Towards General Models of Effective Science Inquiry in Virtual Performance Assessments

Baker, R., Clarke-Midura, J., Ocumpaugh, J.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Abstract

Recent interest in online assessment of scientific inquiry has led to several new online systems that attempt to assess these skills, but producing models that detect when students are successfully practicing these skills can be challenging. In this paper, we study the models that assess student inquiry in an immersive virtual environment, where a student navigates an avatar around a world, speaking to in-game characters, collecting samples, and conducting scientific tests with those samples in the virtual laboratory. To this goal, we leverage log file data from nearly two thousand middle-school students using Virtual Performance Assessment (VPA), a software system where students practice inquiry skills in different virtual scenarios. We develop models of student interaction within VPA that predict whether a student will successfully conduct scientific inquiry. Specifically, we identify behaviors that lead to distinguishing causal from non-causal factors to identify a correct final conclusion (CFC) and to design a causal explanation (DCE) about these conclusions. We then demonstrate that these models can be adapted with minimal effort from one VPA scenario (where students must identify the cause of mutations in a frog population) to a new VPA scenario (where students must identify what is killing a bee population). We conclude with discussions of how these models serve as a tool for better understanding scientific inquiry in virtual environments, and as a platform for the future design of evidence-based interventions.

Keywords: virtual environment, educational data mining, learning analytics, scientific inquiry skill, prediction modeling, designing causal explanations, virtual performance assessment

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Introduction

The last decade has seen an explosion of online learning systems designed to teach and assess science inquiry skills. Many of these systems provide scaffolded simulations that facilitate measurement (Sao Pedro et al., 2012, 2013; Quellmalz et al., 2012). Others have exploited the properties of virtual worlds, allowing students to conduct science inquiry in a more contextualized-environment (Rowe et al., 2009; Clarke-Midura, Dede, & Norton, 2011). In such systems, students navigate their avatars around an immersive virtual environment where they are able to interact with in-game characters, collect data samples, and conduct scientific tests on those samples in a virtual laboratory. Unlike systems in which a student follows a prescribed set of steps to complete an experiment, these virtual environments provide students with greater autonomy, allowing them the freedom to control when their avatar visits different virtual locations and what their avatar does to learn about the problem assigned. While there may be one desired outcome, there are several paths in which to get there, making students responsible for their own instructional decisions (e.g., Lajoie, 1993; Azevedo, 2005).

These contexts offer several advantages for learning and assessment, many of which have been outlined in previous research (Dalgarno & Lee, 2010; Van Joolingen, De Jong, & Dimitrakopoulou, 2007; Gee, 2003; Azevedo, 2005). They better mirror the kind of real-world work of scientists, offering a level of autonomy that may increase engagement and inductive learning opportunities. They also provide a more complex environment where students must search for information and exercise more complex meta-cognition. Inasmuch as these environments facilitate student-to-student interaction, they create more opportunities for the kind of social affordances advocated by Kreijns & Kirschner (2001). These properties make it possible to answer the call of researchers like Webb, Gibson & Forkosh-Baruch (2013), who

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

have encouraged developers to harness “the benefits of embedded continuous unobtrusive measuring of performance while learners are engaged in interesting computerized tasks designed to support their learning,” including the sorts of micro-adaptions described by Kickmeier-Rust & Albert (2010) in their research on digital educational games. At the same time, it can be more challenging to assess students in these conditions than in simulations where student actions are more restricted.

Recent studies have modeled scientific inquiry skills using both learning analytics/educational data mining (LA/EDM) (Baker & Siemens, in press) and knowledge engineering (KE). Many of these models have been developed for environments where student behaviors are highly restricted (e.g. Sao Pedro et al., 2010, 2013 and Quellmalz et al., 2012), but this work has also been extended to virtual environments that offer greater autonomy, including some with many with game-like features (e.g. Rowe & Lester, 2010; Sabourin et al., 2012; Sil et al., 2012; Clarke-Midura & Yudelson, 2013). For example, Rowe and Lester (2010) used KE to model narrative knowledge, strategic knowledge, and content knowledge in a serious game where students explore a virtual world to solve a scientific problem. Sil et al. (2012) used EDM to develop models of inquiry skills and content skills based on the linguistic features of student essays and students interactions within their serious game/virtual world. Clarke-Midura and Yudelson (2013) used EDM to compare a human scored rubric to a machine scored algorithm of students’ causal reasoning.

EDM techniques are considered particularly useful for discovering complicated patterns in ill-defined domains (Lynch, Ashley, Aleven, & Pinkwart, 2008). For that reason, this paper uses EDM to develop automated detectors to predict whether a student will successfully conduct science inquiry in a virtual environment. Specifically, this paper develops models of successful

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

inquiry behaviors within the context of Virtual Performance Assessments (VPA), which assesses whether a student successfully conducts scientific inquiry by presenting students with an authentic science problem in a virtual, simulated context (Clarke-Midura, Dede, & Norton, 2011). VPA exploits the properties of a virtual world, providing students with the opportunity to experience the exploratory nature of science inquiry, where there are multiple paths to correct answers.

There has been considerable concern about the assessment of science inquiry skills, which often requires considerable time and effort on the part of the teacher (Au, 2007; NRC, 2006). Frequently, these assessments are based solely on outcomes, making it difficult to link those results back to the kind of instruction and experiences necessary to improve student learning (Black & Wiliam, 1998). Because many assessments of science inquiry, including those constructed by teachers, reflect this outcome-based approach, the data needed to assess students' thinking processes (Black & Wiliam, 1998) and develop cognitive models of student inquiry (e.g., NRC, 2006; Pellegrino, Chudowski, & Glaser, 2001; Leighton & Gierl, 2007) can be difficult to acquire. Yet these are precisely the kinds of data required if we are going to be able to understand inquiry well enough to produce the sorts of "training wheels" (e.g. Bannert 2000) needed to improve motivation and to develop autonomous, real-world science inquiry skills among our students. Recent work has addressed similar issues by exploring processes related to self-regulation in other online learning domains (e.g. Schoor & Bannert, 2012), but it is also important to determine what behaviors are associated with (and can increase the development of) these complex skills.

The design of VPA allows us to address this gap, providing teachers and researchers with more detailed information about the behaviors and cognitive processes that do (and do not) lead

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

to improved performance of science inquiry. By assessing student performance at conducting scientific inquiry in a setting that allows for complex inquiry behaviors, we collect evidence on whether or not a student demonstrates scientific inquiry skill. Specifically, VPA provides us with an opportunity to harvest the sort of data that can lead to improved designs of both formative assessment and of real-time interventions for struggling students. That is, most students would benefit from targeted and detailed feedback during learning (cf. Narciss & Huth, 2004), but automating that process requires us to identify the learning behaviors that are most useful for predicting successful performances of science inquiry skills.

In this paper, we use EDM methods to study the interactive behaviors associated with successful performance of science inquiry in VPA. A successful student is able to (1) correctly answer the overarching question presented in a virtual scenario (i.e. “Why are the frogs in this farm experiencing mutations?”) and (2) appropriately justify their answer based on causal evidence from within the game (i.e. “What evidence justifies the conclusion that X is causing the mutations among this frog population?”). Each of these models predict whether a student will ultimately demonstrate these skills based upon their earlier interactions within each VPA scenario, a step towards developing the kinds of cognitive models required to design interventions and provide real-time scaffolding. We then study whether these models can be generalized from one virtual scenario (where students investigate frog mutations) to another (where students investigate bee deaths), demonstrating that these models transfer between these two scenarios with very little degradation, findings which offer both practical and theoretical implications.

Models of the behaviors associated with successful scientific inquiry will ultimately help to highlight which student behaviors demonstrate learning, facilitating the development of

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

targeted, automated interventions to modify student behavior. Unfortunately, developing automated detectors can be cost-intensive, and development costs are already quite high for virtual worlds. By showing that the EDM techniques used to model science inquiry performance can be generalized from one scenario to another with minimal changes, we offer a first step towards overcoming this obstacle. That is, we have intentionally designed two VPA scenarios that show different surface problems, but in fact assess the same constructs. This allows us to measure whether we can re-map the same models across virtual assessments, allowing for effective assessment to be more quickly and easily replicated. This approach is both feasible and quick to apply because of VPA's design, which considerably facilitates model transfer compared to prior examples in educational data mining (c.f. Sao Pedro et al., 2012).

Virtual Performance Assessments (VPA)

Virtual Performance Assessments (VPA) are online assessments developed at the Harvard Graduate School of Education in order to study the feasibility of using immersive virtual environments to assess middle school students' performance of science inquiry skills (see http://vpa.gse.harvard.edu). While originally designed as summative assessments, this research develops models that can be used to support formative assessment within VPA. Designed using the Unity game development engine (Unity Technologies 2010), VPA provides students with multiple online assessment scenarios where they interact individually 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.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

VPA's design permits assessment of the kind of authentic practices that facilitate science inquiry (see Shute, 2011 and National Research Council, 2011). As Figure 1 shows, this immersive software mimics the design of many 3-D videogames, but each virtual scenario requires students to apply the scientific method to solve a problem. Unlike more traditional assessments, VPA scenarios simulate real-world environments where there are choices and extraneous information that can lead students down a wrong path. This design provides data on student strategies that are not available within multiple-choice assessments.

[Insert figure 1 here]

This study considers two VPA scenarios, which we refer to as the frog scenario and the bee scenario. In the frog scenario, students are immersed in a virtual world where they must determine the cause of a mutation that results in six-legged frogs. This scenario contains 4 farms, a research kiosk where students can read about 5 possible causal factors, and a laboratory where they can conduct tests on water quality, frog blood and DNA. On each farm, students interact with non-player characters (NPCs) who provide competing opinions about the cause of these mutations. They can also collect tadpoles, frogs, and water samples to test in the laboratory. Possible explanations for the mutations include parasites (the correct causal factor), pesticides, pollution, genetic mutation, and space aliens. Once students think they have sufficient data, they submit a final claim about the cause of the mutation and provide supporting evidence.

In the bee scenario, students immersed in a similar environment are asked to determine what is killing a local bee population. As in the frog scenario, there is a village with four farms, a research kiosk, and a laboratory where they can conduct tests on the nectar, DNA, and protein in

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

the bees. Students can interact with NPCs on each farm who provide conflicting information about the problem, and they also have the ability to collect three types of data that they can then test in the laboratory: larvae, bees, and nectar. The possible explanations for bee deaths include genetic mutation (the correct causal factor), parasites, pesticides, pollution, and space aliens. As with the frog scenario, students have the autonomy to decide when they have sufficient evidence. At that point, they submit their final claim and support it with evidence.

As introduced earlier, the activities in each VPA scenario are deliberately similar, allowing researchers to assess performance of the same inquiry practices in different contexts. Students answer different questions in a different virtual environment, but they perform the same kinds of virtual tasks and undergo similar assessments of learning. These two assessments are structurally similar, even though the contexts (surface level problems and features) of the scenario are quite different. This allows researchers to quickly map the surface features used for developing models in one scenario to the equivalent surface features in another scenario, as shown in Table 1. VPA has been validated in previous research to ensure that it is assessing the performance of science inquiry (Scalise & Clarke-Midura, 2014; Clarke-Midura, McCall, & Dede, 2012; McCall & Clarke-Midura, 2013). Students must understand the causal factors and which data provides evidence for or against the particular claims in each scenario. Information about how these inquiry performances are assessed is presented in Table 2.

[Place Table 1 approximately here]

[Place Table 2 approximately here]

In each VPA, students are free to choose which tasks to carry out and in which order. As in real-world inquiry, each scenario includes extraneous information and the ability to pursue

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

paths that are not helpful for the task at hand. Student interactions with the virtual environment are recorded by the system and stored as log files, facilitating *in situ* evaluation of science inquiry practices. These log files provide researchers the opportunity to stealthily collect and assess process data (student activities) and product data (their final claims). As such, these log files provide richer data on student inquiry processes than might be obtained through more obtrusive observation techniques.

Data and Methods

Students

This study analyzes VPA log files produced by middle school students (grades 7-8) in 138 science classrooms (40 teachers) across the Northeastern and Midwestern United States and Western Canada. The study took place at the end of the school year and was meant to assess inquiry skills students had learned over the course of the year. Students were randomly assigned to begin either the frog or bee scenario; each student was then assigned the other scenario two weeks later. Students were not given any training in virtual environments but were shown a short introductory video prior to each assessment that provided instructions. Students worked within each scenario until they had completed the analysis and produced a final answer for its underlying problem (e.g. why do these frogs have extra legs or why are these bees dying), spending an average of 29 minutes and 29 seconds in the frog scenario (SD = 14 minutes, 30 seconds) and 26 minutes and 5 seconds in the bee scenario (SD = 12 minutes, 27 seconds). Approximately 2000 students completed each scenario (N = 1,985 for the frog scenario, N = 2,023 for the bee scenario), with 1,579 students completing both scenarios. Students generated a

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

total of 423,617 actions within the frog scenario and a total of 396,863 actions within the bee scenario.

Data Logs and Feature Distillation

The log data from VPA includes information about the kinds of actions students perform within each scenario, such as moving from one virtual region to another, picking up or inspecting different objects (e.g. frogs, water, bees, nectar), running laboratory tests (e.g. blood tests, DNA tests, water/nectar quality tests), reading informational pages at the research kiosks, and talking to non-player characters. The log data also record each action's timing and location within the virtual environment and other details about the objects manipulated, the tests run, and the information gathered through reading tasks and interaction with NPCs. The logs of these interactions provide data on students' inquiry processes, but they also provide product data, specifically students' claims about causality and details about the evidence they use to support those claims.

For the purposes of analyses, raw log process data was distilled into 48 semantically meaningful features which either applied to both scenarios (e.g., how many trips a student took the virtual lab) or could be easily substituted from one scenario to the other (e.g., features relating to live bees were treated the same as features relating to healthy frogs, protein tests were treated the same as blood tests, nectar was treated the same as water, and nectar tests were treated the same as water tests). The process of feature engineering is a time-intensive process, particularly when it is the first time that intensive feature engineering has been conducted for a specific class of interactive learning environment (Veeramachaneni, O'Reilly, & Taylor, 2014). However, once thorough feature engineering has been completed for the first time for a specific

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

class of interactive learning environment, the features can often be extensively reused, even in other projects. These features were derived from the log data and sometimes required looking at large periods of time. For instance, individual moves between zones and times spent in zones were distilled into the number of times the learner moved between zones and the time spent in each type of zone (e.g., farms or the lab). Individual laboratory tests (such as a genetic test on a bee) were distilled into counts of how many times the student ran each test. Individual instances of accessing items within the research kiosk or conversing with NPCs were distilled into the total number of times the student accessed each information resource and the total amount of time the student spent on each. These features were then tested as potential predictors of student inquiry skill, and 29 were automatically selected by one or more of the algorithms discussed below. For purposes of analysis, the correct cause in the frog scenario (parasites) was mapped to the correct cause in the bee scenario (genetic mutations). (Tables 3 and 4, presented in our discussion of results, detail the 29 features used to develop and replicate these models.)

Operationalizing Student Success at Inquiry

In highly constrained learning environments, successful science inquiry is often operationalized by identifying known strategies for designing controlled experimental trials (see Sao Pedro et al., 2012, 2013) such as altering one variable at a time (Tschirgi, 1990; Chen & Klahr, 1999). Operationalizing success in less-constrained environments is more challenging, since both successful and unsuccessful solutions may be derived through multiple paths. Previous research on VPA has demonstrated its validity and reliability in assessing science inquiry skills (see Scalise & Clarke-Midura, 2014). In this study, we operationalize student success in the frog scenario and the bee scenario of VPA as the same two measures: (1) a binary (0,1) assessment of

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

whether the student arrived at the correct final conclusion (CFC) and (2) a numerical assessment of the student's performance when asked to design causal explanations (DCE). Measuring whether a student can design causal explanations for why the correct answer is correct, as well as whether they obtain the correct answer, is an important step towards understanding how deep the student's inquiry is; in real-world situations, it is often insufficient to simply know the right answer, it is necessary to justify it to others as well.

For CFC, we identified the students who correctly identified that parasites were causing the frog mutations (28.3%) and the students who correctly identified that genetic mutation was causing bee deaths (29.6%). These conclusions are treated as correct because they are the only hypotheses in each environment without counter-evidence. That is, reaching these conclusions demonstrates that the student has successfully distinguished between causal and non-causal data within the virtual environment, a skill which Kuhn et al. (2000) refers to as the "ability to read data." When only this outcome-based, binary measure is considered, very few students can be said to demonstrate that ability.

In contrast, DCE was operationalized with a point system that awarded partial credit to students who identified the evidence that supported their particular claim, regardless of whether they had reached the correct final conclusion. Students were first asked to identify which data they collected and which tests they ran in the virtual laboratory could support their claim. They were then asked about other available evidence, giving students further opportunities to provide observations that could support their claims. For example, questions about the virtual environment's geography probed students' ability to understand potential environmental effects. These indicators were aggregated into a single measure, resulting in a mean DCE score of 50.00% (SD = 23.33%) for the frog scenario and a mean DCE score of 46.11% (SD = 21.40%)

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

for the bee scenario. This information complements the CFC score, allowing us to better distinguish students who had been lead astray by distracter information (but who understand the principles of scientific inquiry) from those who were completely unsuccessful at demonstrating science inquiry skill. This measure is kept as a numerical variable rather than reducing it to a binary variable, as there is additional information in the degree to which student performance is partially correct.

Constructing Models using Learning Analytics/EDM Techniques

Detectors of both predicted variables were constructed using Rapid Miner 4.6 (Mierswa et al., 2006). The model predicting whether the student reached the correct final conclusion (CFC) was constructed as a binary classification problem. This means that we attempted to model CFC with algorithms that automatically determined which subset of features and patterns combining them predict whether the student is correct (1) or incorrect (0). For each scenario, algorithms that had performed well in previous EDM research were tested, including J48 Decision Trees (Quinlan, 1993), JRip Decision Rules (Cohen, 1995), Step Regression, and K* (Witten & Frank, 2005). Feature selection was conducted by the algorithms themselves; no additional feature selection outer-loop was used. For brevity, only the algorithm with the best performance is reported on in the results.

In conjunction with Leave One Out Cross-Validation (LOOCV) technique (Witten & Frank, 2005) which was applied at the student level, 2 performance metrics (Kappa and A') were used to evaluate potential CFC models. These performance metrics capture two different aspects of a model's success. Kappa (Cohen, 1960) assesses the degree to which a model is better than chance at identifying which student behaviors are associated with ultimately submitting a correct final conclusion. For Kappa, chance performance is 0 and perfect performance is 1; a Kappa of

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

0.31 indicates that the model is 31% better than chance. A' (Hanley & McNeil, 1982) is the probability that when a model is presented with a student who made a correct conclusion and a student who made an incorrect conclusion, the model can correctly distinguish which is which. For A' chance performance is 0.5 and perfect performance is 1.0. Because of flaws in the RapidMiner 4.6 implementation of A', it was calculated with code available at http://www.columbia.edu/ ~rsb2162/edmtools.html.

Because the DCE assessment was numerical, rather than binary, both its construction and evaluation differed from that of CFC. The DCE model was constructed using linear regression. Linear regression is a relatively conservative algorithm that is unlikely to produce an over-fit model (e.g. a model that fits the noise in the data as well as the signal), and for this study, it was implemented using the M5' feature selection procedure (Wang and Witten, 1997) within RapidMiner 4.6. As is also standard, correlation was used as a performance metric for this model.

Readers should note that the performance metrics used to evaluate the optimal models for each assessment (CFC and DCE) were necessarily different. Cohen's Kappa and A' cannot be applied to numerical data (e.g., DCE), and the correlation metrics rest on a set of assumptions that are typically considered inappropriate for binary (0,1) data (Landis & Koch, 1977; Witten & Frank, 2005).

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Results

Models of Correct Final Conclusion (CFC)

Frog Scenario CFC

For the frog scenario, a model developed with the JRip Decision Rules algorithm (Cohen, 1995) performed best (Kappa = 0.548, A' 0.79), achieving results comparable to detectors of science inquiry skill that were developed in much more constrained simulation environments (e.g. Sao Pedro et al., 2010, 2013). The other algorithms attempted (listed above) achieved lower Kappa and A'. Qualitative analysis of this model shows that all of its features are based on the amount of time students spent reading the various information pages available within VPA. As shown in Table 3, students who spent more time reading about evidence that supported the correct hypothesis (that parasites were causing the frogs to grow extra legs) were most likely to ultimately select that option, while those who were distracted by evidence supporting incorrect hypotheses selected other options. It is worth noting that the initial set of significant features is winnowed considerably during model development. This is due to two factors: (1) the relative conservatism of the JRip algorithm, which tends to find simpler (and thus less over-fit) models than competing approaches and (2) the relatively high inter-correlation between features. Many of the features that were not selected captured the same variance in predicting CFC and were therefore unnecessary once a more optimal feature was selected.

Application of Frog Scenario CFC to Bee Scenario

In order to determine the generalizability of the CFC model developed for the frog scenario, this model was tested on data from student interactions with the bee scenario. Context specific features were remapped (e.g. the substitution of reading about viruses for reading about genetic mutations, and so on) before the frog scenario's detector was applied to the data from the

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

bee scenario. Remapping of conceptually similar features has not, to our knowledge, been previously used in generalizing automated detectors between complex scenarios in online learning environments; it is a much simpler and cheaper approach than building entirely new detectors. Results indicate that this model worked well with the new data, achieving a Kappa of 0.332 and an A' of 0.67. (These values are not cross-validated, as they represent application of the model to a new data set, which is comparable to cross-validation; cross-validation is not feasible when applying a model to an entirely new data set.) This represents moderate degradation from the Kappa of 0.548 and A' of 0.79 for the frog scenario but it is still substantially better than chance-level performance, suggesting that the model will be relatively robust to moderate design changes in future VPA scenarios. Indeed, these Kappa and A' values are comparable to models of ill-defined student affect that are predictive of student success years later (e.g. San Pedro et al., 2013).

Bee Scenario CFC

When a new CFC model (not trained on the frog scenario) was constructed for the bee scenario using the same approach as above, its performance was only slightly better (Kappa = 0.417, A' = 0.70) than the performance of the frog scenario model for this data set, and its features were quite similar. As shown in Table 3, the features for the newly constructed model (trained on the bee scenario) are quite comparable to those from the CFC model developed using data from the frog scenario. In both models, students were most likely to achieve the correct final conclusion when they spent more time reading about supporting evidence and less time reading about contradictory evidence. In fact, every feature in the two models involved the amount of time or number of times the student read pages about supporting or contradictory evidence; no other features were selected by either model. We return to this issue in the discussion section.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

[Place Table 3 approximately here]

Models of Designing Causal Explanations (DCE)

Frog Scenario DCE

When DCE was modeled with data from the frog scenario, the final model contained 20 features and achieved a cross-validated correlation of 0.531. As All 20 features are positively correlated with DCE when considered individually, but as Table 4shows, some are negatively correlated in the full model. The switch in sign reflects some degree of collinearity among the variables; allowing this collinearity reduces interpretability of individual feature coefficients but improves model fit, even under stringent cross-validation.

However, since we know that the features will positively correlated with DCE when considered individually, we know that students who pick up more objects, run a broader range of tests, and spend more time reading information pages are likely to have higher values for DCE. Also, despite the negative correlations in the full model, the positive correlations when studied individually tell us that (for example) students who run more genetic tests on non-sick frogs are likely to have higher values for DCE, that students who place more objects in their backpacks are likely to have higher values for DCE, and that students who spend more time exploring farms are likely to have higher values for DCE. The large number of features captured in the DCE model reflects the complexity of DCE skill, showing that it correlates with a range of information-seeking behaviors that extends well beyond those associated with student performance at CFC.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Application of Frog Scenario DCE to Bee Scenario

As with the CFC detectors, the frog scenario's DCE was applied to the bee scenario to test its generalizability, with results (r = 0.401) indicating that the model will be relatively robust to the introduction of new scenarios within VPA. (This correlation is not cross-validated, as it represents an application to new data, which is comparable to cross-validation; cross-validation is not feasible when applying a model to an entirely new data set.) As we found with the CFC detectors, the frog scenario model's performance on the new data demonstrates only modest degradation from that achieved for the original data set (where cross-validated r = .531).

Bee Scenario DCE

When an entirely new model of DCE was developed for the bee scenario (allowing the model to select any features, including those not used in the frog DCE model) performance was comparable to the original performance of the frog model on the frog scenario. The new model achieved a cross-validated correlation of 0.527, essentially equivalent to that achieved for the frog scenario (r = .531) and better than the application of the frog DCE model to the bee scenario (r = .401). Results (Table 4) also show that even though the algorithm selected features without reference to the frog scenario's DCE model, the 19 features selected for the new model were remarkably similar to those in the frog scenario. Approximately 75% of the features in the bee model of DCE map directly to those selected in the frog model of the same construct. As above, it is worth noting that some features switched direction in the full model; this is a normal aspect of the collinearity that occurs with highly complex models. While traditional statistical paradigms try to avoid collinearity to increase interpretability, data mining models tend to include it as it often leads to better fit (even when a model is applied to new data, as here).

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Discussion and Conclusions

Assessing and supporting science inquiry in open-ended virtual environments like Virtual Performance Assessments (VPA) offers educators and researchers the opportunity to monitor science inquiry *in situ*, providing diverse contexts for students to apply, develop, and demonstrate these skills. Finding new models for assessing student performance at inquiry creates new opportunities for the future of assessment.

Although there are clear advantages to assessing science inquiry through systems that offer students greater autonomy, these designs are not without challenges, particularly when it comes to detecting struggling students. Assessing ill-defined science inquiry skills becomes more challenging when there is less scaffolding, as scaffolding can facilitate interpretation of student behavior (e.g. Sao Pedro et al., 2012, 2013). Moreover, monitoring the performance of these skills at scale can prove daunting, especially if the detection system that monitors them must be completely redesigned each time a new context is offered. Yet this is exactly the sort of monitoring that must take place if we are to provide the kind of just-in-time interventions (e.g., Kester et al., 2001; Conlan et al., 2012) that struggling students often need early in the process of developing autonomous inquiry skills.

In this paper we have attempted to chip away at several of these problems, offering models with state-of-the-art performance that predict whether a student will successfully (1) arrive at the correct final conclusion (CFC) and (2) design causal explanations (DCE). These models offer a first step towards understanding the processes successful students go through when conducting scientific inquiry in a complex environment. Moreover, while these models are clearly sensitive to contextual factors (a common problem in many educational interventions, c.f. Dede 2006; Dede et al., 2005), we have shown both that models developed independently for

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

two different VPA scenarios are quite similar and that models developed for one scenario can be applied to a new VPA scenario with minimal modifications and very little degradation. Our approach for remapping based on structural similarities is quick to apply and makes it relatively easy to generalize models between scenarios in the same learning system. It is possible that scenarios that differed in more fundamental ways would need further development for models to generalize; understanding exactly how far the current models will generalize is a valuable area for future research, both in terms of investigating techniques for scaling automatic assessment and in terms of developing cognitive models of science inquiry skills.

The models produced here were generally successful, producing performance comparable to models developed for much more constrained environments (e.g. Sao Pedro et al., 2010, 2013). However, as with any model, there is always room for improvement, particularly with further feature engineering. For example, greater attention to student self-explanation behaviors during data collection and lab work, greater attention to the contexts in which student kiosk reading occurred (e.g. what behaviors did it precede or follow), greater attention to student movement through the environment, and greater attention to the patterns in experimentation and data collection behavior shown by different students, might have produced improvements in eventual model quality. To some extent, feature engineering can go on indefinitely, and deciding when to stop is a judgment call, but it is likely that greater attention to the complexities of student behavior would have led to some model improvement.

Another area where the models could be improved is in making them more usable at run-time. While the factors they identify can already be used as the basis for intervention to some degree, it is possible to build iterative versions of these models that make predictions, from the same features, after a specific number of minutes. Doing so would involve re-fitting the models

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

repeatedly with subsets of data (first minute, first five minutes, first ten minutes, etc.). This process has been carried out for models predicting robust learning in the domain of genetics (Baker, Gowda, & Corbett, 2011), and would be feasible here. This would be a useful area of future work.

One interesting finding is the prominent role that reading about causal factors and evidence plays in the CFC model. This finding shows that even in very new media and interactions such as virtual environments, where students can collect and analyze evidence, critical reading skills are still very important. However, students' success at designing causal explanations (DCE) was associated with a more complex range of student behaviors, indicating that the sophisticated interactions that VPA affords connect more strongly to deep reasoning strategies. These differences demonstrate the challenges of assessing science inquiry through more traditional methods that focus exclusively on product data. Clearly it is important to distinguish between students who have not sought out the right opportunities to collect data from those who are not able to apply it correctly, something which we cannot do if we only consider whether students can figure out the answer to the main question.

There are several ways to apply the models developed here to both automated adaptation and assessment at scale. VPA is already used by thousands of students; the models used here are quick to run in an online system and could conceivably be applied to millions of students. By using automated detection of inquiry skills, feedback and support can be incorporated into the narrative of the virtual world, providing students with real-time feedback, more quickly than it could be provided by individual teachers, who would have to deliver feedback from outside the environment or after the fact.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

For example, stealth interventions (e.g., having an NPC surreptitiously direct students towards resources and evidence that deserve further consideration) could be applied before a student completes the scenario. The autonomy provided by VPA makes certain aspects of such stealth interventions difficult, since one cannot be certain that a student who has not yet explored necessary evidence will not do so eventually. However, it is less problematic to intervene after specific actions that are negatively correlated with desirable outcomes (not just in the model, but individually). A student who is spending excessive amounts of time exploring the space alien information page, for example, could be prompted by an NPC character to view other pages or asked to talk through evidence supporting the space alien hypothesis. While it is true that the relationship found above is for total use of this information page across the scenario, the behavior itself is meaningful, and can be responded to. Each amount of time spent on that activity increases the probability of a negative outcome. Therefore by responding to behaviors such as that one when it occurs, it becomes possible to use these models—which are fairly coarse-grained—for adaptive personalization during learning.

Alternatively, students who complete the scenario unsuccessfully could be redirected based on the findings of these models. That is, by determining features most contributed to either an incorrect CFC or a DCE below a certain threshold, we could design interventions that offer students better opportunities to explore and understand the evidence they previously missed or misinterpreted. By comparing students' performance when allowed to navigate independently (in the first attempt) and their performance when prompted by targeted interventions (in the second attempt), educators and researchers might obtain more Vygotskian (1978) assessment of the student's zone of proximal development for inquiry.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

In conclusion, recent frameworks have placed emphasis on students engaging in the practices of science inquiry. However, developing assessments of science inquiry practices requires new approaches for modeling learning, particularly when students are given considerable autonomy, as they are in VPA. The use of LA/EDM inference models is key to this development, as these methods have enabled researchers to model increasingly ill-defined constructs and behaviors in increasingly open and authentic learning environments. In this research we have shown how we can develop reliable and valid assessments of inquiry learning which can be generalized between scenarios with relatively minimal effort. Such research will help us not only to better understand how students who are successful at science inquiry go about that practice, it should advance efforts to make adaptive personalization available to more students within a wider range of learning situations

Acknowledgments

The research presented here was supported by the Bill and Melinda Gates Foundation. We also thank Chris Dede, Michael Sao Pedro, and Yang Jiang for helpful support and suggestions.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

References

Au, W. 2007. High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Ed. Res.*, 36: 258-267.

Azevedo, R. (2005). Computer environments as metacognitive tools for enhancing learning. *Educational Psychologist*, 40(4), 193–197.

Baker, R.S.J.d., Gowda, S., Corbett, A.T. (2011) Towards predicting future transfer of learning. *Proceedings of 15th International Conference on Artificial Intelligence in Education*, 23-30.

Baker, R., Siemens, G. (in press) Educational data mining and learning analytics. To appear in Sawyer, K. (Ed.) *Cambridge Handbook of the Learning Sciences: 2nd Edition*.

Bannert, M. (2000). The effects of training wheels and self-learning materials in software training. *Journal of Computer Assisted Learning*, *16*(4), 336-346.

Black, P., & Wiliam, D. 1998. Assessment and classroom learning. *Assessment in Education*, 51:7-74.

Chen, K.-C., & Jang, S.-J. (2010) Motivation in online learning: Testing a model of self-determination theory. *Computers in Human Behavior, 26*(4), 741-752.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Chen, K.-C., Jang, S.-J., & Branch, R. M. (2010) Autonomy, Affiliation, and Ability: Relative Salience of Factors that Influence Online Learner Motivation and Learning Outcomes. *Knowledge Management & E-Learning: An International Journal (KM&EL), 2*(1), 30-50.

Chen, Z., Klahr, D. (1999) All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy. *Child Development, 70* (5), 1098-1120.

Clarke, J., & Dede, C. (2009) Design for Scalability: A Case Study of the River City Curriculum. *Journal of Science Education and Technology*, 18 (4), 353-365.

Clarke-Midura, J., Dede, C., & Norton, J. (2011) Next generation assessments for measuring complex learning in science. In D. Plank, J. Norton, C. Arraez, & I. Washington (Eds.), *The road ahead for state assessments.* (pp. 27-40). Cambridge, MA: Rennie Center for Education Research & Policy.

Clarke-Midura, J., McCall, M. & Dede, C. (2012). *Designing Virtual Performance Assessments*. Paper presented at AAAS. Vancouver, Canada, February 18.

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.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

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

Cohen, W. (1995) Fast Effective Rule Induction. *Proceedings of the Twelfth International Conference on Machine Learning.*

Conlan, O., Hampson, C., Koidl, K., Gobel., S. & Mehn, F. (2012) In Kickmeier-Rust, M. D., & Albert, D. (Eds.) *An Alien's Guide to Multi-Adaptive Educational Computer Games*. Informing Science.

Dalgarno, B., & Lee, M. J. (2010) What are the learning affordances of 3-D virtual environments? *British Journal of Educational Technology*, *41*(1), 10-32.

Dede, C. (2006) Evolving innovations beyond ideal settings to challenging contexts of practice. *The Cambridge handbook of: The learning sciences*, 551-566.

Dede, C., Honan, J., & Peters, L., Eds. (2005) *Scaling Up Success: Lessons Learned from Technology-Based Educational Innovation*. New York: Jossey-Bass.

Gee, J. P. (2007). *What Video Games Have to Teach Us About Learning and Literacy*. New York, NY: Palgrave Macmillan.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Hanley, J.A. and McNeil, B.J. (1982) The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology, 143*. 29-36.

Kester, L., Kirschner, P. A., van Merriënboer, J. J., & Baumer, A. (2001) Just-in-time information presentation and the acquisition of complex cognitive skills. *Computers in human behavior*, *17*(4), 373-391.

Kickmeier-Rust, M. D., & Albert, D. (2010). Micro-adaptivity: protecting immersion in didactically adaptive digital educational games. *Journal of Computer Assisted Learning*, *26*(2), 95-105.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006) Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational psychologist*, *41*(2), 75-86.

Kreijns, K., & Kirschner, P. A. (2001) The social affordances of computer-supported collaborative learning environments. In *Frontiers in Education Conference, 2001. 31st Annual* (Vol. 1, pp. T1F-12). IEEE.

Kuhn, D., Black, J., Keselman, A., Kaplan, D. (2000). The Development of Cognitive Skills to Support Inquiry Learning. *Cognition & Instruction*, 18(4), 495-523.

Lajoie, S. P. (1993). Computer environments as cognitive tools for enhancing learning.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Computers as cognitive tools, 261-288.

Landis, J.R., Koch, G.G. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33,* 159-174.

Leighton, J. P., & Gierl, M. J. Eds.. 2007. *Cognitive Diagnostic Assessment for Education: Theory and Applications.* New York, NY: Cambridge University Press.

Lynch, C., Ashley, K., Aleven, V., Pinkwart, N. (2009) Concepts, Structures, and Goals: Redefining Ill-Definedness. *International Journal of Artificial Intelligence in Education, 19,* 253-266.

Martens, R., Gulikers, J., & Bastiaens, T. (2004) The impact of intrinsic motivation on e-learning in authentic computer tasks. *Journal of Computer Assisted Learning, 20,* 368-376.

McCall, M., & Clarke-Midura, J. (2013). *Analysis of gaming for assessment*. Paper presented at the Association of Test Publishers Annual Meeting, Orlando, Florida.

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

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multimedia learning. *Instructional design for multimedia learning*, 181-195.

National Research Council. (2011) *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Academies Press.

National Research Council. 2006. *Systems for state science assessment*. Washington, DC: The National Academies Press.

Pellegrino, J. W., Chudowski, N., & Glaser, R. 2001. *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Quellmalz, E.S., Timms, M.J., Silberglitt, M.D., Buckley, B.C. (2012) Science Assessments for All: Integrating Science Simulations Into Balanced State Science Assessment Systems. *Journal of Research in Science Teaching*, *49* (3), 363-393.

Quinlan, J.R. (1993) C4.5: Programs for Machine Learning. San Francisco, CA: Morgan Kaufmann.

Rowe, J., & Lester, J. (2010) Modeling User Knowledge with Dynamic Bayesian Networks in Interactive Narrative Environments. *Proceedings of the 6th International Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-10),* 57-62.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Rowe, J., Mott, B., McQuiggan, S., Robison, J., Lee, S., Lester, J. (2009) Crystal Island: A Narrative-Centered Learning Environment for Eighth Grade Microbiology. *Proceedings of the AIED'09 Workshop on Intelligent Educational Games,* 11-20.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary educational psychology, 25*(1), 54-67.

Sabourin, J., Rowe, J., Mott, B. W., & Lester, J. C. (2012, January). Exploring inquiry-based problem-solving strategies in game-based learning environments. In *Intelligent Tutoring Systems* (pp. 470-475). Berlin, Germany: Springer.

San Pedro, M.O.Z., Baker, R.S.J.d., Bowers, A.J., Heffernan, N.T. (2013) Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. *Proceedings of the 6th International Conference on Educational Data Mining,* 177-184.

Sao Pedro, M., Baker, R., Gobert, J., Montalvo, O., Nakama, A. (2013) Leveraging Machine-Learned Detectors of Systematic Inquiry Behavior to Estimate and Predict Transfer of Inquiry Skill. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research, 23* (1), 1-39. (2013)

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Sao Pedro, M. A., Baker, R.S.J.d., Montalvo, O., Nakama, A., Gobert, J.D. (2010) Using Text Replay Tagging to Produce Detectors of Systematic Experimentation Behavior Patterns. *Proceedings of the 3rd International Conference on Educational Data Mining,* 181-190.

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., Gobert, J., Baker, R.S.J.d. (2012) The Development and Transfer of Data Collection Inquiry Skills across Physical Science Microworlds. Paper presented at the *American Educational Research Association Conference*.

Scalise, K. & Clarke-Midura, J. (2014). mIRT-bayes as Hybrid Measurement Model for Technology-Enhanced Assessments. Paper presented at the National Council for Measurement in Education Conference. Philadelphia, PA.

Schoor, C., & Bannert, M. (2012). Exploring regulatory processes during a computer-supported collaborative learning task using process mining. *Computers in Human Behavior*, *28*(4), 1321-1331.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer games and instruction*, *55*(2), 503-524.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Sil, A., Shelton, A., Ketelhut, D.J., Yates, A. (2012) Automatic Grading of Scientific Inquiry. *Proceedings of the NAACL-HLT 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7).* Montreal, Quebec, Canada.

Tschirgi, J.E. (1990) Sensible Reasoning: A Hypothesis about Hypotheses. *Child Development, 51* (1), 1-10.

Unity Technologies. Unity Game Engine. (2010)

Van Joolingen, W. R., De Jong, T., & Dimitrakopoulou, A. (2007). Issues in computer supported inquiry learning in science. *Journal of Computer Assisted Learning*, *23*(2), 111-119.

Veeramachaneni, K., O'Reilly, U.M., Taylor, C. (2014) Towards Feature Engineering at Scale for Data from Massive Open Online Courses. arXiv pre-print #1407.5238.

Vygotsky, L. (1978) *Mind in Society*. Cambridge, MA: Harvard University Press.

Wang, Y., Witten, I.H. (1997) Induction of Model Trees for Predicting Continuous Classes. *Proceedings of the European Conference on Machine Learning.*

Witten, I.H., Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques.* Burlington, MA: Morgan Kauffman.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Webb, M., Gibson, D., & Forkosh-Baruch, A. (2013) Challenges for information technology supporting educational assessment. *Journal of Computer Assisted Learning, 29*(5), 451-462.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

FIGURES

Fig. 1. Screen shots of a Virtual Performance Assessment (VPA)

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

TABLES

Table 1. VPA surface feature trait correspondences

Frog Scenario	**Bee Scenario**
Frog	Bee
Tadpole	Larvae
6-legged frog	Dead bee
Blood test	Protein test
Water	Nectar
Water test	Nectar test
DNA test	DNA test
Parasites (correct answer)	Genetic mutation (correct answer)

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Table 2. VPA Product Data: Final Questions.

Final Question	Type of Question
What is your claim for what caused the (6 legged frog, disappearance of bees)?	Multiple choice. (The sole question used to calculate CFC.)
Which farm do you think is the source of the problem?	Multiple choice.
Select the evidence from your back pack that support your claim.	Students click on items in their backpack (e.g. water samples, test results, etc). Students can click on one item, several items, or none.
Is there any evidence from the morphology of the tadpoles/larvae that support your claim? If this data is not evidence, then select "none of the results support my claim."	Multiple choice. Students are provided with images and morphology of 4 different tadpoles/larvae and have to select one or "none."
Is there any evidence from the morphology of the frogs/bees that support your claim? If this data is not evidence, then select "none of the results support my claim."	Multiple choice. Students are provided with images and morphology of 4 different frogs/bees and have to select one or "none."
Select the data from the water/nectar lab results that support your claim. If this data is not evidence, then select "none of the results support my claim."	Multiple choice. Students are provided with water/nectar lab results and have to select one or "none."
Select the data from the DNA test results that support your claim. If this data is not evidence, then select "none of the results support my claim."	Multiple choice. Students are provided with DNA lab results and have to select one or "none."
Select the data from the blood/protein lab results that support your claim. If this data is not evidence, then select "none of the results support my claim."	Multiple choice. Students are provided with blood/protein lab results and have to select one or "none."

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Table 3: Features used in final models predicting Correct Final Claim (CFC)

Training Data	Features Included in Model
Frog Scenario	1. IF the student spent at least 66 seconds reading the parasite information page, THEN the student will obtain the correct final conclusion (confidence = 81.5%) 2. IF the student spent at least 12 seconds reading the parasite information page AND the student read the parasite information page at least twice AND the student spent no more than 51 seconds reading the pesticides information page, THEN the student will obtain the correct final conclusion (confidence = 75.0%) 3. IF the student spent at least 44 seconds reading the parasite information page AND the student spent under 56 seconds reading the pollution information page, THEN the student will obtain the correct final conclusion (confidence = 68.8%) 4. OTHERWISE the student will not obtain the correct final conclusion (confidence = 89.0%)
Bee Scenario	1. IF the student read the genetic mutation information page at least 3 times, AND the student read the pesticides information page no more than 3 times, AND the student read the pollution information page no more than twice, AND the student read the parasites information page no more than once, THEN the student will obtain the correct final conclusion (confidence = 96.1%) 2. IF the student spent at least 29 seconds reading the genetic mutation information page, AND the student read the genetic mutation information page at least 5 times, AND the student read the pollution information page no more than 6 times, AND the student spent no more than 54 seconds reading the parasites information page, THEN the student will obtain the correct final conclusion (confidence = 89.9%) 3. IF the student spent at least 63 seconds reading the genetic mutation information page, AND the student spent no more than 47 seconds reading the parasites information page, AND the student spent no more than 107 seconds reading the pesticides information page, AND the backpack (without repeats) had at least 11 objects at one point, THEN the student will obtain the correct final conclusion (confidence = 83.9%) 4. IF the student spent at least 17 seconds reading the genetic mutation information page, AND the student spent no more than 23 seconds reading the pollution information page, AND the student did not read the parasites information page, THEN the student will obtain the correct final conclusion (confidence = 81.5%) 5. IF the student spent at least 30 seconds reading the genetic mutation information page, AND the student viewed the genetic mutation page at least 8 times, AND the student viewed the pesticides page no more than 9 time, AND the student spent no more than 94 seconds reading the pollution information page, THEN the student will obtain the correct final conclusion (confidence = 85.7%) 6. IF the student viewed the genetic mutation page at least twice, AND the student viewed the pollution page no more than one, AND the student spent no more than 24 seconds reading the pesticides page, THEN the student will obtain the correct final conclusion (confidence = 73.4%) 7. OTHERWISE, the student will not obtain the correct final conclusion (confidence = 85.0%)

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Table 4: Features used in final DCE models. These models predict a DCE score (scaled from 0 to 1). The literal equations are shown here, e.g. multiply -.165 to the maximum number of items (including repeats), add 0.322 multiplied by the maximum number of item (not including repeats), and so on for the rest of the features, add 9.153, and finally divide by 24 to obtain the predicted DCE score for the Frog Scenario.

Model trained on Frog Scenario		**Model trained on Bee Scenario**	
(- 0.165	* Maximum number of items (including repeats) in backpack.	(+ 0.474	* Maximum number of items (including repeats) in backpack, rescaled.
+ 0.322	* Maximum number of items (not including repeats) in backpack.	- 0.258	* Maximum number of items (not including repeats) in backpack, rescaled.
- 0.656	* Average number of items (including repeats) in backpack.	- 1.291	* Average number of items (including repeats) in backpack, rescaled.
+ 0.651	* Average number of items (not including repeats) in backpack.	+ 0.996	* Average number of items (not including repeats) in backpack, rescaled.
+ 3.483	* Maximum degree of coverage for a lab test	+ 1.170	* Maximum degree of coverage for a lab test
- 5.120	* Percentage of time student spent at farms	- 3.355	* Percentage of time student spent at farms
- 0.644	* Ratio between trips to lab and trips to farms (lab trips divided by farm trips)	- 2.878	* Percentage of time student spent in lab
+ 0.542	* Did the student ever run a blood test on the six-legged frog?	- 0.457	* Did the student ever run a genetic test on a dead bee?
+ 0.714	* Did the student ever run a blood test on a non-sick frog?	+ 0.972	* Did the student ever run a protein test on the dead bee?
- 0.657	* Did the student ever run a genetic test on a non-sick frog?	+ 1.082	* Did the student ever run a protein test on a live bee?
- 0.834	* Did the student ever run a water test on farm water?	- 1.544	* Did the student ever run a nectar test on farm nectar?
- 1.137	* Did the student ever run a water test on lab water?	- 0.534	* Did the student ever run a nectar test on lab nectar?
+ 0.044	* How long, on average, did students spend reading information pages? (average per read)	- 0.005	* Standard deviation of time spent reading information pages (per read)
+ 0.009	* How long, in total, did student spend reading information page on parasites (the correct hypothesis)?	+ 0.030	* How long, in total, did student spend reading information page on genetic mutations (the correct hypothesis)?
+ 0.004	* How long, in total, did student spend reading information page on pollution?	- 0.009	* How long, in total, did student spend reading information page on space aliens?
+ 0.799	* Total number of times student accessed information page on parasites (the correct hypothesis)	+ 0.709	* Total number of times student accessed information page on genetic mutations (the correct hypothesis)
- 0.025	* Standard deviation of time spent reading information pages (per read)	- 0.243	* Total number of times student accessed information page on space aliens
- 0.563	* Total number of times student accessed information page on space aliens	- 0.361	* Total number of times student accessed information page on viruses
- 0.197	* Number of different (types of) non-sick frogs student took to the lab at the same time	- 0.279	* Total number of times student accessed information page on pollution

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

+ 1.372 * Did the student take lab water to the lab?	
+ 9.153)	+ 9.316)
/ 24.0	/ 24.0

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)