.R. (1995) Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. User Modeling and User-Adapted Interaction, 4, 253-278.Dede, C. (2010). Comparing frameworks for 21st century skills. *21st Century Skills: Rethinking How Students Learn*, *20*, 51–76.

Dede, C. (2010). Comparing frameworks for 21st century skills. *21st Century Skills: Rethinking How Students Learn*, *20*, 51–76.

Gobert, J. D. (2015). Inq-ITS: design decisions used for an inquiry intelligent system that both assesses and scaffolds students as they learn. *Handbook of cognition and assessment. New York: Wiley/Blackwell.*

Gobert, J.D., Sao Pedro, M.A., Baker, R.S.J.d., 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.

Greiff, S., Wüstenberg, S. & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education, 91*, 92–105.

Greiff, S., Wüstenberg, S. & Funke, J. (2012). Dynamic problem solving: a new measurement perspective. *Applied Psychological Measurement, 36*, 189–213.

Jonassen, D. H. (Ed.), *Learning to solve complex scientific problems*. New York: Lawrence Erlbaum.

Klahr, D. (2002). *Exploring science: The cognition and development of discovery processes.* Cambridge: MIT press.

Klahr, D. & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*(1), 1–48.

Kuhn, D. (2012). The development of causal reasoning. *WIREs Cognitive Science, 3*, 327–335.

Kuhn, D., Black, J., Keselman, A. & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction, 18*, 495–523.

LaMar, M. (under review). Markov decision process measurement model. Manuscript under review.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L. & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, *53*(2), 215–233.

Murray, T. (2003). An Overview of Intelligent Tutoring System Authoring Tools: Updated analysis of the state of the art. In *Authoring tools for advanced technology learning environments* (pp. 491–544). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-94-017-0819-7_17.

NGSS Lead States. (2013). The Next Generation Science Standards: For States, By States. Achieve, Inc. on behalf of the twenty-six states and partners that collaborated on the NGSS. Retrieved from http://www.nextgenscience.org/next-generation-science-standards.

OECD (2014). *PISA 2012 results. Creative Problem Solving*. Paris: OECD.

Paquette, L. & Baker, R.S. (under review) Comparing Machine Learning to Knowledge Engineering for Modeling SRL Behaviors: A Case Study in Gaming the System. Manuscript under review.

Perkins, K., Adams, W., Dubson, M., Finkelstein, N., Reid, S., Wieman, C. & LeMaster, R. (2006). PhET: Interactive simulations for teaching and learning physics. *The Physics Teacher*, *44*(1), 18–23.

Popper, K. R. (1959). The logic of scientific discovery. *London: Hutchinson*.

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. New York: John Wiley & Sons, Inc. Retrieved from http://dl.acm.org/citation.cfm?id=528623.

Rafferty, A. N., LaMar, M. M. & Griffiths, T. L. (2015). Inferring Learners' Knowledge From Their Actions. *Cognitive Science*, *39*(3), 584–618. https://doi.org/10.1111/cogs.12157.

Sao Pedro, M., Baker, R.S.J.d., Gobert, J. (2012) Improving Construct Validity Yields Better Models of Systematic Inquiry, Even with Less Information. *Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UMAP 2012)*, 249–260.

Sao Pedro, M. A., Baker, R. S. & Gobert, J. D. (2013a). 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.

Sao Pedro, M.A., Baker, R.S.J.d., Gobert, J., Montalvo, O. Nakama, A. (2013b) 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.

Sao Pedro, M., Jiang, Y., Paquette, L., Baker, R.S., Gobert, J. (2014) Identifying Transfer of Inquiry Skills across Physical Science Simulations using Educational Data Mining. *Proceedings of the 11th International Conference of the Learning Sciences*.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53–76. https://doi.org/10.1016/j.asw.2013.04.001.

Vollmeyer, R., Burns, B. D. & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science, 20*, 75–100.

Wieman, C. (2015). Comparative cognitive task analyses of experimental science and instructional laboratory courses. *The Physics Teacher*, *53*(6), 349–351.

Wüstenberg, S., Greiff, S. & Funke, J. (2012). Complex problem solving. More than reasoning? *Intelligence, 40*, 1–14.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*, 172–223.