Biology Microworld to Assess Students' Content Knowledge and Inquiry Skills and Leveraging Student Modeling to Prescribe Design Features for Scaffolding Learning

By
Matthew Knapp Bachmann
A Thesis
Submitted to the Faculty
of the
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the
Degree of Master of Science
in
Computer Science
May 2012

APPROVED:
Janice Gobert, Advisor

____

Joseph Beck, Co-Advisor

____

Craig Wills, Head of Department

____

Abstract

It is the underlying presupposition of the Science Assistments research (http://www.scienceassistments.org) that students need to leave school with a basic understanding of science and grounding in inquiry skills (NSES, 1996; NRC, 2011). We also believe that the current standard for assessing these skills, the Massachusetts Comprehensive Assessment System, is inadequate in terms of the rote-oriented multiple-choice tests.

This thesis describes the creation of a simulation, or microworld, of an animal cell. This content is aligned with the Massachusetts science frameworks for middle school Life Science (Massachusetts Department of Education, 2006). Our microworld, Simcell, gives students an opportunity to form hypotheses, design experiments to test these hypotheses, and analyze their data collected during the experiment. The microworlds track students' actions in log files that can be analyzed by the system to provide fine tuned assessments of students, and based on these assessments, in the future, we will provide dynamic help though scaffolds to students who are struggling with inquiry (Gobert et al, 2007; 2009; Gobert et al, in press).

Over the course of two studies, this biology microworld was designed, developed, and fined tuned through the use of domain experts and student pilot data. We also analyzed the student logs in order to try to model students' learning so we can predict useful times for the system to come in and help. In study one we identify a potential point to remediate struggling students. In study two we conducted a series of logistic and linear regressions to predict student knowledge. However, due to the large number of different variables and the relatively small size of the dataset, we could not be confident in the results that were obtained. Many attempts to reduce the number of variables used in the model were tried, but these methods did not yield more promise than the original set.

Finally, we finish this report with a new path for researchers to consider, namely, looking at the data in different ways in order to find a way of viewing the data that would allow for known successful student modeling techniques such as Bayesian Knowledge Tracing.

Acknowledgements

First I would like to thank my advisors Janice Gobert and Joe Beck for guiding me through my time at WPI, with a special note to Janice who went above and beyond in the final steps. Thanks to Dave Brown for guiding my work in its early stages and providing helpful advice to the microworld itself. Additionally, thank you to Kathi Fisler for her insightful comments on the full draft. Thanks to the other WPI professors for giving me the tools I will be using though out my life. Thanks to the Department of Education and the National Science Foundation for funding me and many of my fellow students though graduate studies and for making this and other projects possible. Thanks to my fellow classmates for their support as things got difficult and for giving me some of the best times of my life. Finally, I want to thank my parents without whom none of this would be possible.

Table Of Contents

List Of Figures And Tables

Introduction

According to the US Department of Labor, the growth areas of our economy are ones that come from the STEM disciplines: Science, Technology, Engineering, and Math (US Department of Labor, 2007). However, student test scores in these areas show that we are not giving students the skills they need to succeed in these disciplines (National Center For Education Statistics, 2009). When students leave school, they will enter a world where being knowledgeable about scientific topics is critical in order to form informed opinions on world issues such as the energy crisis and the environmental impacts we have on the world. In addition, students need to have honed inquiry skills so they can reason critically in order to make scientific decisions about these issues and in their own lives (NRC, 9116; 2011).

The National Science Education Standards (NSES, 1996), the U.S. Dept. of Education (1993), and the National Assessment of Educational Progress (NAEP; Champagne, A., K. Bergin, R. Bybee, R. Duschl, and J. Gallagher 2004) recognize inquiry as an important aspect of scientific literacy and concur on a definition that includes these components: 1) identifying questions that guide investigations; 2) designing and conducting investigations; 3) using technology to improve investigations; 4) formulating and revising explanations and models using logic and evidence; 5) recognizing and analyzing alternative explanations; and 6) communicating and defending a scientific argument. The needs are clearly defined for us, yet as a country we seem to fall behind other countries on international comparisons of science. Specifically, according to the Program of International Assessments' 2006 study, 15-year-old students from the United States performed below average in science literacy tests when compared to fifty-seven other countries (Snyder, 2009). We assess student knowledge in state and national standardized tests such as the MCAS, despite the National Research Council's *National Education Standards* call for science to be taught as a way of thinking rather than as an index of facts (NSES, 1996). Clearly, work needs to be done in order to achieve the goals set out for our students.

In the Science Assistments Project (Gobert et al 2007, 2009, 2011) http://www.scienceassistment.org/), we seek to help students acquire critical science process skills by engaging them in inquiry-based learning with *microworlds*. Microworlds are computer simulations that students can work with to perform virtual science inquiry in a way that traditionally required a full lab setup for each student. These microworlds allow each student to perform experiments and work with systems for a fraction of the cost and time that it would take to set this up using real world materials. Think of how phase change is demonstrated with a beaker, a heat source, and a thermometer. This setup must be put together in advance for each student or small group of students. Then the teacher needs to make sure students use the materials in a safe manner. Use of microworlds can simplify this process for teachers so they can spend less time preparing the lab and more time preparing and aiding with the lesson.

In addition these microworlds allow students do things that were simply not possible in a high school science lab-- like observing the effects of dropping objects in environments with different forces of gravity, or working with animal cells by building their own.

How Modern Machines Can Improve Microworlds

The benefits of computer simulations have been available to teachers for decades (Howel, Scott, & Diamond, 1987). What makes the Science Assistments project different is how we are taking steps to turn these science microworlds into a science tutoring system. We want these microworlds to be able to assist students who are struggling with inquiry processes and to challenge students who are able to easily complete it. The way we are able to do this is by taking advantage of the power of modern computers and the Internet to gather large amounts of data. We then can use this data to build models that can be used to assess the student's progress. With these models we can provide both detailed reports on students to help teachers focus on skills/areas needed to help the student, and adaptive scaffolds to help students directly. The microworlds in the Science Assistments project collects detailed logs of student activity in such detail that the student's work can be replayed using them. Using similar logs, Gobert et al. (2006; Buckley, Gobert, et al, 2010) were able to identify student behavior as either "systematic" or "haphazard." This ability to categorize students who are able to perform complicated inquiry processes demonstrates the power of using log data to capture students' inquiry skills. The Science Assistments project hopes to take this logging to the next level of scoring students' inquiry skills computationally and on the fly.

The Assistments project, upon which Science Assistments was built, has seen success by applying student modeling and real time behavioral tracking in the field of mathematics (Razzaq, Feng et. al. 2005, Feng, Heffernan, Koedinger 2006). Having a sufficient model of the student can allow the system to predict the behavior of that student and accurately assess the student's current level of proficiency. The Assistments team has successfully created a mathematics assessment system that gives teachers full reports, helps students, and can even predict the student's performance on MCAS scores (Razzaq, Feng et. al. 2005, Feng, Heffernan, Koedinger 2006). This assessment system, and other tutors in this domain, have accomplished learning gains by identifying common mistakes students make and providing step-by-step help though scaffolds that support the student in eventually learning the skills required by the question. In addition, these tutors have leveraged computational student modeling to identify when a student has acquired skills.

While the Science Assistments project leverages a lot of the work in the Assistments project we are trying to solve a fundamentally different problem. Math is a well-defined domain and problems in Assistments tend to be multiple-choice problems that are solved with a simple right or wrong and the middle school level a computer can solve most of the problems itself so grading computationally is often easy. Science, on the other hand, is a more ill-defined domain. Even at the middle school level. When assessing a student there it can be much harder to identify good or bad behavior without direct human intervention and the grade must not be a simple right or wrong. It should come to no surprise that the techniques in Assistments and previous ITS systems do not easily port over.

Problems Applying Common Techniques from ITS

The most common modeling approach is Bayesian Knowledge Tracing (BKT). BKT is a two state network. A student has either learned the skill or not. In addition BKT has four different parameters. These are: initial knowledge $P(l_0)$ or the probability that the student knew the skill before starting the problem set, learn P(t) or the probability that the student will learn the skill after an application of the skill, guess P(g) or the probability that the student will guess the problem correctly if they do not know the skill, and slip P(s) or the probability the student will make a mistake and get the problem wrong even though the know the skill (Corbett, Anderson, 1995.)

While BKT has been the core of student modeling research, recently a different approach, PFA (Performance Factor Analysis), has gained some traction in the field. PFA has been shown to have advantages over BKT, such as better estimation of student knowledge parameters (Gong, Beck, Heffernan 2010). PFA is a logistic regression model that predicts student performance with each skill. A skill consists of three parts. A value representing the difficulty of the skill and two variables that scale with the number of prior successes and failures in applying that skill (Pavlik, Cen, Koedinger 2009).

Both PFA and BKT have been successful in the field. However there is a fundamental problem in applying these models to scientific inquiry: science is a far less defined field than Math. Primary and secondary level mathematics has the advantage of easily determining if a student has gotten a problem correct or not. Both BKT and PFA models depend on easily determining if the student's answer was correct or incorrect. While the Science Assistments Project has seen some successes in computationally defining correctness, it continues to be a research question (Montalvo, Baker et al 2010).

However the largest problem in leveraging these models in our system is that a student does not get very many practice attempts at skills in the time it takes to complete a reasonable science assessment activity. BKT and PFA require a large number of practice attempts to gain any accuracy in their predictions. In our system students generally have only one to two chances to apply many of the skills we would like to scaffold such as forming hypotheses, analyzing data, etc. This fact makes models such as PFA or BKT not viable.

While these models are not necessarily appropriate, there have been attempts to model the student such that we could potentially distinguish students in a way that allows us to scaffold them. One attempt was student transition analysis (STA). STA was a basic Markov model that tried to trace a student though an entire inquiry activity. The goal with this method was to track the student's path and identify when they have gone down problematic paths (Bachmann, Gobert, Beck 2010). This model did not prove successful in predicting student performance.

However, the method when applied to an activity with greater possibility of divergence of student choices may prove to be a useful one. While a lot has been done on the topic of finding appropriate models for science environments it is clear that we have yet to find an appropriate baseline on which to build.

The work presented here is another attempt to model students in the hopes of finding a method with which to scaffold and assess them. In addition this work extends the Science Assistments project into a new domain: cellular biology. Cellular biology is a domain where students have many misconceptions, the source of which is the inability think at the microscopic level. Specifically, we try to address the problem students have with the cell being a dynamic system of organelles that work together to sustain a cell's life (Berthelsen, 1999). To learn about the cell, students must work with a simulation of the cellular environment (Flores, Tovar, & Gallegos, 2003).

There is a lot to consider when designing a new microworld. The microworld created as part of this work (Simcell) was designed to fit into the Massachusetts State Framework, which states that students must understand the functionality of the various organelles in the cell. In addition to trying to establish this content knowledge in students' minds we also seek to get students applying sound inquiry methods when working with this activity. However, on top of this we need to provide a clear, clean interface that can communicate this knowledge without causing too much cognitive load, which may overwhelm the student and hinder learning (Sweller, 1994).

This work includes two studies, the first of which works with the original Simcell microworld designed by a group of WPI undergraduates' for their Interactive Qualifying Project (IQP). I assisted Krista Baker, Sana Hashmi, and Fioledia Prifti in the design of the microworld

(Baker, Hashmi, Prifti 2010). From the beginning of their IQP, I served as a technical consultant, assisting with their design effort as well as being the sole programmer. The second study works with a revision of the microworld where changes were made based on the results and logs of the first study, in addition to the critical feedback from students working with the system, educators, and a professor knowledgeable in the field of HCI. This paper will discuss the design of the microworld and the attempts to develop student models to scaffold students who use the activity in the future.

Development Of Concept

The goal for this microworld was to create an inquiry activity aligned with the Massachusetts curricular frameworks to address the function and structures of organelles within the life sciences strand (Massachusetts Department of Education, 2006). One of the goals of the Science Assistments project is to assess student knowledge *and* leave students knowing more about the domain after completion of the activity. We decided that a "hands-on" portion for the activity would allow the student to directly work with the cell itself and could engage them on a deeper level since research has shown that allowing students to interact with a simulation of a cellular environment gives students a deeper understanding of that cell (Flores, Tovar, & Gallegos, 2003). Additionally, since we collect and analyze students' log files, we are able to use these as fine-grained performance assessments of inquiry in this domain. One small feature, which we think was a huge leap of making Simcell feel hands-on was the usage of a drag and drop user interface. The microworld is described in detail later in the document.

Technology

The Science Assistments project utilizes a whole suite of technologies to create the activities that make it into schools. Each is described in turn.

Assistments

Assistments is a web based software system which provides students with instruction and structured practice for mathematics. In addition, it provides teachers with detailed reports on students' performance and acquisition of skills (Feng et al 2010). Science Assistments uses this framework to embed our microworlds; we also utilize this system to assist us in the logging of student actions. Assistments is written in a combination of Ruby on Rails, JavaScript, HTML, and CSS.

OpenLaszlo

OpenLaszlo is an open source language that uses an XML like structure combined with JavaScript to make dynamic websites. This is the primary language of the Science Assistments project. The most useful aspect of this language is that it is a free language that compiles into an Adobe Flash SWF file, which we can easily embed into a web-based lesson created in the Assistment system.

Science Assistment Core Architecture

The Science Assistment core architecture is the back-end for any microworld made for the project. It provides the frame for the student's inquiry steps as well as handling any communication between the microworlds and the Assistment system. It is written in OpenLaszlo and was created by Mike Sao Pedro, Andy Montalvo, and Ermal Toto.

Development Of Simcell

Simcell was the result of two years of iterative development. The original design of Simcell One was a collaborative effort between an IQP team and myself, as previously mentioned. No one on the IQP team had any programming knowledge so I was the sole developer of the microworld. Working with the tools provided to me from the other members of the science assistment team, I created the microworld that would eventually appear in middle schools.

Upon the completion of the study one I also coded the tools that the IQP team used to analyze the log files. The microworld was written in OpenLaszlo. This language allowed us to embed the microworlds easily into assessments and provided tools for the easy creation of graphical user interfaces. The tools to parse the logs were written in Java.

When all the analyses for Simcell One were completed, the IQP team had finished their work and left the project. From this point I was the sole designer of the system. I decided to rewrite the microworld from scratch as I had gained a greater understanding of OpenLaszlo and wanted to try to write the microworld in a way that would allow future students to expand upon the microworld more easily. I designed the code to take advantage of OpenLaszlo’s data tools, which would allow me to specify more of the microworld in XML so that changes in configuration could be made simply by changing the XML file rather than having to work with the code.

More than just changing the code, I decided to change up the microworld as well. I talked with my advisors, fellow students, developers, teachers, students, and an HCI expert to gather ideas and feedback in the design and implementation of every stage of the inquiry cycle: Observing, Hypothesizing, Experimenting, and Analyzing. This new Simcell served as the basis for study two and was also used as the base for other Biology microworlds.

Finally I expanded on the logging for the microworld to provide logs in a format that could more easily be parsed by future researchers. Using these new logs, I wrote tools to analyze data given by the new microworld to help make sense of the data gathered as students worked with the system.

Integrating Simcell With Assistment

In order for the software and us to properly analyze student actions we need to capture them in a useful format. The way we do this is with a logging infrastructure that hooks into the Assistments logging infrastructure. When a microworld starts up a logger is created and registered. Whenever the user does an action in our microworld we take this action parse it into a usable format. Then we pass this information to Assistments, which treats it as any other assistment log. The action ends up in a database that we access at our end. From there we are able to utilize these logs for offline analyses.

Study One

The purpose of this study is to give the Simcell microworlds its first usage in the field in order to collect data to help fine tune the microworld itself and assess its ability to convey the material to real world students. We also expected to use the data collected in this study to find methods to scaffold students working through the activity.

Method

Participants

Participants were 85 middle school students from a large urban school in Massachusetts. Analyses were performed on data from 71 students who completed the activity in full. Over 50% of the students from this school are on free or reduced price lunch programs. Less than 50% of the students are Caucasian, and at the time of the study 92% of students achieved below "proficient" on their MCAS science performance test.

Materials

Science Assistments

Science assistments is a web based tutoring system that includes tools for assessing and honing students' inquiry skills. It includes a logging infrastructure that captures all students' actions within the learning environment (Gobert et al, 2007; 2009; in press).

Inquiry Pre and Posttest

The inquiry pre and posttests consists of 27 multiple-choice questions and one open response question (see appendix). The goal of this test is to get an assessment of a student's inquiry skills: such as designing experiments, forming hypotheses, analyzing data, and communicating findings.

Domain Pre and Posttest

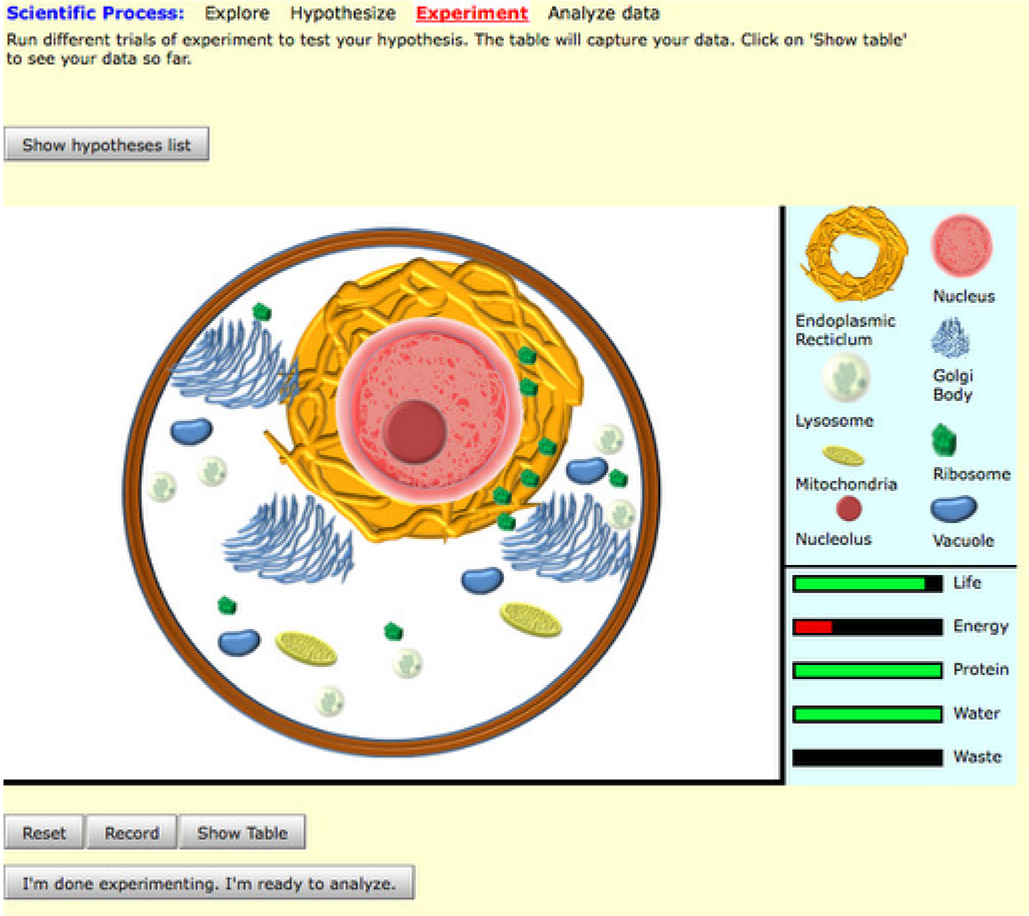
The domain pre and posttests consist of 16 multiple choice questions designed to measure students' knowledge of cellular biology related to cell structure and function (see appendix A). The goal of this test was to gather a students' overall knowledge of cells. It included questions on animal cell, plant cell, and more general cell biology content. Simcell itself only targeted animal cells. The more general pre/post test served to capture the student's broader knowledge of the topic.

Simcell Microworld

The Simcell microworld presents a cell membrane to students and allows them to change the cell by dragging and dropping cell organelles into or out of it (figure 1). The microworld gives students instant feedback on their actions through *gauges*. The gauges tell the student the current level of waste, water, energy, protein, and life of the cell. By hitting a record button, students are able to record onto a data table the organelles (and their respective amounts) that are currently in the cell and the corresponding gauge levels.

Problem Scenarios

We created two problem scenarios for this activity. *Scenario one* presents students with a cell that is lacking mitochondria, which is shown by a low energy gauge (figure 1). *Scenario two* presents students with a cell that is low in vacuoles and lysosomes, which is shown by low protein and water gauges.



Run different trials of experiment to test your hypothesis. The table will capture your data. Click on 'Show table' to see your data so far.

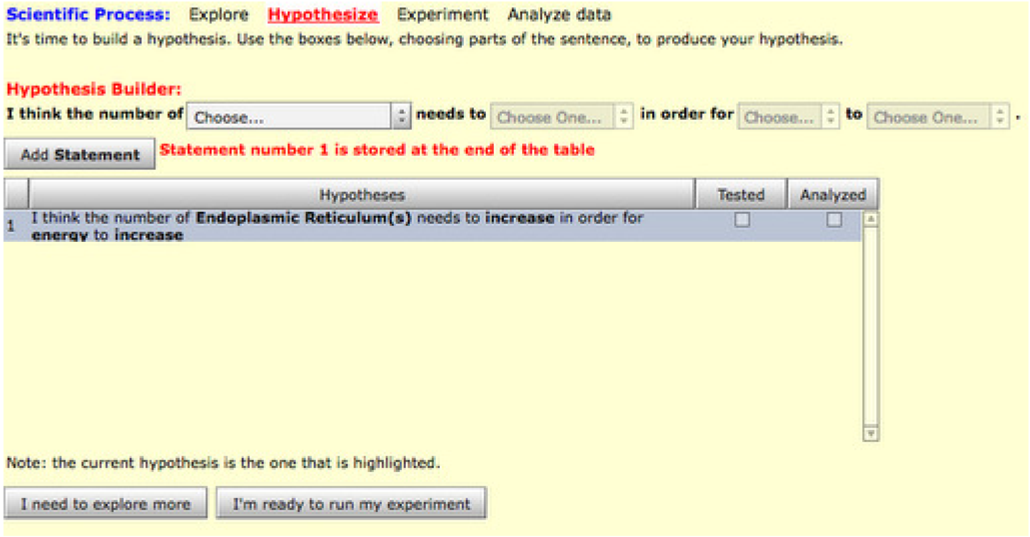
Figure 1: The Simcell 1 microworld as configured for scenario one. The energy gauge is low due to a lack of energy.

The system guides students through a scenario by first having them diagnose the problem with an open response text box. Next they test their hypothesis in the microworld using the hypothesis widget, and finally they use the data analysis widget.

Hypothesis Widget: the method students used to form hypotheses

The hypothesis widget (Figure 2) presents students with a series of drop down menus and asks them to form a hypothesis in the following form: “I think the <Independent Variable (An organelle)> needs to <DIRECTION> in order for <Dependent Variable (A gauge)> to

<DIRECTION>”. An example of a hypothesis that would be consistent with the domain and relevant to the problem presented would be: “I think the number of mitochondria needs to increase in order for energy to increase”.



It's time to build a hypothesis. Use the boxes below, choosing parts of the sentence, to produce your hypothesis.

Hypothesis Builder:

I think the number of Choose... **needs to** Choose One... **in order for** Choose... **to** Choose One... .

Statement number 1 is stored at the end of the table

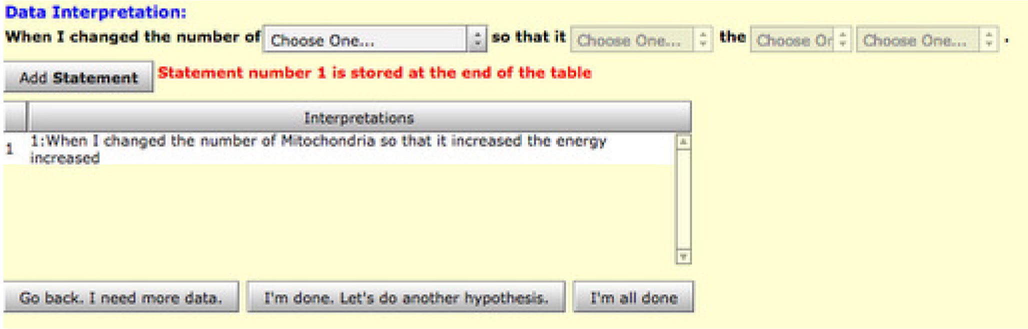
	Hypotheses	Tested	Analyzed
1	I think the number of **Endoplasmic Reticulum(s)** needs to **increase** in order for **energy** to **increase**	□	□

Note: the current hypothesis is the one that is highlighted.

Figure 2: **The Simcell 1 Hypothesis Widget.**

Data Analysis Widget: the method students use to analyze their data

The data analysis widget (Figure 3) shows students the data table they created while working with the microworld and asks the students to make some analyses. Analyses are created using a series of drop down menus in the following form: “When I changed the number of <Independent Variable (An organelle) > so that it <DIRECTION> the <Dependent Variable (a gauge)> <DIRECTION>”. An example of an analysis that would be considered consistent with the domain and relevant to the problem presented would be: “When I changed the number of Mitochondria so that it increased, the energy increased”.


Data Interpretation:

When I changed the number of Choose One... so that it Choose One... the Choose Or Choose One... .

Statement number 1 is stored at the end of the table

	Interpretations
1	1:When I changed the number of Mitochondria so that it increased the energy increased

Add Statement
Go back. I need more data.
I'm done. Let's do another hypothesis.
I'm all done

Figure 3: The Simcell 1 Data Analysis Widget

Procedure

The activity was designed to run in one day. We allowed a second day for students who missed the first day or did not finish. Assistments first presented students the inquiry pretest followed by the domain pretest. Then students were presented a blank cell in the Simcell microworld. This served as an orienting task for students to learn the interface; students were told to explore the interface and observe the effects of different organelles on the cell. Next, students completed the two problem scenarios described above. Finally, students completed the domain posttest followed by the inquiry posttest.

Data Coding

Pre and Posttests

Inquiry pre and posttests were autoscored by the Assistments system (Razzaq et al, 2005).

Students' logs

We categorized students based on their actions as recorded in the log files. Categories are defined by the state of the cell and what organelles the student used. We created four categories:

- *Excellent*: students who built a cell in which all health gauges were at optimal levels. All the gauges on the UI were full except for waste, which was empty.
- *Good*: students who built a cell in which water, energy, protein, and life gauges were all greater than 70% and waste was less than 40%. All the gauges were green on the UI.
- *Fair:* Students who used every organelle provided to them but their cell's health was insufficient to earn a good rating.
- *Poor:* Students who did not use every organelle.

Students' Hypotheses

We autoscored student hypotheses checking for correctness and relevance. A relevant hypothesis is defined as a hypothesis in which the dependent variable is one of the health gauges that is at an unhealthy level at the start of the problem. A correct and relevant hypothesis is a relevant hypothesis that happens to be true. For example: if the cell is low on energy, a hypothesis that says, "I think the mitochondria needs to increase in order for energy to increase" or "I think the mitochondria needs to decrease in order for energy to decrease" is coded as correct and relevant. A student's number of correct and relevant hypotheses was summed over both problem scenarios and used as a variable in our statistical analyses.

Correct?	**Relevant?**	**Hypothesis**
Yes	Yes	"I think the number of mitochondria need to increase in order for energy to increase"
Yes	No	"I think the number of Ribosomes need to decrease in order for protein to decrease"
No	Yes	"I think the number of mitochondria need to increase in order for energy to decrease"
No	No	"I think the number of Ribosomes need to increase in order for energy to increase"

Table 1: Examples of various hypotheses possible in Simcell 1

Students' Data Interpretations

We autoscored data analyses for correctness and relevance. Correct and relevant data analyses are defined the same way as they were for hypotheses. For this we did not take into account whether or not students recorded data that backed up their analyses. A student's number of correct and relevant analyses is summed over both problem scenarios and used as a variable in our analyses.

Correct?	**Relevant?**	**Data Interpretation**
Yes	Yes	"When I changed the number of Mitochondria so that it increased the energy increased"
Yes	No	"When I changed the number of Ribosomes so that it decreased the protein decreased"
No	Yes	"When I changed the number of Mitochondria so that it increased the energy decreased"
No	No	"When I changed the number of Ribosomes so that it increased the energy increased"

Table 2: Examples of various data interpretations in Simcell 1

Research Questions

1. Do group differences in exploration behavior, as measured by the exploration category they were placed in based on their behavior, yield differences in inquiry and/or domain knowledge at post-test?

2. Do differences in exploration behavior, as measured by the exploration category they were placed in based on their behavior, yield a greater number of correct and relevant hypotheses and/or analyses?

3. Did students learn from the activity as measured by the difference between pre and posttest score?

Results

Do group differences in exploration behavior yield differences in inquiry and/or domain knowledge?

A repeated measures MANOVA was conducted using exploration category during the orienting task (where the goal was for students to learn the interface) as the independent variable and the domain and inquiry pre- and post-tests as repeated measures. It yielded a marginally significant within-subjects effect for time (F(3,66) = 3.112, p = .051, part. η^2 = .086). The interaction between category and time was not significant (F (3, 67) = 3.122, p = .657, part. η^2= .030). When collapsing over time, we found significantly different test scores by exploration group for domain knowledge (F(3, 67) = 3.76, p = .015, part. η^2= .14) and for inquiry knowledge (F (3, 67) = 2.98, p = .038, part. η^2= .12). Planned contrasts found that students in the excellent exploration category had significantly greater pre- and post-inquiry and domain test scores than those in the poor exploration category (**Excellent**: Domain Pre: *M* = .59, *SD* = .32, Domain Post: *M* = .66, *SD*= .27, **Poor**: Domain Pre: *M* = .33, *SD* = .18, Domain Post: *M* = .65, *SD* = .23).

Inquiry Tests	**Pretest**		**Posttest**	
Exploration Group	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
Excellent	.33	.19	.49	.21
Poor	.36	.12	.28	.12

Table 3: Inquiry test results

Domain Tests	Pretest		Posttest	
Exploration Group	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
Excellent	.33	.19	.49	.21
Poor	.36	.12	.28	.12

Table 4: Domain test results

Do differences in exploration behavior yield a greater number of correct and relevant hypotheses and/or analyses?

A MANOVA using exploration category as the independent variable and the total correct and relevant hypotheses and analyses in problem scenario one as the dependent variables, using domain pretest as a covariate, yielded a significant result for exploration category (F(6,71) = 3.25, p=.004, η^2 = .133). Students in different exploration categories had a significantly different number of correct and relevant hypotheses in scenario one (F(3,71) = 5.97, p=.001, η^2 = .214). However, students in different exploration categories did not have a significantly different number of correct and relevant analyses in scenario one (F(3,71) = 2.17, p=.1, η^2 = .09). Planned contrasts were conducted in order to see which groups were significantly different from each other. It was found that students in the good category had significantly different number of correct and relevant hypotheses in scenario one than did those in the fair category (p = .000). Students in the fair category had significantly different correct and relevant hypotheses and analyses in scenario one than those in the poor category (p = .020).

A MANOVA using exploration category as the independent variable and the total correct and relevant hypotheses and analyses in scenario two as the dependent variables, using domain pretest as a covariate, found a marginally significant result for exploration category (F(6,71) =

5.09, p = .051, η^2 = .091). Students in different exploration categories had a significantly different number of correct and relevant hypotheses in scenario two (F(3,71) = 4.43, p=.007, η^2 = .167). However, students in different exploration categories did not have a significantly different number of correct and relevant analyses in scenario two (F(3,71) = 2.03, p=.12, η^2 = .085).

Planned contrasts were conducted in order to see which groups were significantly different from each other. Here it was found that students in the excellent category had significantly more correct and relevant hypotheses in scenario two than students in the good category (p = .023). Students in the good category had significantly more correct and relevant hypotheses than students in the fair category (p = .008). Students in the fair category had significantly more correct and relevant hypotheses than students in the poor category (p = .027).

A MANOVA using exploration category as the independent variable and the total correct and relevant hypotheses and analyses over the activity as the dependent variables (dvs) yielded a significant result (F(6,71) = 3.57, p =.003, η^2 = .142). Students in different exploration categories had significantly more correct and relevant hypotheses over the whole activity (F(3,71) = 6.86, p =.0001, η^2 = .304). Students in different exploration categories had marginally significantly different number of correct and relevant analyses over the whole activity (F(3,71) = 2.7, p =.053, η^2 = .129). *Planned contrasts:* Students in the excellent category had significantly more correct and relevant hypotheses over the entire activity (p = .021). Students in the good category had significantly more correct and relevant hypotheses over the entire activity than students in the fair category (p = .000). Students in the fair category had significantly more correct and relevant hypotheses *and* analyses over the entire activity than those in the poor category (p = .000).

Correct And Relevant Hypotheses	Scenario One		Scenario Two	
Exploration Group	Mean	Standard Deviation	Mean	Standard Deviation
Excellent	.82	.4	1.27	.65
Good	.56	.5	.69	.78
Fair	.18	.39	.36	.58
Poor	.00	.00	.00	.00

Table 5: MANOVA results over problem scenarios for hypotheses

Correct And Relevant Hypotheses	Entire Activity	
Exploration Group	Mean	Standard Deviation
Excellent	1.73	1.27
Good	1.41	1.04
Fair	.55	.86
Poor	.00	.00

Table 6: MANOVA results over the entire activity for hypotheses

Correct And Relevant Analyses	Scenario One		Scenario Two	
Exploration Group	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
Excellent	.73	.47	1	.89
Good	.75	.44	.66	.79
Fair	.50	.51	.36	.49
Poor	.17	.41	.00	.00

Table 7: MANOVA results over problem scenarios for analyses

Correct And Relevant Analyses	Entire Activity	
Exploration Group	**Mean**	**Standard Deviation**
Excellent	1.73	1.27
Good	1.41	1.04
Fair	.86	.83
Poor	.17	.41

Table 8: MANOVA results over entire activity for analyses

Are Students Learning From The Activity?

Domain Pre and Posttests

A paired samples t-test comparing student domain pre and posttests showed that students had significantly higher domain posttest scores (t(70) = 2.13, p = .037, Cohen's d=.2).

Inquiry Pre and Posttests

A paired samples t-test comparing student domain pre and posttests showed that students did not have significantly different inquiry test scores (t(70) = -1.55, p = .126, Cohen's d = -.18).

Discussion

We designed this study with the goal of testing how Simcell would do in a real classroom, and to collect data to begin exploring methods of modeling, and in future, scaffolding students' inquiry processes.

Do group differences in exploration behavior yield differences in inquiry and/or domain knowledge?

The results we have presented suggest that students' exploration behaviors do indicate some differences in their inquiry and domain knowledge. Specifically, students in the "excellent" exploration category had higher pre and post scores for both inquiry and domain knowledge than those in the "poor" exploration category. In brief, students who built more complete cells, as shown in the logs, came into the activity with greater prior knowledge than those who built less complete cells.

These are exciting results because they suggest that the explore section of the activity can be used as a method to categorize students early on in the activity since their inquiry behavior in this early phase predicts their subsequent learning. This is an attractive feature if time is limited and one would like to be able to have students use the systems without losing time with an old fashioned multiple choice test.

Do differences in exploration behavior yield a greater number of correct and relevant hypotheses? And do differences in exploration behavior yield a greater number of correct and relevant data interpretations?

Our results indicate that students who built more complete cells went on to make more correct and relevant hypotheses than those who built less complete cells. This may suggest that students are actively learning about the system presented to them and are using information they learned to assist them in further inquiry. This also provides more support for the notion that using the explore section as a method to categorize students early in the activity is a productive one. Since students in the different exploration groups only had a marginally different number of correct and relevant analyses, it seems that for data interpretation, the exploration phase played less of a role. Specifically for scenario two there was not a significant difference in the correct number of correct and relevant analyses. This could be a reflection of the increased difficulty of scenario two, as it requires a multi-part solution. Or it could simply be sign of students being confused over the analysis activity. Overall, it seems students who made more complete cells were more likely to demonstrate higher knowledge gains compared to students who made less complete cells.

Are Students Learning From The Activity?

Our results indicate that there was at least some learning in this activity, as there was some gains in the domain tests. This was encouraging and gave us some evidence that we may be on the right track in the design of this activity. That being said, the effect size was very small and improvement on these gains is necessary.

There was no significant change in inquiry scores, which may suggest that Simcell 1 is not doing a good enough job honing students' inquiry skills. It is important to note that inquiry skills are difficult to learn and it is possible that more than one learning activity or microworld is necessary to hone students' inquiry skills. Another possibility is that the inquiry multiple-choice tests are not aligned closely with the skills that we are assessing in the activity. This possibility seems plausible since one of the main goals of the Science Assistments project is to create activities that develop richer models of student inquiry skills than do multiple-choice tests.

That being said, it is still disappointing that we did not see gains in the inquiry tests because these are the types of questions that a student is likely to encounter in the MCAS, which students need to pass in order to graduate. Although our activities may be trying to reach for a higher standard than the MCAS, there is no reason we cannot try to improve MCAS scores while doing so.

Problems in Simcell 1 And How Simcell 2 Addressed Them.

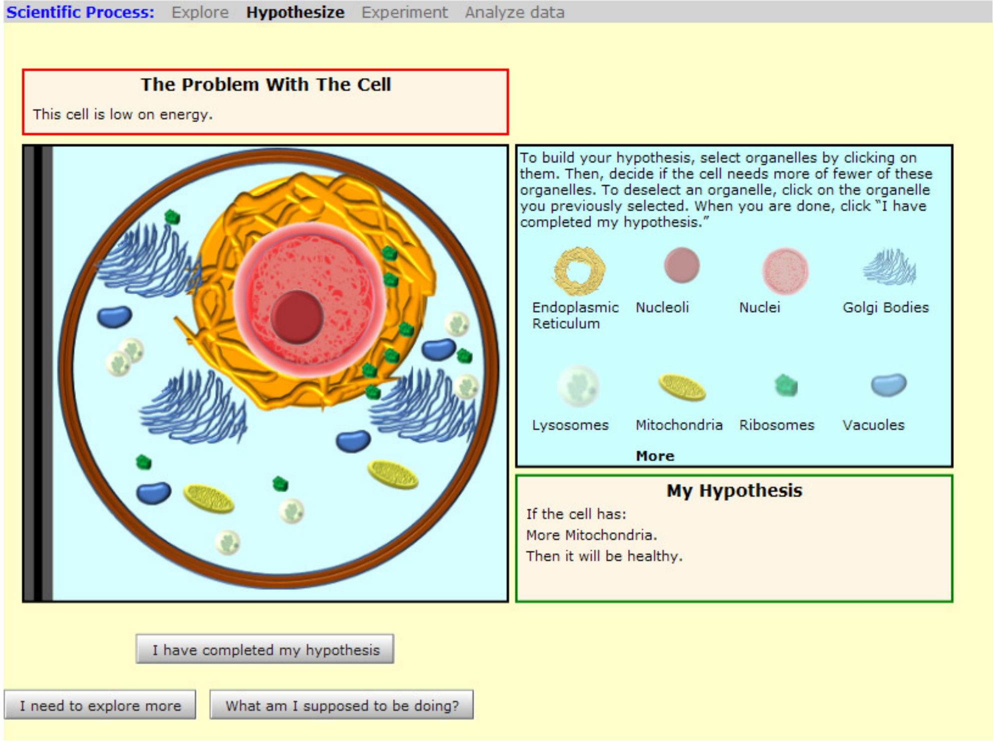
Many sources were analyzed with the goal of improving the initial microworld. We looked at the student data, gathered feedback from professors, and talked with both graduate students and middle school students. In addition, an expert in Human Computer Interaction was consulted. The above process led to many improvements to Simcell that made it a far better activity. The following section describes some of the problems of Simcell 1. For each section we describe the problems, and show their solutions in Simcell 2.

Hypothesis Widget

The hypothesis widget used in Simcell 1 (see figure 2) was taken from the physical sciences microworlds developed by our group. While it served the purposes for the activity, it did not do so optimally. One problem is that it took too many steps to make a hypothesis. It requires four different drop down boxes. Based on student response data, it was clear that many students would just click through these drop downs quickly without being mindful of what they were doing.

In addition, the hypotheses widget did not provide the expressive power to students required to fully hypothesize all of the problems of the cell. For example: scenario two required adding two separate organelles to fix the problem and form a complete cell. In order to properly communicate this in Simcell 1, a student would have to make two separate hypotheses. Very few students even bothered to try and make two hypotheses. Seeing this problem we designed the hypothesis phase to allow students to make one hypothesis covering all the possible interventions to the cell.

The hypothesis phase of Simcell 2 presents (Figure 6) students with an animal cell, which is unhealthy in that it has one or more problems and asks students to form a hypothesis about what needs to be done to make the cell healthy. Hypotheses are formed by clicking on the organelles and then deciding if the cell needs more or fewer of the selected organelles. From this phase, students can either go to the explore phase if they do not think they can make an informed hypothesis or they can go to experiment once they are ready to test a hypothesis.



The Problem With The Cell

This cell is low on energy.

To build your hypothesis, select organelles by clicking on them. Then, decide if the cell needs more of fewer of these organelles. To deselect an organelle, click on the organelle you previously selected. When you are done, click "I have completed my hypothesis."

My Hypothesis

If the cell has:
More Mitochondria.
Then it will be healthy.

Figure 4: The Simcell 2 Hypothesis Widget

This improves on Simcell 1's design in that rather than being bound to specifying only one organelle, the student may select as many organelles as they believe necessarily to correct the problem presented to them. The form that hypotheses in Simcell 2 are in is: "If the cell has: <more/fewer> <organelle one>, … , <more/fewer> <organelle n>, then the cell will be healthy".

Explore Activity and Experiment Phase

A substantial problem in Simcell 1 are the gauges in the explore activity and the experiment phase (displayed in figure 5). Simcell 1 has five gauges: life, energy, protein, water, and waste. It uses those five gauges to display the functions of the eight organelles. This led to a few cases where the function of an organelle was not properly conveyed by the microworld. This made it impossible for the student to gain any knowledge about that organelle. One example of this flaw is the vacuole. Middle school students are taught that the vacuole stores nutrients for the cell. However, in the Simcell interface, this functionality is expressed by increasing the protein gauge when more vacuoles are present.

In addition, these gauges did not do a great job in communicating to the student that the cell has too many of one type of organelle. In Simcell 1, when putting in an organelle, the relevant gauges will increase until a certain threshold is reached. If the student continues to add organelles beyond this threshold, the relevant gauges will begin to decrease. This creates an issue where the gauge is at the same level when the cell has too many of an organelle *and* too few of the same organelle. This failure of the UI can leave the student with uncertainty with how to proceed without assessing any relevant inquiry skills or domain knowledge.

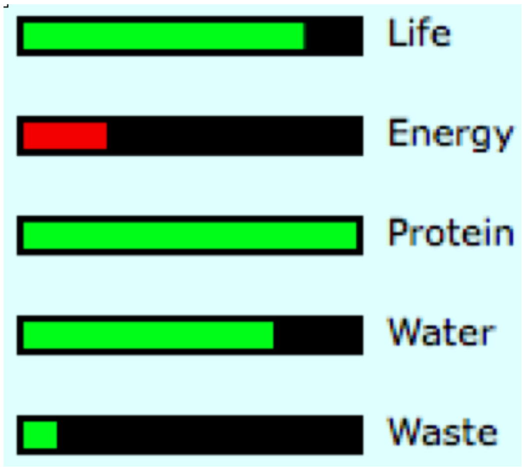


Figure 5: The Simcell 1 Gauges

The most significant problem with the experiment phase of the activity is how we ask the student to collect data. In Simcell 1, a student has the option to record the state of the cell at any time by clicking a record button located at the bottom of the screen. The problem with this method is that we do not require the student to use this button. Looking at the log data it seems very few students made use of this button and it seems likely that many students did not even notice it. If a student does not collect sufficient data in this phase, then completing the final phase of the activity is impossible. In turn, the student is left confused and the activity fails to capture the students' skill at interpreting data. Leaving the student with no way of completing the next phase of the activity.

The explore phase of Simcell 2 (Figure 6) gives the student an empty animal cell and asks students to work with the cell enough to be able to create informed hypotheses. This serves as the task to orient students towards the kind of work they will be doing throughout the rest of the activity. Organelles can be added to the cell by dragging them from the toolbar into the cell. Organelles can be removed by dragging them out of the cell and into the toolbar. The student can see the results of their changes by clicking 'Scan Cell'. This 'Scan Cell' button serves the student by allowing them to update the state of the cell. It also provides a critical point for researchers as it gives us a sense of when the student believes that they *need* to update the state. This can be a potential point for scaffolding or assessing the student. The student is unable to collect data during this phase and when students are ready they can return to the hypothesis phase.

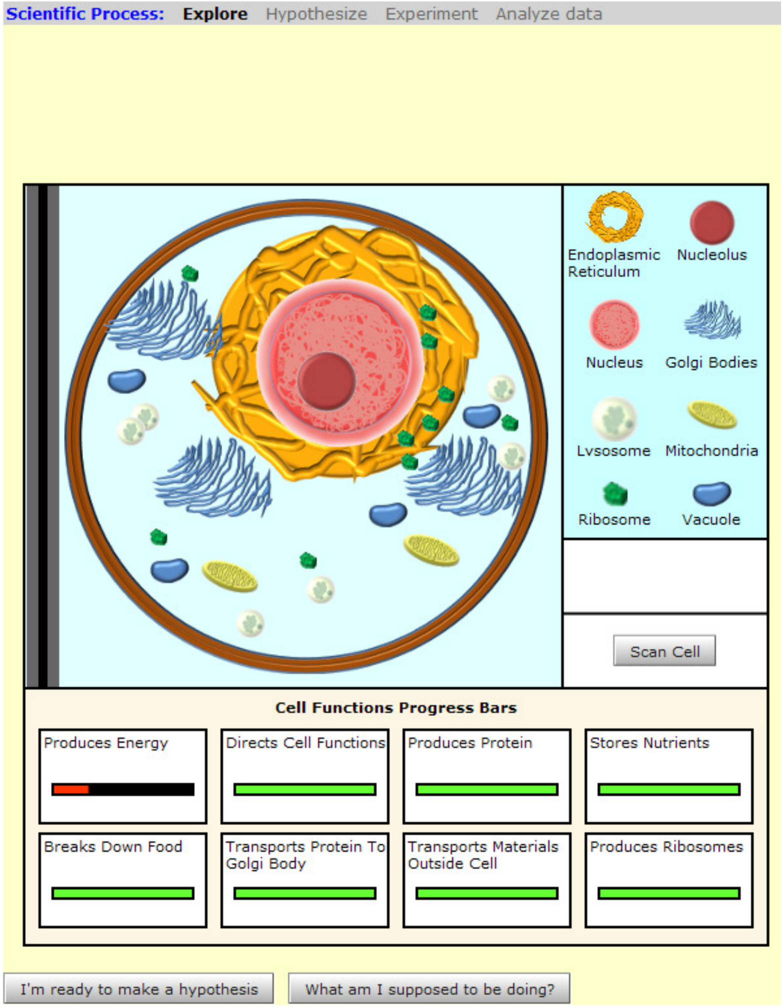
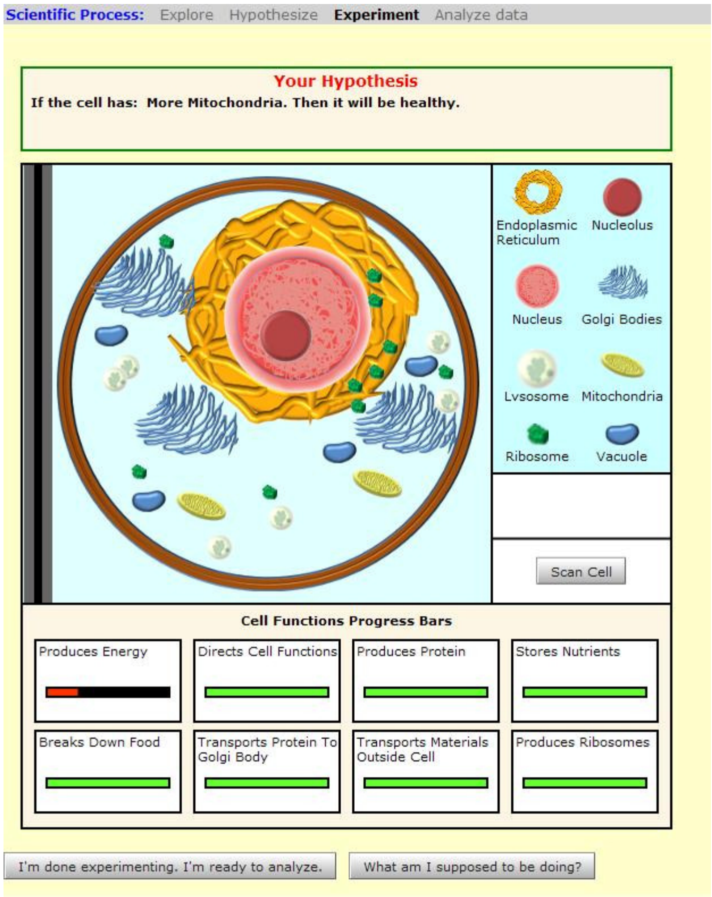


Figure 6: The Simcell 2 Explore Phase

The user interface of the experiment phase of Simcell 2 (Figure 7) is the same as the explore phase. Students test their hypothesis by either adding or removing organelles to the cell and hitting ‘Scan Cell’ to see the results of these changes. When a student hits ‘Scan Cell’, the state of the cell after the scan is recorded as data that students can use in the analyze phase. Then when they are finished testing their hypotheses, the students can go to the analyze phase.



Your Hypothesis

If the cell has: More Mitochondria. Then it will be healthy.

Scientific Process: Explore Hypothesize Experiment Analyze data
Endoplasmic Reticulum
Nucleolus
Nucleus
Golgi Bodies
Lvsosome
Mitochondria
Ribosome
Vacuole
Scan Cell
Cell Functions Progress Bars
Produces Energy
Directs Cell Functions
Produces Protein
Stores Nutrients
Breaks Down Food
Transports Protein To Golgi Body
Transports Materials Outside Cell
Produces Ribosomes
I'm done experimenting. I'm ready to analyze.
What am I supposed to be doing?

Figure 7: The Simcell 2 Experiment Phase

Explore Activity / Experiment Phase Improvements

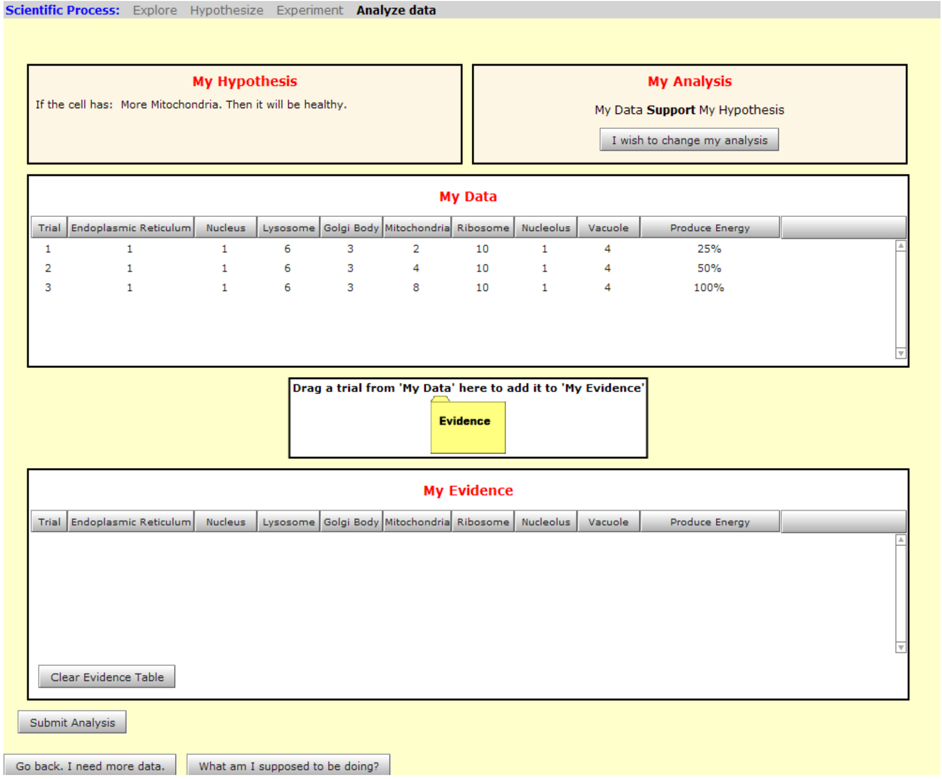
The experiment phase of the Simcell 2 activity went through many iterations to try to improve on its predecessor. The first improvement was to increase the number of gauges from five to eight: one for each organelle. Each gauge was given a short description that *directly* described the behavior of an organelle. The hope is that this would help the student make connections between the organelles and the descriptions of their respective functions. When a student adds/removes parts and the hits ‘Scan Cell’ they will see gauges increase/decrease, they have the potential to form a connection between the organelles they are adding/removing and the gauges those organelles affect. This was not feasible with the original design, as the five gauges did not convey enough information to discern the function of the eight organelles.

The method in which the organelles are updated was also redesigned in order to help us get better information from the student and to help the student collect data required to successfully complete the analyze phase. In Simcell 1, the gauges automatically updated as organelles were added or removed from the cell. Now, in Simcell 2, the gauges can only be updated through a scan procedure that the student initiates by hitting the “Scan” button. This scanning process is indicated through an animation that shows the student’s cell being scanned. Once the animation is complete, the gauges are updated. For each gauge a status is placed above the gauge that either says “Increased/ decreased, too much” depending on if the cell’s capability to perform the functionality desired in the gauge description either improved, got worse, or if the student simply has overwhelmed the cell with a particular organelle. If there was no change, the gauge does not receive a status. These status messages were added in order to help offload short term memory by giving a student an easy way to keep track of how their most recent change affected the health of the cell.

Analysis Phase

The analysis phase has the same issues as the hypothesis phase. The widget simply did not have the expressive power required to describe the interventions students could make to the cell. Therefore the interface would get in students’ way when trying to solve the problem presented to them. Once again, in order to have a completely proper analysis in scenario two, the student was required to make two disconnected data interpretations.

The analyze phase of Simcell 2 (Figure 9) shows the students the data they collected during the experiment phase. First they are asked to make a statement about whether their data either supports or refutes their hypothesis. After that they are asked to point to trials from their experiment that show their statement to be accurate. Students select trials by dragging them into a folder labeled “Evidence”.



Scientific Process: Explore Hypothesize Experiment Analyze data

My Hypothesis

If the cell has: More Mitochondria. Then it will be healthy.

My Analysis

My Data **Support** My Hypothesis

I wish to change my analysis

My Data

Trial	Endoplasmic Reticulum	Nucleus	Lysosome	Golgi Body	Mitochondria	Ribosome	Nucleolus	Vacuole	Produce Energy
1	1	1	6	3	2	10	1	4	25%
2	1	1	6	3	4	10	1	4	50%
3	1	1	6	3	8	10	1	4	100%

Drag a trial from 'My Data' here to add it to 'My Evidence'

My Evidence

Trial	Endoplasmic Reticulum	Nucleus	Lysosome	Golgi Body	Mitochondria	Ribosome	Nucleolus	Vacuole	Produce Energy

Figure 8: The Simcell 2 Analysis Phase

The analyze phase of Simcell 2 went through the largest redesign of the activity. This improves on the analysis phase on Simcell 1 in that it provides the expressive power required to fully form data interpretations. The choice to have students identify their data as either supporting or refuting their hypothesis separately from the identification of trials that demonstrate this fact, gives researchers two indices of student’s reasoning processes. For example if the student believes their data supports their hypothesis but is incorrect, the system can potentially jump in and suggest the student collect more data before going on. If the student ignores this message, we have another opportunity to intervene when the student attempts to select trials to back up their false claim.

Study Two

The purpose of study two was to judge the effectiveness of Simcell 2 and to collect data for the eventual addition of help for the student in the form of adaptive scaffolds.

Participants

Participants were 92 middle school students from four classes of a large urban school in Central Massachusetts. This school has approximately 38% Caucasian students, 37% Hispanic students, 14% African American students, and 8% Asian students. Their MCAS science performance is typically low -- approximately 70% of students scoring “below proficient” on their most recent MCAS Science test. Analyses were performed on the students who completed the relevant portions of the activity.

Materials

Inquiry Pretest

The inquiry pre and posttests were dropped from this study to allow students more time to complete the activity.

Domain Pretest

This pretest was completely rewritten from that which was used in Study 1. The test for Study 2 consisted of eight multiple choice questions that are more closely aligned to the activity. Both scenarios target one of more of the questions in this test.

SimCell 2.0

Explore Phase

This section places students in the explore phase in the Simcell app cycle. They cannot go to the other phases during this section.

Problem Scenario One

The first problem scenario shows the students an animal cell that is unable to produce enough energy due to a lack of mitochondria. This uses the Simcell app cycle to have students form a hypothesis, test that hypothesis, and analyze their data. This scenario targets one of the problems in the domain pre and posttests.

Problem Scenario Two

The second problem scenario is identical to the first problem scenario except students are shown an animal cell that cannot store nutrients or break down food. This scenario targets two of the problems in the domain pre and posttests.

Domain Posttest

The domain posttest was identical to the domain pretest. During previous studies researchers have seen that students have less motivation to take the posttest seriously due to the questions being the same. Researchers have observed that some students think they are being asked these questions by mistake. Based on student scores it seems that this was not a problem in this study as there was not an overall drop in test scores and the average score (72% correct) was way above the expected score guessing randomly (25%), suggesting that students took the posttest seriously and did not simply guess their way through it.

Procedure

The activity was designed to run in one day. The activity started with the pretest. Afterwards students were given the Simcell 2 explore section, followed by Simcell 2's scenario 1 and scenario 2. The activity concluded with the posttest.

Data Coding

Pre and Posttests.

Pre and posttests were autoscored by the Assistments system.

Student Logs

While students go through the activity, the system logs every event that occurs in interface. This creates a large amount of data per student. To make this data set useful, it was broken down into individual variables that could be analyzed with various statistical methods. Each variable is described below. Where applicable, there is one instance of a variable for each scenario.

Test And Overall Activity Variables

Pretest Targeted Questions Score

The pretest targeted problems score is the student's score on the pretest if only questions directly targeted by scenarios from this activity are taken into account. A targeted pre- (or post-) test question asks about content that the problem scenario is designed to assess. For example: "What do the Mitochondria do?" is an example of a targeted pretest question because in problem scenario one the cell presented to students is low on energy and can be fixed only by adding Mitochondria. The question "What does the Endoplasmic Reticulum do?" is a non-targeted question because there is no problem scenario that requires the student to work with this organelle. There is a potential chance that the student could gather this information because they *can* choose to see the effects of the Endoplasmic Reticulum on the cell in any problem scenario or during the explore phase of the activity. In the pre- and post-test there are three targeted questions and eight non-targeted questions. Non targeted questions were part of the test to allow us to get a more full baseline of the students' knowledge of these topics and to see if working with the microworld could perhaps lead to gains, by means of inference-making, in these areas not directly tested.

Pretest Time

This is the time spent completing the pretest.

Total Time Activity

This is the time spent completing the activity.

Explore Variables

Explore Time

This is the time spent during the explore section of the activity.

Completed Cell In Explore Section

This variable tracks whether or not the student built a perfect cell in the explore section.

Number Of Scans In Explore Section

This is the number of times the student hit scan cell in the explore section of the activity.

Number Of Actions In Explore Section

This is the number of actions the student took in the explore section of the activity. An action is adding an organelle, removing an organelle, moving an organelle, scanning the cell, or clicking any of the buttons on the interface.

Inter Scan Variance In Explore Section

This variable tracks the number of *different* organelles the student interacts with between scans. Rather than reporting any individual scan we count the number worked in the last three scans and report an average. The last three scans are used since we wished to see if a student ended with more systematic inquiry trials (cf. Buckley et al, 2010). The rationale behind the use of this variable is that it tracks how well the student is controlling for variables during their experimental trials. Simultaneously working with more organelles leads to higher confounding data.

Mean Action Time In Explore Section

This variable tracks the mean of the time in between actions. This variable potentially acts as a sign of the student's mindfulness. If the mean action time is low there is not a lot of time for thought between actions. If it is higher, then it may suggest that the student is actively processing the material (Sternberg, 1969).

Standard Deviation Of Student Action Time In Explore Section

This variable tracks the standard deviation of the time between actions. This tracks how consistently the student spends time in between their actions.

Scenario Variables

Scenario Time

This variable is the time spent completing the scenario.

Returned To Explore

This variable tracks if the student decided to return to the explore section of the activity.

Time To Form Hypothesis

This variable is the time the student spent forming their hypothesis.

Time Reading Instructions In Hypothesis Phase

This variable is the time spent reading the instructions in the hypothesis phase of the scenario.

Hypothesis Consistent With The Domain

This variable is whether or not the student's hypothesis was one that happens to be scientifically accurate. For example: if the cell is low on energy, a hypothesis that states that the cell needs more mitochondria would be consistent with the domain. Any other hypothesis would not be consistent with the domain. This variable is autoscored using a knowledge-engineered grading rule. More specifically, for each scenario there are only subsets of possible hypotheses that are consistent with the domain, so the auto grading rules check if the student's hypothesis matches one of those.

Mean Action Time In Hypothesis Phase

This variable tracks the mean of the time in between actions. See: "*Mean Action Time In Explore Section."*

Standard Deviation Action Time In Hypothesis Phase

This is the standard deviation of time spent between actions. See: "*Standard Deviation Time In Explore Section."*

Completed Cell In Experiment Phase

This variable tracks whether or not the student completed the cell in the experiment phase of the scenario.

Number Of Scans In Experiment Phase

This variable tracks the number of scans the student does in the experiment phase of the scenario.

Number Of Actions In Experiment Phase

This variable tracks the number of actions the student performed in the experiment phase of the scenario. See: *"Number Of Actions In Explore Section"*.

Did Student Test Their Hypothesis?

This variable tracks if the student tested their hypothesis during the experiment phase of the activity. This is done using a knowledge-engineered grading rule that looks at the state of the cell after each scan in order to determine if the student made changes that would be a reasonable test of the hypothesis they provided. More specifically, given a set of trials (each trial is recorded when the student hits "scan cell"), the grader determines if the student hypothesis is tested based on the following criteria: 1) Whether or not the cell is in the state presented to the student at the start of the problem, and 2) If the hypothesis is consistent with the domain (See: *Hypothesis*

Consistent With The Domain). If the hypothesis is not consistent with the domain, the grader looks for a trial where the student either added or removed all the parts the student specified in their hypothesis.

First Scan In Experiment Unrelated To Hypothesis

This variable checks to see if the student's first scan involved an organelle that is in the student's hypothesis.

Inter Scan Variance In Experiment Phase

This variable is the inter scan variance in the experiment phase. See: *"Inter Scan Variance in Explore Section"*.

Mean Action Time In Experiment Phase

This variable tracks the mean of the time in between actions. See: "*Mean Action Time In Explore Section"*.

Standard Deviation Action Time In Experiment Phase

This is the standard deviation of time spent between actions. See: "*Standard Deviation Time In Explore Section"*.

Conclusion Correct In Analysis Phase

This variable is whether or not the student was able to correctly identify if their hypothesis is consistent with the domain. See: *"Hypothesis Consistent With The Domain"*.

Number Of Trials Used As Evidence In Analysis Phase

This variable is the number of trials the student selected from their data set to back up their conclusion in the analysis phase.

Evidence Backs Up Conclusion In Analysis Phase

This variable is an autoscored attempt to see if the student's selected trials back up the conclusion that the student made. This is computed using a knowledge-engineered grading rule that checks for two types of occurrences to determine that the student has backed up their conclusion: 1) If the student's conclusion is one that could be backed up with evidence that could *potentially* be gathered in the environment, and 2) If the trials that the student selected as evidence represent a tested hypothesis. See: "*Did Student Test Their Hypothesis"* to see how the grader checks if a set of trials represent a hypothesis test.

Too Many Trials In Analysis Phase

This variable checks to see if the student had more than the three trials. This is an easy way of seeing if the student is not mindfully picking trials since they should only need 2-3 trials to back up their analysis.

Mean Action Time in Analysis Phase

This variable tracks the mean of the time in between actions. See: "*Mean Action Time In Explore Section"*.

Standard Deviation Action Time In Analysis Phase

This is the standard deviation of time spent between actions. See: "*Standard Deviation Time In Explore Section"*.

Open Response Code

The open responses were graded by a human coder and given one of four scores: Correct, Incorrect, Irrelevant, and Blank.

Open Response Length

The number of characters in the student's open response.

Research Questions

1. Do students learn from the updated activity, as evidenced by gains on the post-test

2. Can we predict student performance on posttests based on their actions in the activity?

Results

Do students learn from the updated activity?

	Pretest Mean	Pretest Standard Deviation	Posttest Mean	Pretest Standard Deviation
Complete Test	.68	.22	.73	.23
Targeted Problems	.74	.27	.85	.25
Non Targeted Problems	.64	.25	.65	.26

Table 9: Results for Simcell 2 Pre-post tests

A paired-samples t-test was conducted to compare students' pre and posttest scores. There was a significant difference in scores from pretest to posttest: t(65) = -2.18, p = .033, d = 0.21.

A paired-samples t-test was conducted to compare students' scores on pre and post problems directly targeted by the microworld scenarios. There was a significant difference in scores from pretest targeted problems to posttest targeted problems; t(66) = -3.06, p = .003, d = 0.42, thus, there is a much larger effect for the items targeted by the activity than the test as a whole.

A paired-samples t-test was conducted to compare students' pre and post problems not directly targeted by the problem scenarios. There was no significant difference in scores from pretest to posttest; t(65) = -.31, p = .76.

Can we predict student performance on posttests based on their actions in the activity?

Original Variables

A forward regression using all the variables coded to predict students' targeted posttest score found three predictors: the students' targeted pretest score (x_1), if the student returned to explore in scenario 1 (x_2), and the standard deviation of student action time in explore (x_3). The final model was: $.38x_1 + -.24x_2 + -.23x_3$ ($R^2 = .29$, Adjusted $R^2 = .25$).

Due to the large number of variables and comparatively small number of data points, a test was run to attempt to check the validity of the above result. To do this, four data files were generated with the same number of variables as the data set. Each row of the new data files was filled in with a random value using Excel's rand() function. Using each of these data files, a forward regression was run attempting to predict one of the variables using the other variables in the random data sets. The models that came out had comparable R^2 values (.5, .34, .6).

Variables Selected By Domain Expert

After the above test it was clear the number of variables had to be decreased. To do this we consulted a domain expert. She selected variables that she deemed would capture student knowledge and inquiry processes. The following variables were selected:

- Open Response Code
- Pretest Score on Non Targeted Problems
- Explore Time
- Scenario Time
- Completed Cell In Explore Section
- Number Of Scans In Explore Section
- Number of Actions In Explore Section
- Returned To Explore
- Hypothesis Consistent With The Domain
- Completed Cell In Experiment Phase
- Number Of Scans In Experiment Phase

- Number Of Actions In Experiment Phase
- Did Student Test Their Hypothesis
- Number Of Trials Used As Evidence In Analysis Phase
- Evidence Backs Up Conclusion In Analysis Phase

A forward regression using all the selected variables over both scenario one and two to predict students' targeted posttest score found two predictors: the students' targeted pretest score (x_1), and if the student returned to the explore phase (x_2). The final model was: $.45x_1 + -.23x_2$ ($R^2 = .3$. Adjusted $R^2 = .28$).

A forward regression using the selected variables from scenario one, explore, and the pretest to predict students' targeted posttest score found two predictors: the students' targeted pretest score (x_1), and if the student returned to explore (x_2). The final model was: $.45x_1 + -.23x_2$ ($R^2 = .3$. Adjusted $R^2 = .28$).

A forward regression using the selected variables from scenario two, explore, and the pretest to predict students' targeted posttest score found one predictor: the students' targeted pretest score (x_1). The final model was: $.5x_1$ ($R^2 = .25$. Adjusted $R^2 = .24$).

Factor Analyses

In an attempt to further reduce the number of variables and hopefully get more predictive accuracy, the selected variables were used in a series of factor analyses. The first factor analysis used all of the selected variables. The factor analysis gave eight factors, which were put into a forward regression model that found zero predictors.

The second factor analysis used the selected variables from scenario one, explore, and the pretest. The analysis gave five factors that were put into a forward regression model that found one predictor: factor one (x_1). The final model was $.277x_1$ ($R^2 = .08$, Adjusted $R^2 = .06$).

The final factor analysis used the selected variables from scenario two, explore, and the pretest. The analysis gave five factors that were put into a forward regression model that found zero predictors.

Discussion

Did Students Learn From The Activity?

Based on the paired samples t-test there was at least some learning in this activity. Based on the t-test on targeted problems and non-targeted problems, it seems that all of this learning was on the problems targeted by the microworld scenarios. This result has positive and negative connotations. The positive aspect of this is that it seems our microworld had some success in teaching students the material required. The less than positive news is that students' limited interactions with the non-targeted organelles in the explore phase did not lead to even a small gain in those non-targeted areas as hoped.

Did The Interface Improvements Make This A Better Activity?

This question is difficult to answer quantitatively. However, an argument can be made that Simcell 2 was more successful as a microworld than Simcell One. Looking at the test scores between study one and study two, students showed learning gains for both microworlds. Not only are the microworlds themselves quite different, but the pretests written for them are very different as well. This means that comparing them empirically is not possible. That being said, the data does seem to show Simcell 2's potential as an inquiry environment.

This does not begin to answer if Simcell 2 was *better*. Considering the results of the targeted pretest questions, it seems Simcell 2 did a great job at teaching students the materials that scenarios target. Due to the fact that Simcell 2 has one gauge per organelle, it is possible to have scenarios that can target each of the animal cell functionality that middle school frameworks expect students to know. This was not possible in Simcell One due to the fact that the gauges presented to the students obfuscated the functionality of many organelles.

That being said, Simcell 2 was not ideal. The largest problem encountered in the field was that students were confused when they reached the analysis phase, specifically in the section where students had to drag trials that backed up their analysis. Many students complained about this and it was one of the major failures of Simcell 2.

Can we predict Students Posttest Performance Based On The Activity?

Based on the analyses ran for this study, we do not have a successful method to predict student posttest performance. The predictive power of even our most successful methods was limited. Another cause for concern is that upon analyzing the results it became clear that it was not possible to distinguish if the models were actually predicting student performance or if they were simply modeling noise; this is primarily due to the relatively low number of participants and the high number of variables used. Because of this we cannot make any substantial claims about what interactions affect student performance in Simcell 2. This does not necessarily mean it is impossible to predict student performance; what we have demonstrated here is the modeling of students in a microworld is a difficult problem that needs further study.

While this is a disappointing result, there is a potentially interesting possibility. Perhaps the inability to predict the results of a multiple choice test demonstrates that the tasks required in our microworld *are* significantly different from what the MCAS and other multiple choice tests assess. While we would like to succeed in having students do well on this multiple choice test *and* go beyond them, it is a potentially good sign that our microworlds do not align perfectly or tightly with them. Otherwise there would be little gain from having students use our system when we could simply give them flash cards to memorize the rote facts tested.

Conclusions, Contributions, and Future Work

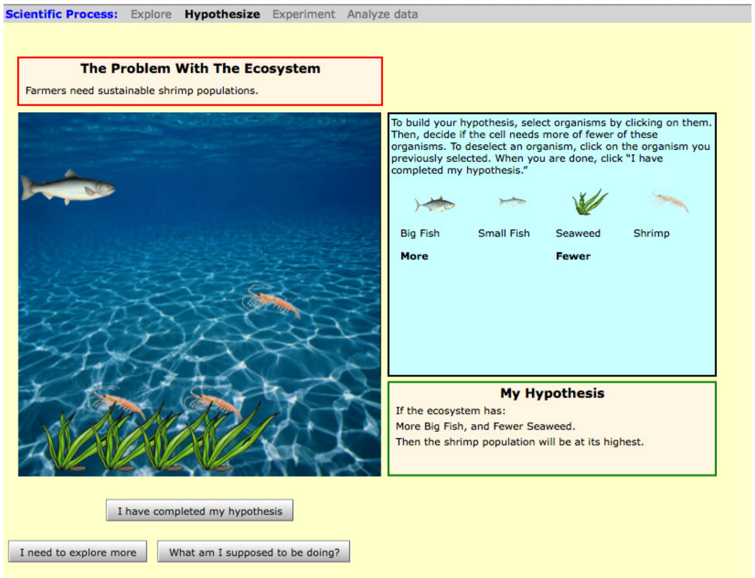
Identifying points for scaffolding

The results from study one and study two fail to indicate clear and precise moments for scaffolding. However, the results of study one did provide evidence that students can be assessed early in the activity and this finding can potentially lead to a scaffolding system that can catch students who are having problems with inquiry early on and possibly help put them on a productive path to inquiry learning. This would allow the system to know where students stand while still providing a worthwhile learning experience. Study two was less successful in finding methods to directly scaffold students. The data simply did not provide enough actionable intervention points. The results demonstrate the difficulties in student modeling in inquiry environments. These results should serve as a warning sign to future researchers to be careful with what they present as results to the community. It would have been very easy to look at these results and write up a document about the effects shown in them without noticing the underlying flaws in the methodology that were exposed. Specifically, it is now clear that when analyzing a dataset with so few data points, there is a need to be careful of the model you try to fit since it becomes very easy for statistical methods to simply find patterns in noise.

Simcell Microworld

The results of study one and two demonstrate that the microworlds was at least partially successful in helping students learn the material. We had learning gains in both studies and study two demonstrated gains directly related to the specific topics covered by the problem scenarios. The quantitative results presented here do not necessarily demonstrate that the redesign improved the microworlds. The learning gains were similar in both studies and we lack the comprehensive study comparing the two microworlds directly as it was outside the scope of this work. That being said, there are some aspects of Simcell 2 that demonstrate its superiority. The first of which is that the one to one relationship between an organelle and a gauge un-confounded the functionality of certain organelles. With one gauge per organelle each topic can be expressed clearly instead of obscuring the functionality of some organelles by trying to express them in gauges that do not clearly communicate their functionality. Another point is that the log data demonstrates that the addition of the scan cell button in Simcell 2 forced students to record data as they progressed though the experiment phase, thus affording them the tools they needed to complete the analysis phase correctly. The design of the Simcell one made recording data optional and the majority of students did not collect sufficient data to move forward.

This microworld will be one of many that serve as the Science Assistments life science microworld suite, eventually providing teachers a set of tools to pinpoint areas they need to reinforce in class by assessing their students regularly. Simcell 2 has also served as the basis for other microworlds in this suite. Specifically, the recently developed an Ecolife microworld that re-uses the codebase and UI for Simcell 2 as a starting template (see figure 10).



The Problem With The Ecosystem

Farmers need sustainable shrimp populations.

To build your hypothesis, select organisms by clicking on them. Then, decide if the cell needs more of fewer of these organisms. To deselect an organism, click on the organism you previously selected. When you are done, click "I have completed my hypothesis."

Big Fish: More
Seaweed: Fewer

My Hypothesis

If the ecosystem has:
More Big Fish, and Fewer Seaweed.
Then the shrimp population will be at its highest.

Figure 9: Ecolife Microworld based on Simcell 2

Future Work

There is a lot of fodder here for future researchers to build on. The first of which is the microworld itself. Based on time in the classroom and the log data, it seems that students still struggle with the analysis phase of this activity. There is still a very real possibility that the UI for this phase of Simcell 2 is still far too complicated.

The system in this work was designed for a specific cell. Future programmers, designers, and researchers can extend this work to include other types of cells to allow Simcell to be a more general-purpose microworld for cellular biology.

The modeling approach used in this work was simply unsuccessful. Future work in this front can revisit this aspect and experiment with different modeling approaches. One possibility is revisiting the assertions made earlier in this document suggesting that Bayesian Knowledge Tracing and Performance Factor Analysis are inappropriate for this environment. A more fine-grained analysis could be done in the experiment section, potentially identifying skills that have enough practice attempts to achieve some reasonable predictive power. Another idea would be to attempt a more coarse-grained analysis. For example, take the skill being measured to be "inquiry" and identify virtually every action in the system that can be either correct or incorrect as an action. Perhaps with this we can get some sort of prediction that can be used as a marker identifying if the student needs help or is successfully inquiring about the cell.

References

Bachmann, M., Gobert, J., & Beck, J. (2010). Tracking Students' Inquiry Paths through Student Transition Analysis. *Proceedings of the 3rd International Conference on Educational Data Mining* , 269-270.

Baker, K., Hashmi, S., & Prifti, F. (2010). *Developing Technology-Based Biology Assessments for Cell Structures and Functions.* Interative Qualifying Project WPI Worcester Massachussetts.

Berthelsen, B. (1999). Naive Conceptions in Life Science. *MSTA Journal, 44* (1), 13-19.

Buckley, B. C. (2006). Using Log Files To Track Students' Model-based Inquiry. *Proceedings of the Seventh International Conference of the Learning Sciences*, 57-63.

Buckley, B., Gobert, J. & Horwitz, P. (2006). Using Log files to Track Students' Model-based Inquiry. In the *Proceedings of the Seventh International Conference of the Learning Sciences (ICLS),* Mawah: NJ: Erlbaum, pp.57-63.

Champagne, A., K. Bergin, R. Bybee, R. Duschl, & J. Gallagher. (2004). *Science Framework Development: Issues and Recommendation.* Washington, DC.

Corbett, A., & Anderson, J. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 253-278.

Feng, M., Heffernan, N., & Koedinger, K. (2006). Predicting State Test Scores Better with Intelligent Tutoring Systems: Developing Metrics to Measure Assistance Required. In Mitsuru I, Ashley K, Chan T-W (Eds.)*, Intelligent Tutoring Systems*. Heidelberg Germany, Springer Berlin, pp 31-40.

Feng. M., H. N. (2010). Can We Get Better Assessment From A Tutoring System Compared to Traditional Paper Testing? Can We Have Our Cake (better assessment) and Eat it too (student learning during the test). *Proceedings of the 3rd International Conference on Educational Data Mining*, 41-50.

Flores, F. T. (2003). Representation of the Cell and Its Processes in High School Students: An Integrated View. *International Journal of Science Education, 25* (2), 269-286.

Gobert, J., Heffernan, N., Koedinger, K., and Beck, J. (2009). *Assistments Meets Science Learning AMSL.* Proposal (R305A090170) funded by the US Department of Education.

Gobert, J., Heffernan, N., Ruiz, C., & Kim, R. (2008). *Assistments Meets Inquiry.* Annual Report submitted to the National Science Foundation for NSF-DRL# 0733286.

Gobert, J., Sao Pedro, M., Baker, R.S., Toto, E., & Montalvo, O. (in press). *Leveraging educational data mining for real time performance assessment of scientific inquiry skills within microworlds,* Journal of Educational Data Mining.

Gong, Y. B. (2010). How to Construct More Accurate Student Models: Comparing and Optimizing Knowledge Tracing and Performance Factor Analysis. *International Journal of Artifical Intelligence in Education*, *21(1),* 27-46.

Howell, R. D., Scott, P. B., & Diamond, J. (1987). The Effects of "Instant" Logo Computing Language on the Cognitive Development of Very Young Children. *Educational Computing Research*, 249-260.

Lawson, A. E. (2010). *Teaching Inquiry Science in Middle and Secondary Schools.* Thousand Oaks: SAGE Publications, Inc.

Massachusetts Department of Education. (2006). *Massachusetts Science and Technology/Engineering Curriculum Framework.* Malden: Massachusetts Department of Education.

Montalvo, O. B. (2010). Identifying Student' Inquiry Planning Using Machine Learning. *Proceedings of the 3rd International Conference on Educational Data Mining*, 141-150.

National Center for Education Statistics. (2009). *The Nations Report Card: National Assessment Of Education Progress At Grades 4, 8, 12.* United States Department of Education.

Natonal Research Council. (1996). *National Science Education Standards.* Washington, DC: National Academy Press.

Pavlik Jr., P. C. (2009). Learning factors transfer analysis: Using learning curve analysis to automatically generate domain models. *Proceedings of the The 2nd International Conference on Educational Data Mining*, 121-130.

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

Razzaq, L., Mingyu, F., Nuzzo-Jones, G., Heffernan, N. T., Koedinger, K., Brian, J., et al. (2005). Blending Assessment and Instructional Assisting . *Proceedings of the 12th Artificial Intelligence In Education*, 555-562.

Snyder, T. D., Dillow, S., & Hoffman, C. (2009). *International Comparisons of Education Digest of Education Statistics 2008.* National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica 30*, 276–315.

Sweller, J. (1994). Cognitive Load Theory, Learning Difficulty, and Instructional Design. *Learning and Instruction*, 295-312.

United States Department Of Labor. (2007). *The STEM Workforce Challenge: The Role of the Public Workforce System in a National Solution for a Competitive Science, Technology, Engineering, and Mathematics (STEM) Workforce.* U.S. Department of Labor, Employment and Training Administration by Jobs for the Future.

Appendix A

Appendix B