Looking inside the black box: assessing model-based learning and inquiry in BioLogica™

Barbara C. Buckley*

WestEd,
400 Seaport Court, Suite 222,
Redwood City, CA 94063, USA
E-mail: bbuckle@wested.org
*Corresponding author

Janice D. Gobert

Social Sciences and Policy Studies Dept.,
Worcester Polytechnic Institute, Atwater Kent Labs,
Worcester, MA 01609 USA
E-mail: jgobert@wpi.edu

Paul Horwitz

The Concord Consortium,
25 Love Lane,
Concord, MA 01742, USA
E-mail: phorwitz@concord.org

Laura M. O'Dwyer

Boston College,
Lynch School of Education,
Campion Hall, Room 336E,
Newton, MA 02467, USA
E-mail: odwyerl@bc.edu

Abstract: The Modeling Across the Curriculum Project (MAC; IERI # 0115699, Oct 2001–2006) used real-time assessments to facilitate student learning and model-based inquiry among high school students. We developed technology, materials, and processes that enabled us to monitor and respond to students' actions. MAC learning activities engage students in a progressive model-building approach (Gobert, 2008; White and Frederiksen, 1990). Formative assessments were seamlessly embedded in scaffolding designed to guide model-based learning and inquiry. Because instruction and assessment were integrated, we were able to measure model-based inquiry skills *in situ*, thus circumventing the problem of assessing inquiry separate from its context (Mislevy et al., 2002). After identifying useful log file data and developing algorithms for analysing that data on a large scale, we identified productive inquiry strategies that correlated with learning gains. Our findings have immediate applicability to the design of tasks intended to elicit and support rich inquiry learning.

Keywords: technology-based science learning; performance assessment; embedded assessment; formative assessment; model-based learning; genetics learning; problem-solving; model of inheritance; data-mining.

Reference to this paper should be made as follows: Buckley, B.C., Gobert, J.D., Horwitz, P. and O'Dwyer, L.M. (2010) 'Looking inside the black box: assessing model-based learning and inquiry in BioLogica^{TM}', *Int. J. Learning Technology,* Vol. 5, No. 2, pp.166–190.

Biographical notes: Barbara Buckley is a Senior Research Associate at WestEd, where she leads content teams in biological, earth and physical science as they develop simulation-based assessments and inquiry activities for middle school classrooms. Her primary focus is on the use of technology for supporting model-based learning and assessment in science.

Janice Gobert is an Associate Professor of Learning Sciences and Psychology in the Social Sciences and Policy Studies Department and the Computer Science Department at Worcester Polytechnic Institute. Her areas of expertise are technology-based learning and assessment in science, in particular with visualisations such as simulations and student's epistemologies of models in science and their interactions with learning.

Paul Horwitz directs the Modeling Center at the Concord Consortium. He received his PhD in Theoretical Physics from New York University. His educational research interests centre around the use of technology for helping students learn to use mental models in science and mathematics.

Laura M. O'Dwyer is an Assistant Professor in the Department of Educational Research, Measurement and Evaluation at Boston College where she teaches experimental design and advanced data analysis, and is a Senior Research Associate at CSTEEP. She has examined educational issues such as the impact of tracking in middle school mathematics classrooms, the use of educational technologies as a teaching and learning tool, and the impact of virtual algebra courses on high school achievement.

1 Introduction

With the enactment of the No Child Left Behind Act of 2001 and the Education Science Reform Act of 2002 (Public Law 107–110), the current level of accountability in education demands evidence-based research and higher levels of performance for students at all skill levels. This makes it critical to be able to assess students' learning reliably, (Fadel et al., 2007) and places a particularly high value on timely formative assessments that can assess students' understanding and thereby help produce the desired learning gains. In science education, this means being able to assess not only students' content knowledge, but also their model-based inquiry skills since these will enable them to reason about science content and will support future science learning (Gobert et al., 2007a). While important progress has been made on assessing students' content knowledge (National Research Council, 2002a), the assessment of process skills such as inquiry model-based inquiry, equally important for scientific literacy (Perkins, 1986), has lagged. Inquiry skills are considered an important component of scientific literacy because it is through these skills that students acquire new knowledge and are able to transfer their knowledge to unfamiliar domains (National Research Council, 1999, 2000, 2000b; Gobert et al., 2007b).

Despite the widely acknowledged need for inquiry at all levels of the science curriculum (National Research Council (US), 1996), very few assessments exist for measuring or quantifying inquiry. Existing large-scale assessments fail to address inquiry skills (Quellmalz and Pelligrino, 2009). The situation is further complicated by the fact that it is difficult to separate inquiry from context. Inquiry skills developed in rich scientific contexts must be assessed within the scientific domain and context in which they are embedded (Mislevy et al., 2002). Like many others, we have found that traditional assessments fail to capture either the complex understanding or inquiry skills needed to conduct and learn from inquiry (Buckley et al., 2002; Ayala et al., 2002; Shavelson et al., 2002). Given the current emphasis on accountability, the need for and importance of assessing inquiry skills *in situ* has important ramifications for students, teachers, schools, and policy makers as well as for science education reform efforts.

2 Modeling Across the Curriculum (MAC) project

This paper describes the efforts of the IERI-funded project MAC to create technology-enhanced assessments grounded in a theory of model-based learning (Buckley, 1992, 2000; Gobert and Buckley, 2000), embedded in computer-based learning activities guided by model-based scaffolding (Buckley, 2000; Buckley et al., 2004; Gobert and Buckley, 2000), and enabled by Pedagogica™ (Horwitz and Burke, 2002). The experimentation with models demanded of students in MAC learning activities is a task much more analogous to real-world scientific methods than the act of answering a collection of unrelated multiple-choice questions. The inferences we make by analysing the log files automatically generated while students perform model-based inquiry tasks produce insightful and rigorous formative assessments that can guide learning without disrupting it.

The MAC project explored the potential of a powerful new approach to instruction, assessment, and educational research: one that combined the collection of extremely fine-grained student performance data with virtually unlimited scalability. We conducted this work in three different scientific domains: genetics, gas laws, and Newtonian mechanics. We offered students problem-solving activities supported by manipulable computer models linked to context-sensitive scaffolding. We then logged their actions as they attempted to solve the problems. The resulting log files provided a wealth of data bearing on the students' content learning as well as their inquiry skills and their ability to reason with models within these three different scientific domains. In the interest of coherence this paper focuses on just one domain, Genetics, as implemented in BioLogica™.

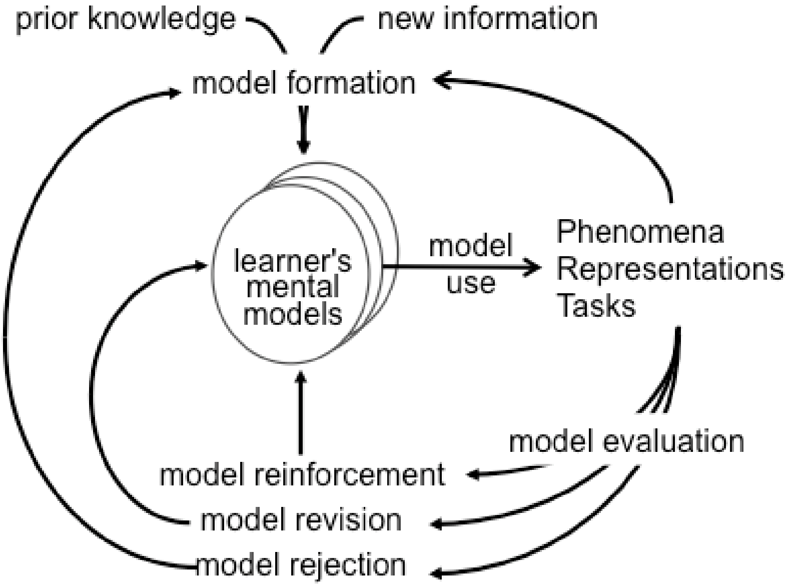
The central technological achievement of the MAC project was the development of *hypermodels* – manipulable computer models linked to text, scaffolding, and formative assessments. Students' interactions with these hypermodels were monitored and used to provide immediate tailored feedback to the students. At the termination of each computer session, data in the form of XML-tagged log files were encrypted and uploaded to a central server where they were decrypted, parsed, and used to populate a database. We implemented a set of data-mining tools and used them to analyse the logged data and to produce classroom-level reports for teachers and to generate input for a standard statistics package. With multiple points of redundancy between client and server software, the

MAC technology comprised a robust, distributed computing environment designed *ab initio* for scalability and adaptable to the delivery of different kinds of educational applications and assessments.

3 Theoretical framework: model-based learning

The MAC project organised research, learning activities, and assessment around model-based learning (Buckley and Boulter, 2000; Clement, 1989; Gobert and Buckley, 2000; Gobert and Clement, 1999), a theory of science learning that integrates basic research in cognitive psychology and science education.

Figure 1 Theoretical framework: model-based learning and reasoning in science

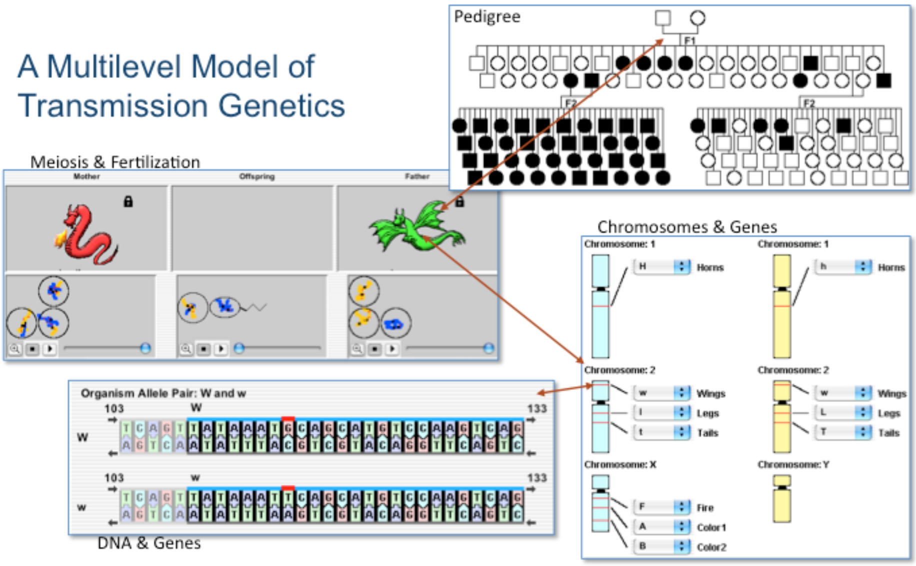


The tenets of model-based learning are based on the hypothesis that understanding requires the construction of mental models of the phenomena under study, and that all subsequent problem-solving, inference making, or reasoning are done by ‘running’ and manipulating these mental models (Johnson-Laird, 1983). We view mental models as internal, cognitive representations used in reasoning of many kinds (Brewer, 1987; Rouse and Morris, 1986). As shown in Figure 1, mental models, like prior knowledge, influence our perceptions of phenomena and our understanding of information. Interactions with phenomena, representations, and tasks, in turn, influence our mental models (Gentner and Stevens, 1983; Johnson-Laird, 1983). Thus, we define model-based learning as a dynamic, recursive process of learning by constructing mental models of the phenomenon under study. It involves the formation, testing, and subsequent reinforcement, revision, or rejection of those mental models (Buckley et al., 2002; Gobert and Buckley, 2000). This is analogous to hypothesis development and testing seen among scientists (Clement, 1989) and therefore, we argue the reasoning needed in inquiry. These higher order skills – hypothesis generation from working with or observing the model or phenomenon, testing that hypothesis, and interpreting the data – are critical to inquiry but difficult to assess.

4 Model-based instruction and assessment with BioLogica

In the MAC project our learning activities were designed to foster the development of students' mental models of the structures involved in transmission genetics and their role in supporting the inheritance of traits from one generation to the next. Using BioLogica™, students interacted with a multi-level hypermodel of genetics [see Figure 2; (Horwitz and Christie, 1999, 2000)]. BioLogica™ grew out of an earlier hypermodel called GenScope™ (Horwitz et al., 1996). It consists of a suite of models dealing with DNA, genes, chromosomes, germ cells, meiosis and fertilisation, organisms and traits, pedigrees, and populations. Each model offers a representation of its subject domain as well as specific affordances enabling students to effect relevant manipulations.

Figure 2 The linked, manipulable models at the core of the BioLogica hypermodel (see online version for colours)



In the BioLogica hypermodel, manipulations made at any level can affect any other level, much the way alterations in a single cell of a spreadsheet can 'ripple' through and affect other cells. Thus, the alteration of a single nucleotide at the DNA level may produce a mutation that causes an alteration of phenotype visible at the organism level. The mutation may be transmitted to a gamete during meiosis and may be expressed in an offspring through fertilisation. The statistical likelihood of the appearance of the altered phenotype can be studied at the pedigree level. If the mutation conveys a selective advantage, the frequency of the mutated allele will increase at the population level.

In the MAC project, we embedded BioLogica within a scripting environment called Pedagogica™, which supported the scaffolding and logging of students' actions (Horwitz and Burke, 2002; Horwitz et al., 2008). Pedagogica provides an authoring tool for creating the scripts, using the Javascript language, and also offers runtime support, including a communication channel between the script and the underlying model. This enables the script to react to runtime events, such as a student running fertilisation to create a new organism. The script can then examine the genotype of the new organism and react accordingly.

Using this scripting environment, we developed 12 learning activities for this domain, building on the earlier work of the NSF-funded GenScope [NSF#9725524, Horwitz Principal Investigator (PI)] and BioLogica [NSF#0087579; Horwitz and Gobert, (PIs)] (Horwitz and Barowy, 1994; Horwitz and Christie, 1999, 2000; Horwitz et al., 2008, 1996; Buckley et al., 2004, 2006).

4.1 Scaffolding students interactions with BioLogica's models

Research has shown that students have difficulty interpreting and reasoning with external models and representations and that scaffolding is needed to support students' learning (e.g., Gobert et al., 2004; Gobert and Clement, 1999; Kindfield, 1993, 1994; Larkin, 1989; Larkin and Simon, 1987; Lowe, 1993). This is consistent with a study of the GenScope project in which students who used worksheets to scaffold their interactions with the GenScope model outperformed students whose interactions were not scaffolded (Hickey et al., 2003). In addition to generic scaffolding such as pre- and post-organisers, orienting tasks, and glossary items, MAC activities scaffold learners' model-based learning of the domain by supporting learning in the following ways:

- *Representational* scaffolds focus students' attention on the perceptual cues of the representations and make links with other representations such as terminology.
- *Model components acquisition* scaffolds support students' acquisition knowledge about one or more aspects of the phenomenon (e.g., spatial, causal, functional, temporal).
- *Model components integration* scaffolds help students combine model components in order to come to a deeper understanding of how they work together.
- *Model based reasoning* scaffolds support students in reasoning with their models.
- *Reconstruct, reify and reflect* scaffolds require students to use what they have learned and apply it to another context or task (Gobert et al., 2004).

Scaffolding of each type was implemented in the form of questions, tasks, or explanations that focused on aspects of model-based learning. Scaffolding questions and tasks were transformed into assessments by taking advantage of Pedagogica's ™ data capture capabilities (Horwitz and Burke, 2002) and customisable feedback to students based on the data captured.

4.2 Formative assessments

The PADI project (Mislevy et al., 2002) aided our conceptualisation of the assessment tasks with its focus on the student model (knowledge, skills, and abilities to be assessed), task model (tasks to elicit behaviour from which we can infer the state of a student's model) and evidence model (data that provides evidence from which we can infer the state of a student's model). In MAC this led us to pose the following questions:

- What mental models and inquiry skills do we want students to develop?
- What mental models and inquiry skills do we want to be able to assess and provide feedback on?
- What tasks would engage students in progressive model-building and provide data?
- What data would provide evidence from which we could infer the state of students' models and inquiry skills?

Our research also built on the analysis conducted when Kindfield and Hickey created the pre- and post-tests that measured learning gains in the GenScope project (Hickey et al., 2003, 1998; Kindfield et al., 1999). We used two of their reasoning dimensions: one that distinguished within and between generations (what we would consider reasoning *with* models of meiosis and fertilisation vs. models of inheritance), and a second that distinguished between reasoning from cause to effect vs. effect to cause, which in turn builds on the work of Stewart et al. (Stewart and Hafner, 1991; Stewart et al., 1992). We found the second dimension particularly helpful in analysing the tasks presented to students in the MAC learning activities.

We use three tasks from BioLogica's Monohybrid activity to illustrate. *Monohybrid* is a central instructional activity in the biology sequence. It is the fourth activity out of 12 and was designed to help students integrate their models of meiosis and fertilisation (developed in the first two activities and assessed in activity three) into a model of inheritance. It introduces students to the pedigree level and the interactive Punnett square – features of BioLogica that combine visual representations with powerful domain-specific tools that support reasoning and inference-making. The activity concludes with four tasks that develop and assess students' skills at using these tools as well as their mental models of inheritance. We discuss the last three of these tasks in detail below.

Table 1 Comparison of exemplars from BioLogica

Task name	*Task description*	*Purpose*	*Inquiry skills*	*Reasoning required*
Task 2 – Predict	Predict offspring of two-legged dragons and test prediction.	Diagnose common misconception.	Using Punnett square and pedigree	Cause to effect, two generations
Task 3 – Produce	Modify parental genotypes such that all offspring have two legs.	Assess model of meiosis, fertilisation, inheritance.	Using Punnett square and pedigree	Effect to cause, two generations
Task 4 – Skip	Determine parental genotypes that result in traits appearing to skip a generation.	Assess model of meiosis, fertilisation, inheritance	Design experiment, interpret data, conduct experiment.	Effect to cause, three generations

As shown in Table 1, the three structured and scaffolded instructional tasks focus on the same content and use the same tools, but they require and elicit different reasoning. Task 2 – Predict guides students' investigation of the distribution of traits among the offspring of two-legged parents. It requires that students reason from cause to effect using their models of inheritance of the Legs characteristic in the model dragon species. To scaffold their reasoning, BioLogica steps them through creating and using a Punnett square. In contrast, Task 3 – Produce requires students to reason from effect to cause by asking, what parental genotypes would result in all the offspring having two legs? Students must set the genotypes of the parents, breed them, and check the result. Task 4 – Skip is an unscaffolded transfer task that asks the students to demonstrate the genetic mechanism that causes traits to appear to skip a generation. To do this, students must reason in both directions over three generations. All of these monohybrid tasks have fixed initial states and correct answers, which makes it straightforward to monitor students' inquiry processes.

4.2.1 Task interface and data capture

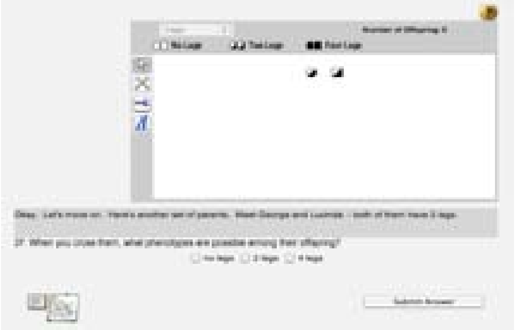
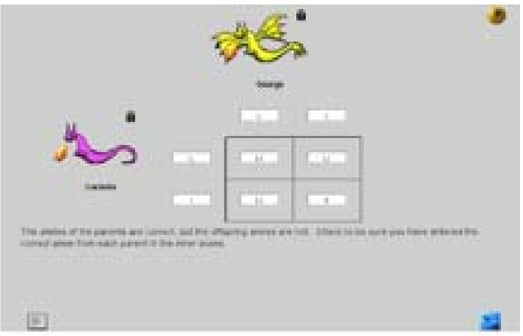
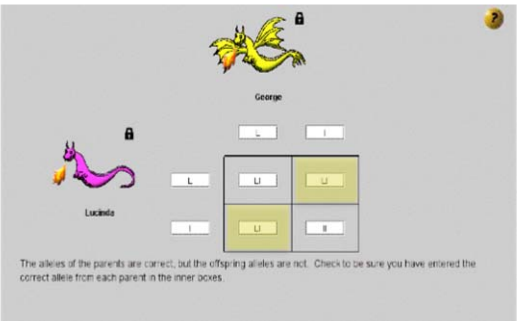
To illustrate the task interface and the data collected when students are working on monohybrid learning activities, we describe Task 2 – Predict.

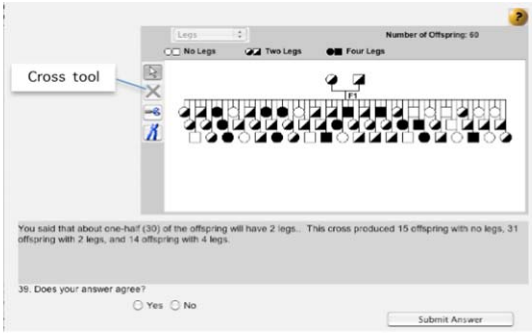
In Figure 3, the sequence of screen shots shows the steps students must take to check their prediction. The first screen presents the task; the pedigree tools have been disabled and are not yet usable by the student in this stage of the task. After they make their prediction, students must complete the Punnett square. They can, but are not prompted to, use the Chromosome tool to examine the chromosomes of the parents, then fill in the alleles of the parents and the offspring. BioLogica checks to see if they did so correctly and provides appropriate feedback. For instance, if they filled in the parental alleles incorrectly but did not examine the parents' chromosomes, the computer would suggest that they check the chromosomes and try again. Students were allowed three attempts before BioLogica presented them with a correctly completed Punnett square. Once they had passed that stage, either on their own or with assistance, they were asked to select those cells in the Punnett square that represent 2-legged offspring. They were allowed an unlimited number of attempts with feedback until they did this correctly. They were then asked to predict, based on their observation of the Punnett square with some cells highlighted, what fraction of the offspring would have two legs. Finally, students were asked to breed the parents using the Cross tool and to check the result against their predictions.

As described above, we were able to log students' actions as they worked on inquiry tasks of this kind. That function, though non-trivial from a technology standpoint, was straightforward to implement. The hard part was the analysis of the data collected. In the next section we describe what went into the log files and how we used that data to make inferences concerning students' inquiry skills and understanding of content, and to provide appropriate feedback to them.

The MAC activities were created with objects that automatically logged the same types of data each time they were used. For example, each time a student uses the Cross tool that action is logged. The log includes the time when the cross occurred as well as the genotype and generation number of each parent and the number of offspring produced. Data of this kind can be used to trigger feedback to the student as described above (e.g., for students who did not use the chromosome tool before completing their Punnett square). Of greater import for this paper, the structured log files that were generated by these self-logging objects captured data to be used for offline analysis.

Figure 3 Sequence of steps in Task 2 – Predict (see online version for colours)

<table>
<tr>
<td>

</td>
<td>The first screen presents the task; the tools of the pedigree are not usable yet. Students make a prediction by answering a survey question. The correct answer is all three.</td>
</tr>
<tr>
<td>

</td>
<td>After students make their prediction, they must complete the Punnett square. They may examine the chromosomes of the parents if they choose, then fill in the alleles of the parents and the offspring. Students are given three attempts before BioLogica presents them with a correctly completed Punnett square.</td>
</tr>
<tr>
<td>


The alleles of the parents are correct, but the offspring alleles are not. Check to be sure you have entered the correct allele from each parent in the inner boxes.
</td>
<td>Students select cells corresponding to 2-legged offspring. They are given unlimited attempts with feedback until they obtain the correct answer. Students are then asked to predict what percentage of the offspring will have two legs based on the Punnett square.</td>
</tr>
<tr>
<td>


You said that about one-half (30) of the offspring will have 2 legs.. This cross produced 15 offspring with no legs, 31 offspring with 2 legs, and 14 offspring with 4 legs.

39. Does your answer agree?

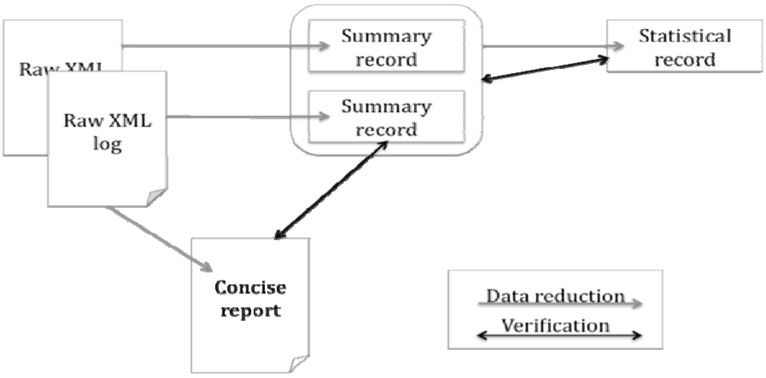
○ Yes ○ No
</td>
<td>Finally, students breed the parents using the cross tool of the pedigree and check their results against their predictions.</td>
</tr>
</table>

In order to provide appropriate feedback to students and to conduct MAC research we had to be sure that the data from which we were generating inferences and conclusions were accurate traces of students' actions. We conducted a series of verification and reduction steps, beginning with action-by-action comparison of log files to video of computer screens taped as students used learning activities (Buckley et al., 2004). When omissions or duplications were discovered, we revised the relevant script and tested it again. Because we were unsure what data would be needed, we erred on the side of capturing as much data as possible.

After we were certain that the log files accurately captured student actions, including their answers to embedded questions, we began the process of reducing them to forms and formats from which we could develop algorithms for analysing student performance. Each student session generated hundreds of pages of raw log files, which would have been intractable were it not for the XML tags used to structure the output. MAC's data-mining tools enabled us to extract from the log files database those that were relevant to a particular class of students, a particular activity, or a particular student and then generate a variety of reports from any given log file at different levels of detail.

Figure 4 diagrams the flow of data from raw log files to statistical records.

Figure 4 Overview of data processing during verification and reduction



The reduction from raw XML to concise reports preserved all the detail existing in the raw log, but formatted it for human examination. Table 2 provides an example of each.

Concise reports provide a time-stamped record of a student's actions and the data associated with those actions. As shown in Table 2, at 13:23:10 while observing the wings pedigree, this student crossed two dragons producing 40 offspring. We can read the genotypes of the parents for all traits and observe that they are the first pair of dragons (generation 0) for this pedigree. With the concise report we can determine what the students did before and after this action and how they answered questions.

If our objectives had been restricted to using this technology in case study research we could have stopped with the concise report. However for scalability and statistical analyses, we needed to summarise each student's performance in a format that could be examined in spreadsheet form and imported into a statistical analysis program. Thus, all subsequent data reduction involved data analysis as well. This required iterative cycles of human coding and analysis, creation of algorithms for machine coding, and verification of machine coding, followed again by human analysis (Buckley et al., 2006). We provide some detail about these iterations below.

Table 2 Comparison of raw XML log and concise report formats

Format and features	*Example of data for one cross*

Raw XML log
• Hundreds or pages
• Not easily read by humans
	Characteristic is being observed: trait: wings
	    </message>
	  </action>
	  <action priority="middle">
	    <date> 2005.02.15.13.23.10&nbsp; 02/15/05 | 13:23:10 </date>
	    <message>
	      Genotype of mother: Hh, SS, ww, Ll, Tt, pp, Ff, Aa, BB
	    </message>
	  </action>
	  <action priority="middle">
	    <date> 2005.02.15.13.23.10 &nbsp; 02/15/05 | 13:23:10 </date>
	    <message>
	      Genotype of father: Hh, SS, WW, Ll, Tt, p, F, a, B
	    </message>
	  </action>
	  <action priority="middle">
	    <date> 2005.02.15.13.23.10 &nbsp; 02/15/05 | 13:23:10 </date>
	    <message>
	      Generation of mother: 0
	    </message>
	  </action>
	  <action priority="middle">
	    <date> 2005.02.15.13.23.10 &nbsp; 02/15/05 | 13:23:10 </date>
	    <message>
	      Generation of father: 0
	    </message>
	  </action>
	  <action priority="middle">
	    <date> 2005.02.15.13.23.10 &nbsp; 02/15/05 | 13:23:10 </date>
	    <message>
	      number of offspring: 40
	    </message>
	  </action>

Concise report • Chronological record of student actions and answers • Readable by humans	

Elapsed time	*Interval (sec)*	*Action*	*Trait.node question ID*	*Mother's genotype/ student response*	*Generation/ score*	*Father's genotype/ student response*	*Generation/ score*	*# offspring*
13:23:10		Cross	wings	HH, SS,ww, Ll,Tt, pp, Ff, Aa, BB	0	HH, SS, WW, Lt, Tt, p, F, a, B	0	40

A group of five researchers began by individually reading a complete set of log files from one randomly-selected student to identify those actions critical to successful completion of various tasks. From this, we developed the initial specifications for data extraction, reduction, and further analysis. For each activity, specifications were created to transform the chronological concise report into a record that summarised the students' actions using variables such as the amount of time taken or the number of attempts needed to complete the task, whether the student was successful, what input variables, tools, or other resources (e.g., Punnett squares) the student used. To verify that the summary records were accurate, we compared them to their corresponding concise reports for a selected subset of records. This process often uncovered unexpected student actions not anticipated in our original specifications. In these cases we revised the summary report generator and tested it with a different set of log files. Specifications for producing the statistical record had to include algorithms or rubrics for characterising and evaluating student performances. These are described in more detail when we examine each task. Because some students required more than one class period to complete an activity, we also specified how to aggregate multiple summary records for a student into a single statistical record.

The summary and statistical record formats resulting from this process are described and illustrated in Table 3.

Table 3 Comparison of summary and statistical record formats

Summary record	*Statistical record*
• One record per log file	• One record per student per activity
• Includes autoscoring	• Aggregates student use of an activity
• Basis for teacher reports	• Concatenates answers and aggregates autoscoring
	Used for statistical analysis

Student ID	*Class ID*	*Date*	*Total duration (min)*	*T4 time*	*Success*	*Q42A*	*Tries*	*Crosses*	*T4 cat*	*F1 crosses*	*Cross*	*Chromo*	*Snip 1*
15021	5174	Tue Feb 15 12:52:32 CST 2005	34.2	11.3	1	I	3	WW x Ww, ww x Ww, ww x WW	8	1	4	11	12

As Table 3 shows, we captured how long a student was involved in a session and in a particular task (Task 4 in this case). In this example, Student 15021 successfully completed the task, but required three tries. Based on the crosses listed, the student's performance was assigned to category B[1] (successful in 2–3 tries with no repeated crosses). The student made one cross of the second generation (F1), used the cross tool four times, examined chromosomes 11 times, snipped 12 individuals and two families from the pedigree (to clear the way for another attempt), read the genome chart once for four seconds, did not use the Punnett square tool, and viewed the task description for a total of 95 seconds. If the student had worked on this task in more than one session, the data would have been summed in the case of times and tries, concatenated in the case of crosses, and characterised on the basis of all the data. Our goal was not only to characterise students' performances, but also to capture more data about their learning experiences and inquiry processes.

4.3 Analysis of student actions during problem-solving and inquiry

Autoscoring students' responses to multiple-choice questions was straightforward, once the correct answer was specified. On the other hand, analysing students' actions during problem-solving and inquiry was much more complex and revealed significant differences in inquiry strategies employed by different students. In our early observations of students and perusals of concise reports, we observed that some students were haphazard and unfocused as they tried to accomplish tasks; others were systematic and mindful (Gobert, 1994, 1999; Thorndyke and Stasz, 1980). For example, we found that some students bred the same two organisms over and over, evidently in the hope that they would eventually, by random chance, achieve the distribution of offspring traits they sought. These students appeared to have understood that there is a random component to each meiotic and fertilisation event, but had failed to understand the *statistically* predictable nature of characteristics of offspring produced by the same parents. Faced with the same task, other students, by their unprompted use of the chromosome tool, demonstrated that they were spontaneously reasoning at both phenotypic and genotypic levels.

We observed situations in which students succeeded at a task in one try, apparently because they reasoned with their mental models and did what needed to be done. On the other hand, if students demonstrated flawed mental models or poor reasoning abilities, we expected that the use of a systematic inquiry strategy might help them accomplish the task, whereas lack of systematicity was likely to hamper problem-solving and learning from inquiry. With this in mind, we set out to develop algorithms for detecting these different strategies.

Taking each monohybrid task in turn, we now describe in detail how we analysed the log file data to evaluate students' procedural, schematic, and strategic knowledge.

4.3.1 Task 2 – Predict: what legs phenotypes will the offspring of a pair of 2-legged parents have?

Recall that the Task 2 – Predict required students to predict the phenotypes of the offspring of a pair of 2-legged dragons, complete a Punnett square to help them reason about the offspring, select offspring genotypes that produce 2-legged dragons, and estimate the proportions of offspring before breeding the 2-legged dragons and checking the results against their predictions.

- Evidence. Table 4 summarises the data captured during Task 2 – Predict and the scoring rubric used to process the data.

Table 4 Task evidence for Task 2 – Predict

Task component	*Data captured*	*Scoring rubric*
Prediction of offspring phenotypes	Check boxes for no legs, 2 legs, 4 legs	1 point if all three boxes checked
Punnett square completion	Contents (alleles in each cell) Number of attempts (max = 3)	2 points if completed correctly the first time, 1 point if completed in 2–3 tries
Punnett square selection	Alleles of selected cells Number of attempts (unlimited)	2 points if correct on first attempt, 1 point if correct in 2–3 tries, 0 if more than 3 tries.
Estimated number of offspring with two legs	Multiple choice	1 point if correct

- Analysis. Task 2 – Predict provides fine-grained evidence of students' procedural knowledge as demonstrated by completing and interpreting the Punnett square. It also provides evidence of students' model of inheritance of the Legs characteristic (schematic knowledge) as demonstrated by the initial prediction and by the selection of the Punnett square cells that represent 2-legged offspring. Responses to the initial prediction also enabled us to identify the approximately 10% of the students who held the naïve conception that 2-legged parents produce only 2-legged offspring.
- Findings. As shown in Table 5, 78% of the students completed the Punnett square correctly the first time. If they hadn't done so, students were given a correct Punnett square and asked to select the squares with the correct allele combinations. 76% of the students succeeded the first time. There was no limit to tries, which frustrated some students, because they could not proceed to the probability question of approximately how many offspring will have two legs. When they finally succeeded in selecting the cells with genotypes for 2-legged offspring in the Punnett square, 88% were able to estimate the proportion of 2-legged offspring. Their scores for each step of the process were totalled to create the T2score (max = 6) for statistical comparisons, but the disaggregated scores for the task are part of the summary and statistical records and could be used to provide diagnostic information to teachers and students. The distribution of scores is shown in Table 5.

Table 5 Distribution of student scores for each step of task (280 students, 2005–2006 data)

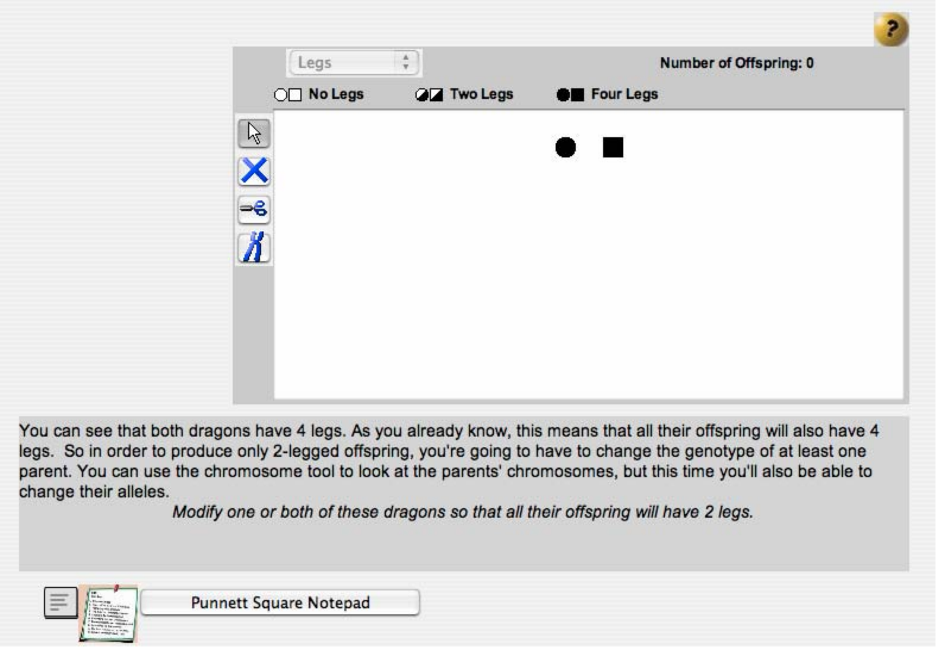
Score	*Prediction*	*Complete Punnett square*	*Select Punnett square cells*	*Estimate offspring*
2	n/a	219	214	n/a
1	145	17	44	247
0	135	44	22	33
% students correct first time	0.52	0.78	0.76	0.88

We infer from these results that while roughly half of the students had inadequate models of Legs inheritance as shown by their predictions, most of the students had mastered the procedural skills for using Punnett squares to predict probable offspring and demonstrated that they could reason from cause to effect when scaffolded. Reasoning from effect to cause is a more challenging problem (Stewart and Hafner, 1991, 1994) as we will see in the following tasks.

4.3.2 Task 3 – Produce: can a pair of parents have only 2-legged offspring?

Before we give students access to any pedigree tools shown in Figure 5, we ask, 'Is it possible for 2 parents to produce only 2-legged offspring?' and 'What would their genotypes have to be?' We give them a pair of 4-legged dragons and ask them to produce only 2-legged offspring.

Figure 5 Task 3 – Produce interface (see online version for colours)



You can see that both dragons have 4 legs. As you already know, this means that all their offspring will also have 4 legs. So in order to produce only 2-legged offspring, you're going to have to change the genotype of at least one parent. You can use the chromosome tool to look at the parents' chromosomes, but this time you'll also be able to change their alleles.

Modify one or both of these dragons so that all their offspring will have 2 legs.

- Evidence. Table 6 summarises the evidence provided by Task 3 – Produce, organised by tool use. It shows the data captured for each tool use and the variables calculated from tool usage for the entire task. Success is indicated by the correct parental cross (LL x ll).
- Analysis. We expect this task to be more difficult than Task 2 – Predict because it requires students to reason from effect to cause. It requires that they use the chromosome tool to change the alleles of one parent to recessive 'l' alleles, then breed them using cross tool. Students also have access to the Punnett square pad and the dragon genome chart at bottom left. Knowing how to use the tools demonstrates procedural knowledge, while reasoning with model of legs inheritance to determine what the parental alleles should be is evidence of schematic knowledge. This task

may also elicit strategic knowledge if the student is not successful on the first attempt. The student must then evaluate the data and determine what is needed to succeed.

- Findings. Since Task 3 – Produce and Task 4 – Skip employ similar analyses, we will present Task 3 – Produce findings with those of Task 4 – Skip.

Table 6 Task evidence and processing for Task 3 – Produce

Time-stamped tool usage	*Data captured*	*Variables calculated for the summary record*
Cross tool use	Genotype, gender, generation of parents, number of offspring	List of all crosses made How many crosses made How many repeated crosses Successful cross (LL x ll)
Chromosome tool use	Genotype, gender, generation of inspected dragon	Times used
Punnett square use	Contents of Punnett square	Times used
Dragon genome chart	Time between open and close	Total time open

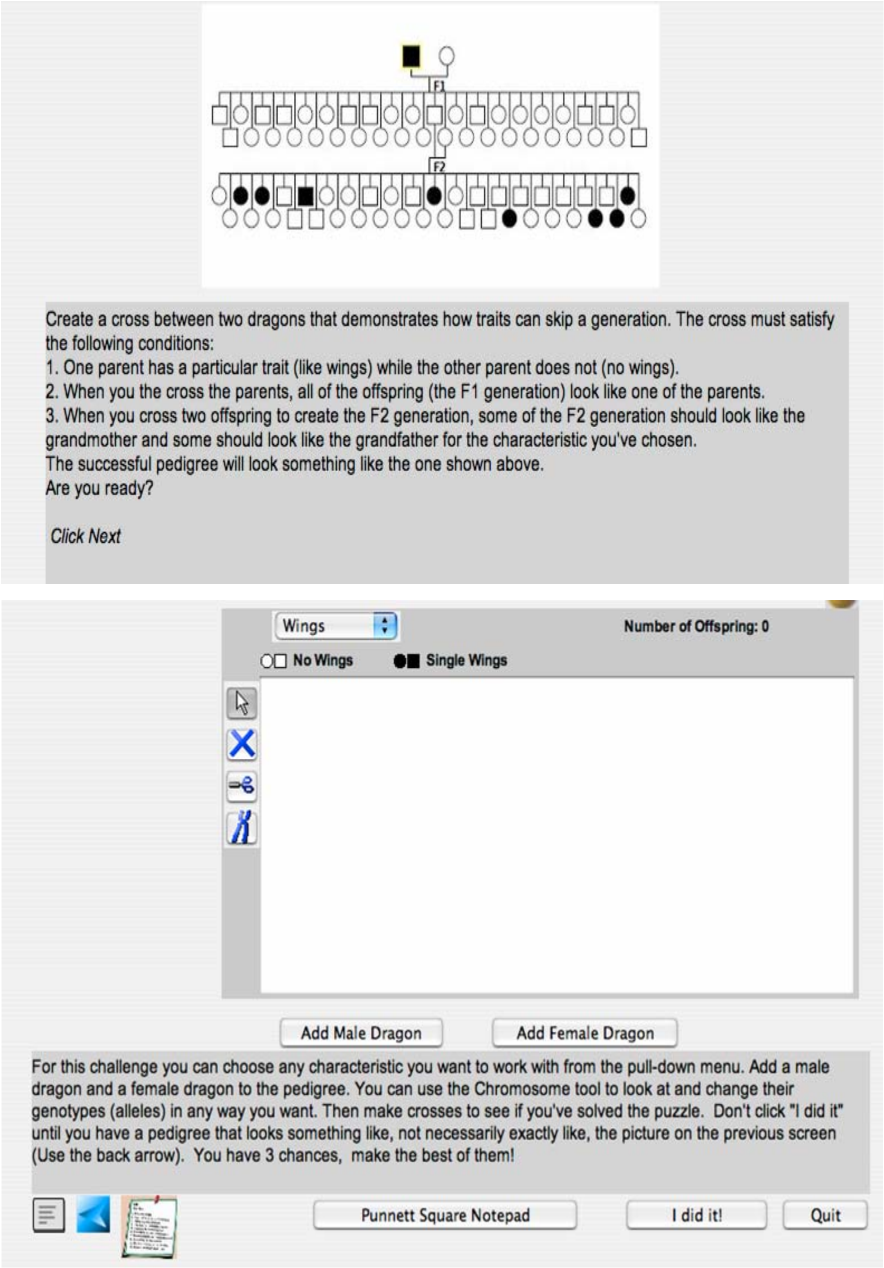
4.3.3 *Task 4 – Skip: create a 3-generation pedigree that demonstrates why traits appear to skip a generation*

As shown in Figure 6, we ask students to create a cross between two dragons that demonstrates how traits appear to skip a generation. Students must add dragons, select a characteristic (horn, wings, or tail) for which the inheritance pattern is simple dominance (as opposed to legs, which are inherited as an incompletely dominant characteristic), change the alleles so that one parent is homozygous dominant and the other is homozygous recessive, cross them and cross a pair of the F1 generation. We then ask them to explain how they did it.

- Evidence. Log files captured the same kinds of data shown in Table 6 for Task 3 – Produce. In the summary record for Task 4 – Skip, success is indicated if the student made the correct parental cross (DD x rr) followed by a cross of any two offspring.

- Analysis. The task requires model-based reasoning between generations in both forward and backward directions. Two heterozygous parents (cause) will express the dominant trait but will produce offspring with both traits (effect). To get two heterozygous parents (effect), the original pair must be homozygous; one with the dominant trait, the other with the recessive trait (cause). Students must extend their reasoning to encompass a third generation to successfully accomplish Task 4. It is also a transfer task in two respects. Because it involves traits that are examples of simple dominance models of inheritance, it requires transfer to a different model of inheritance. The first step in accomplishing the task is identical to Task 3's successful solution, so students have an opportunity to transfer procedural knowledge. If this step is completed successfully, then the third generation automatically follows when any two offspring are crossed. As in Task 3 – Produce, using the tools is evidence of procedural knowledge while reasoning with one' model is evidence of schematic knowledge. This task may also elicit strategic knowledge if the student is not successful on the first attempt.

Figure 6 Task 4 – Skip interface screens (see online version for colours)



Create a cross between two dragons that demonstrates how traits can skip a generation. The cross must satisfy the following conditions:
1. One parent has a particular trait (like wings) while the other parent does not (no wings).
2. When you the cross the parents, all of the offspring (the F1 generation) look like one of the parents.
3. When you cross two offspring to create the F2 generation, some of the F2 generation should look like the grandmother and some should look like the grandfather for the characteristic you've chosen.
The successful pedigree will look something like the one shown above.
Are you ready?

Click Next

For this challenge you can choose any characteristic you want to work with from the pull-down menu. Add a male dragon and a female dragon to the pedigree. You can use the Chromosome tool to look at and change their genotypes (alleles) in any way you want. Then make crosses to see if you've solved the puzzle. Don't click "I did it" until you have a pedigree that looks something like, not necessarily exactly like, the picture on the previous screen (Use the back arrow). You have 3 chances, make the best of them!

4.3.4 Scoring Task 3 – Produce and Task 4 – Skip

Determining if a student succeeded was straightforward for Task 3 – Produce and Task 4 – Skip: if they made the correct cross or two, they must have changed the alleles appropriately. Characterising their inquiry skills is harder. After extensive examination of the concise reports for a variety of students and much debate, we identified a simple measure of systematicity – that is, no repeated crosses. Because the pedigree displays the cumulative results of all crosses, repeating a particular cross yields no additional useful data. Therefore, we deemed accomplishing the task without any repeated crosses systematic. We arrived at a 6-point ordinal scale to rank the students' performances on both tasks (see Table 7). At each end of the scale were the students who succeeded or failed with just one attempt. In between were the students who were either successful or not and systematic or not. Based on our hypothesis that students who were systematic were more likely to be reasoning with their models, we ranked students who were systematic higher than those who were haphazard and favoured success over failure.

4.3.5 Findings Task 3 – Produce and Task 4 – Skip

When we examine the distribution of students among the ordinal categories for Task 3 – Produce and Task 4 – Skip, we see, as predicted, that Task 4 – Skip is more difficult than Task 3 – Produce: the percentage of students who succeeded on the first try for Task 3 – Produce is over twice that for Task 4 – Skip. Notice also the reproducible results for 2005 and 2006.

Table 7 Criteria for auto-coding student performance on Task 3 – Produce and Task 4 – Skip and distribution of students

Code	*Success*	*Tries*	*Repeated crosses*	*Description*	*Task 3 (%)*		*Task 4 (%)*	
					2005	*2006*	*2005*	*2006*
6	1	= 1		Successful, 1st try	41	47	19	22
5	1	> 1	*no* repeated crosses	Systematic	29	29	28	29
4	1	> 1	repeated crosses	Haphazard	21	16	7	7
3	0	> 1	*no* repeated crosses	Systematic	2	1	13	12
2	0	> 1	repeated crosses	Haphazard	6	6	20	18
1	0	= 1		Unsuccessful, 1 try	< 1	1	13	11
				Total N	405	281	353	246

5 Monohybrid findings

Based on data from students who used the monohybrid activity in 2005–2006 and took both the pre- and post-tests, we found all three tasks were moderately and significantly correlated with total pre-test scores and with each other.

Table 8 Correlations among total pre-test scores and task scores

	Total score pre-test	*Task 2 – Predict*	*Task 3 – Produce*	*Task 4 – Skip*
Total score pre-test	1.00	-	-	-
Task 2 – Predict	0.29*	1.00	-	-
Task 3 – Produce	0.29*	0.56*	1.00	-
Task 4 – Skip	0.24*	0.34*	0.31*	1.00

Note: * Correlation is significant at the 0.01 level (2-tailed).

All three task scores were also significant predictors of students' post-test scores, holding pre-test scores constant. As the three tasks were correlated, to avoid colinearity among the measures, we ran individual regressions of post-test scores on task performance, holding pre-test scores constant. Pre-test scores and task scores were entered as separate blocks; block 1 included prior achievement only and block 2 included prior achievement and a single task score. This approach allowed us to examine the percentage of variance in the post-test scores explained by a single task score over and above the variance explained by prior achievement. The results of the regression analyses in Table 9 show that each task score is a significant predictor of the post-test scores after controlling for prior knowledge.

Table 9 Summary of individual regressions of task performance on post-test, holding pre-test constant

	Total post-test acores			
	β	*t*	*Sig.*	*Total R^2*
Model 1				
Block 1 – prior achievement only				
Total pre-test scores	0.47	9.43	< .001	21.90%
Block 2 – prior achievement and task scores				
Total pre-test scores	0.36	7.60	< .001	
Task 2 – Predict scores	0.39	8.13	< .001	35.30%
Model 2				
Block 1 – prior achievement only				
Total pre-test scores	0.47	9.50	< .001	22.10%
Block 2 – prior achievement and task scores				
Total pre-test scores	0.38	7.72	< .001	
Task 3 – Produce scores	0.33	6.89	< .001	32.10%
Model 3				
Block 1 – prior achievement only				
Total pre-test scores	0.47	8.86	< .001	21.80%
Block 2 – prior achievement and task scores				
Total pre-test scores	0.41	7.81	< .001	
Task 4 – Skip scores	0.24	4.53	< .001	27.00%

Task 2 – Predict explained the greatest variance in the post-test scores after controlling for pre-test scores (13.40%). Task 3 – Produce explained 10% of the variance in the post-test scores after controlling for pre-test scores. Task 4 – Skip explained only 5.2% of the variance in the post-test scores after controlling for pre-test scores.

The results of the regression models suggest that being able to use the Punnett square to support reasoning and learning may be essential to the development, not only of students' procedural knowledge, but also to their schematic knowledge and mental models. This is consistent with Kindfield's belief that students' understanding of genetics and of domain-specific representations co-evolve during learning (Kindfield, 1993, 1994). It also supports our belief that the Punnett square is a powerful tool and representation for helping students learn models of inheritance through model-based reasoning.

6 Discussion

Over the five years of the MAC project we developed a technology platform, curricular and assessment materials, and a reporting system that enabled us to monitor, assess, and respond to students' actions. This paper has focused on just three out of the many tasks that comprise the learning activities of the biology strand of the project. In it, we have demonstrated how we were able to assess students' inquiry skills and problem-solving approaches as well as their naïve conceptions and mental models of monohybrid inheritance.

Assessments in the MAC project served three purposes:

1 to provide data concerning overall learning gains

2 to guide immediate feedback to students

3 to provide fine-grained analysis of students' reasoning and inquiry skills in order to better understand how inquiry supports learning.

This paper focuses primarily on how data that is both fine-grained and large-scale can be analysed in order to get rich evidence of students' reasoning and inquiry skills in this domain.

It has become routine to warn inexperienced researchers not to attempt to collect too much data for fear that the analysis will prove time-consuming and ultimately ineffective. The much dreaded 'dribble file' that reports every action of the student is often sited as a trap for the unwary. The MAC data came close to being just such a dribble file, yet with an appropriate mix of human ingenuity, hard work, and the use of sophisticated tools for data-mining, we were able to uncover a wealth of useful information from it. This paper has described the methods that enabled us to process approximately 1.5 gigabytes of extremely complex data that include performance parameters, embedded assessments, surveys, and pre- and post-tests. The log files captured data that documented how nearly 12,000 students in over 50 schools, taught by 127 teachers, used MAC learning activities in three domains over a period of three years.

We have described the affordances of three BioLogica tasks for assessing students' understanding and reasoning. These tasks are located at a critical point in the progression of model-based learning activities used to teach genetics. In the course of determining what students learn with our activities we investigated novel ways of assessing their understanding and reasoning by capturing and analysing their actions as they solved problems with BioLogica hypermodels. Our findings have implications not only for formative, real-time classroom assessment of students' understanding, but also for large-scale, high-stakes testing.

We have demonstrated the ability not only to assess students' ability to complete inquiry tasks successfully, but also to identify non-productive strategies adopted by some students, as well as systematic inquiry and problem-solving approaches used by others. We have correlated student performances on these three tasks with learning gains.

This project and paper has described a rigorous method for collecting fine-grained data on a very large scale and analysing these data using computer algorithms. Our procedures for data reduction and methodology for analysis are informed by theoretically-motivated assessment frameworks (Shavelson et al., 2002; Mislevy, 2002). Thus, the work reported on occupies a fruitful middle ground, merging automated data acquisition and data-mining techniques suitable for very large scale use with theoretical frameworks for assessment based on much finer-grained experiments.

6.1 Limitations

Our research methodology had unavoidable limitations. For instance, the variance in students' manipulation of our computer-based models may be explained both by a corresponding variance in their mental models and inquiry skills and by differences in their familiarity with the software itself. We attempted to minimise this effect by providing, particularly in the early activities, scaffolding designed to introduce students to the software tools. Nevertheless we cannot be certain that all of our students became equally adept at using the software. Moreover, many of our participating teachers encouraged cooperative learning styles in their classrooms, which resulted in substantial collaboration between students as they worked on our learning activities. Our instructions to the teachers were not to allow such 'cross talk' on the pre- and post-tests, which therefore can be taken as reflecting the work of individual students, but the collaborative atmosphere of the classroom necessarily introduces some uncertainty into our analysis of students' manipulations of the computer model. For example, did students succeed on their first attempt because their neighbour told them what to do?

6.2 Using log files for formative assessment

There are two aspects to this topic: how to create useful log files and how to use the reports that can be generated from them to guide curricular decisions in the classroom. Theory-driven assessments are important to creating useful log files. We have explored three tasks that assess different aspects of model-based learning and inquiry, different mixtures of knowledge types, and different kinds of schematic and strategic reasoning. The tasks must be matched to the knowledge about the student one wishes to acquire and the nature of the evidence that will support sustainable inferences. It many sometimes prove necessary, for example, to constrain what students can do to accomplish a task, as we discovered when we observed several students who tried to create the desired pedigree of Task 4 – Skip by carefully snipping out all the offspring that had the undesired traits!

What data you decide to capture is crucial to gathering the evidence you need. You really don't need every keystroke or mouse click, but if in doubt, don't leave it out. It's relatively easy to program the computer to ignore certain data, but impossible to fill it in if it was never collected. We have also found that time on task, although sometimes unreliable for a variety of reasons, can help to distinguish between productive and unproductive behaviour.

Our model-based assessments provide formative data that, if reported in a timely fashion, could be used by teachers to inform curricular or instructional decisions. On the MAC project we were unable to take full advantage of this feature because we did not yet know how to analyse the data; we didn't learn that until the research phase of the project was completed[2]. Nevertheless, our research clearly points the way toward the creation of useful formative assessments based on real-time analysis of students' actions during problem-solving and inquiry activities. Such assessments can be useful both as short-cycle formative assessments, used to mediate students' interactions with the computer, and as input to classroom teachers both during and after class. Our research contributes to the knowledge base necessary to develop such formative assessments. Indeed, the [Logging Opportunities in Online Programs for Science (LOOPS), NSF # 0903243] and Calipers projects [NSF # 0454772 and # 0741729; Quellmalz, (PI)] among others are already building on our research. Eventually, one can hope, even the large-scale, high-stake assessments that dominate so much of the education enterprise may come to rely on the same interactive approach (Quellmalz and Pelligrino, 2009).

Acknowledgements

The Modeling Across the Curriculum project was funded by the National Science Foundation, the US Dept. of Education and the National Institutes of Health under a grant awarded to the Concord Consortium (IERI #0115699). Any opinions, findings, and conclusions expressed are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Ayala, C.A., Shavelson, R.J., Yin, Y. and Shultz, S. (2002) 'Reasoning dimensions underlying science achievement: the case of performance assessment, *Educational Assessment,* Vol. 8, No. 2, pp.101–121.

Brewer, W.F. (1987) 'Schemas versus mental models in human memory', in P. Morris (Ed.): *Modelling Cognition,* pp.187–197, John Wiley and Sons, Chicester.

Buckley, B.C. (1992) 'Multimedia, misconceptions and working models of biological phenomena: learning about the circulatory system', Unpublished doctoral dissertation, Stanford University.

Buckley, B.C. (2000) 'Interactive multimedia and model-based learning in biology', *International Journal of Science Education,* Vol. 22, No. 9, pp.895–935.

Buckley, B.C. and Boulter, C.J. (2000) 'Investigating the role of representations and expressed models in building mental models', in J.K. Gilbert and C.J. Boulter (Eds.): *Developing Models in Science Education,* pp.105–122, Kluwer, Dordrecht, Holland.

Buckley, B.C., Gobert, J.D. and Christie, M.T. (2002) 'Model-based teaching and learning with hypermodels: what do they learn? How do they learn? How do we know?', *Paper presented at the American Educational Research Association,* New Orleans.

Buckley, B.C., Gobert, J.D. and Horwitz, P. (2006) 'Using log files to track students' model-based inquiry', *Paper presented at the 7th International Conference of the Learning Sciences,* Bloomington, IN.

Buckley, B.C., Gobert, J., Kindfield, A.C.H., Horwitz, P., Tinker, B., Gerlits, B., et al. (2004) 'Model-based teaching and learning with hypermodels: what do they learn? How do they learn? How do we know?', *Journal of Science, Education and Technology,* Vol. 13, No. 1, pp.23–41.

Clement, J. (1989) 'Learning via model construction and criticism: protocol evidence on sources of creativity in science', in J.A. Glover, R.R. Ronning and C.R. Reynolds (Eds.): *Handbook of Creativity: Assessment, Theory and Research,* pp.341–381, Plenum Press, New York.

Fadel, C., Honey, M. and Pasnick, S. (2007) 'Assessment in the age of innovation', *Education Week,* Vol. 26, No. 3, pp.4–40.

Gentner, D. and Stevens, A.L. (Eds.) (1983) *Mental Models,* Lawrence Erlbaum Associates, Hillsdale, NJ.

Gobert, J. (1994) 'Expertise in the comprehension of architectural plans: contribution of representation and domain knowledge', Unpublished doctoral dissertation*,* University of Toronto, Canada.

Gobert, J. (1999) 'Expertise in the comprehension of architectural plans: contribution of representation and domain knowledge', in J.S. Gero and B. Tversky (Eds.): in *Visual and Spatial Reasoning in Design '99,* Key Centre of Design Computing and Cognition, University of Sydney, AU.

Gobert, J. (2008) 'The affordances of model-based learning: conceptual, phenomenological, ontological and epistemological considerations', *Discussant on Designing and Assessing Modeling and Visualisation Technology-Enhanced Learning,* B. Zhang, (Organiser), *International Conference of the Learning Sciences*, June 24–28, Utrecht, The Netherlands.

Gobert, J.D. and Buckley, B.C. (2000) 'Introduction to model-based teaching and learning in science education', *International Journal of Science Education,* Vol. 22, No. 9, pp.891–894.

Gobert, J.D. and Clement, J.J. (1999) 'Effects of student-generated diagrams versus student-generated summaries on conceptual understanding of causal and dynamic knowledge in plate tectonics', *Journal of Research in Science Teaching,* Vol. 36, No. 1, pp.39–53.

Gobert, J. and Clement, J. (1994) 'Promoting causal model construction in science through student-generated diagrams', *Paper presented at the Annual Meeting of the American Educational Research Association,* New Orleans.

Gobert, J., Buckley, B.C. and Horwitz, P. (2007a) 'Through the looking glass and what we found there: a logging infrastructure for investigating students' inquiry processes', *Paper presented at the Annual Meeting of the American Educational Research Association,* April 9–13, Chicago, IL.

Gobert, J., Buckley, B. and Clarke, J.E. (2004) 'Scaffolding model-based reasoning: representations, cognitive affordances and learning outcomes', *Paper presented at the American Educational Research Association,* San Diego, CA.

Gobert, J., Heffernan, N., Ruis, C. and Kim, R. (2007b) *AMI: ASSISTments Meets Inquiry,* Proposal funded by the National Science Foundation (NSF-DRL# 0733286).

Hickey, D.T., Kindfield, A.C.H., Horwitz, P. and Christie, M. (2003) 'Assessment-oriented scaffolding of student and teacher performance in a technology-supported genetics environment', *American Educational Research Journal,* Vol. 40, No. 2, pp.495–538.

Hickey, D.T., Wolfe, E.W. and Kindfield, A.C.H. (1998) 'Assessing learning in a technology-supported genetics environment: evidential and consequential validity issues', *Paper presented at the Annual Meeting of the American Educational Research Association,* April, San Diego.

Horwitz, P. and Gobert, D. (2000) *Fostering Transfer from Open-Ended Exploration to Scientific Reasoning,* funded by the National Science Foundation (NSF-REC# 0087579).

Horwitz, P. and Barowy, W. (1994) 'Designing and using open-ended software to promote conceptual change', *Journal of Science Education and Technology,* Vol. 3, No. 3, pp.161–185.

Horwitz, P. and Burke, E.J. (2002) 'Technological advances in the development of the hypermodel', *Paper presented at the American Educational Research Association,* New Orleans.

Horwitz, P. and Christie, M. (1999) 'Hypermodels: embedding curriculum and assessment in computer-based manipulatives', *Journal of Education,* Vol. 181, No. 2, pp.1–23.

Horwitz, P. and Christie, M. (2000) 'Computer-based manipulatives for teaching scientific reasoning: an example', in M.J. Jacobson and R.B. Kozma (Eds.): *Innovations in Science and Mathematics Education: Advanced Designs for Technologies of Learning,* pp.163–191, Lawrence Erlbaum and Associates, Hillsdale, NJ.

Horwitz, P., Gobert, J. and Buckley, B.C. (2008) 'From computer-based manipulatives to hypermodels', in M.J. Jacobson (Ed.): *Designs for Learning Environments of the Future: International Learning Sciences Theory and Research Perspectives,* Manuscript under contract, Springer.

Horwitz, P., Neumann, E. and Schwartz, J. (1996) 'Teaching science at multiple levels: the GenScope program', *Communications of the ACM,* Vol. 39, No. 8.

Johnson-Laird, P.N. (1983) *Mental Models*, Harvard University Press, Cambridge, MA.

Kindfield, A.C.H. (1993/1994) 'Biology diagrams: tools to think with', *Journal of the Learning Sciences,* Vol. 3, No. 1, pp.1–36.

Kindfield, A.C.H., Hickey, D.T. and Yessis, L.M. (1999) 'Assessing student understanding of genetics: the NewWorm(c) assessment', *Paper presented at the Annual Meeting of the National Association for Research in Science Teaching,* Boston, MA.

Larkin, J.H. (1989) 'Display-based problem solving', in D. Klarh and K. Kotovsky (Eds.): *Complex Information Processing: The impact of Herbert A. Simon,* pp.319–341, Lawrence Erlbaum Associates, Hillsdale, NJ.

Larkin, J.H. and Simon, H.A. (1987) 'Why a diagram is (sometimes) worth ten thousand words', *Cognitive Science,* Vol. 11, pp.65–99.

Lowe, R. (1993) *Successful Instructional Diagrams*, Kogan Page, London.

Mislevy, R.J., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., Haertel, G. et al. (2002) *Design Patterns for Assessing Science Inquiry,* Unpublished manuscript, Washington, DC.

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

National Research Council (1999) *How People Learn: Brain, Mind, Experience, and School*, J.D. Bransford, A.L. Brown and R.R. Cocking (Eds.), Committee on Developments in the Science of Learning, National Academy Press, Washington, DC.

National Research Council (2000) *Inquiry and the National Science Education Standards*, National Academy Press, Washington, DC.

National Research Council (2002a) *Knowing What Students Know: The Science and Design of Educational Assessment,* National Academy Press, Washington, DC.

National Research Council (2002b) *Technology and Assessment: Thinking Ahead: Proceedings of a Workshop,* National Academy Press, Washington, DC.

Perkins, D.N. (1986) *Knowledge as Design,* Lawrence Erlbaum Associates, Hillsdale, NJ.

Quellmalz, E.S. and Pelligrino, J.W. (2009) 'Technology and testing', *Science,* 2 January, Vol. 323, pp.75–79.

Rouse, W.B. and Morris, N.M. (1986) 'On looking into the black box: prospects and limits in the search for mental models', *Psychological Bulletin,* Vol. 100, No. 3, pp.349–363.

Shavelson, R.J., Li, M., Ruis-Primo, M.A. and Ayala, C.C. (2002) 'Evaluating new approaches to assessing learning', *Paper presented at the Keynote Address: Joint Northumbria/EARLI Assessment Conference,* University of Northumbria at Newcastle, Longhirst Campus, UK.

Stewart, J. and Hafner, B. (1991) 'Extending the conception of problem solving', *Science Education,* Vol. 75, No. 1, pp.105–120.

Stewart, J. and Hafner, R. (1994) 'Research on problem solving: genetics', in D. Gabel (Ed.): *Handbook of Research on Science Teaching and Learning*, pp.284–300, Macmillan, New York.

Stewart, J., Hafner, R., Johnson, S. and Finkel, E. (1992) 'Science as model building: computers and high-school genetics', *Educational Psychologist,* Vol. 27, No. 3, pp.317–336.

Thorndyke, P. and Stasz, C. (1980) 'Individual differences in procedures for knowledge acquisition from maps', *Cognitive Psychology,* Vol. 12, pp.137–175.

White, B.Y. and Frederiksen, J.R. (1990) 'Causal model progressions as a foundation for intelligent learning environments', *Artificial Intelligence,* Vol. 42, No. 1, pp.99–157.

Notes

[1] These categories later became an ordinal scale.

[2] The MAC teacher reports did include students' answers to explicit questions embedded in the learning activities. Teachers could also access the raw log files containing detailed, time-stamped information regarding students' actions. Missing, however, from the teacher reports were the automated analyses of the kind we have described in this paper, since the algorithms for producing such analyses was not developed in time.