Using Log Files To Track Students' Model-based Inquiry

Barbara C. Buckley, Janice D. Gobert, & Paul Horwitz,
The Concord Consortium, 10 Concord Crossing Suite 300, Concord, MA 01742
bbuckley@concord.org

Abstract: We report on the Modeling Across the Curriculum Project (mac.concord.org), a 5-year study funded by the IERI program (IERI # 0115699). In this project we are using qualitative model-based tools for inquiry in three science domains (Genetics, Gas Laws, & Newtonian Mechanics) to promote scientific literacy on a broad scale, namely, content knowledge, process skills, and epistemologies of science (Perkins, 1986). The process skills we focus on in this project are model-based inquiry skills; epistemological constructs of interest here are students' epistemologies of models as a subset of epistemology of science (Gobert & Discenna, 1997; Schwartz & White, 1999). Our technological infrastructure, Pedagogica ™ (Horwitz & Burke, 2002) logs all students' interactions with our hypermodels (Horwitz & Christie, 1999). We use students' log files as performance assessments of model-based inquiry on inquiry tasks called "hot spots" and track changes in inquiry skills over time both within and across domains.

Project Overview

The Modeling Across the Curriculum project is a scalability project for which we have developed a technology platform, a reporting system, and curricular materials. There are four levels of research being conducted. Level 1 is focused on improving the scaffolding design through individual interviews of students and teachers. Level 2 is focused on classroom-based studies to evaluate the impact of amount of scaffolding. Level 3 is a longitudinal study of our dependent variables (content, inquiry skills, attitudes towards science, and epistemology of models) with the same students across 3 years in all three domains in our Partner Schools. Level 4 addresses what supports are necessary to scale this to many more schools.

Our curricular activities present students with content using a progressive model-building approach (White & Frederiksen, 1990; Raghavan & Glaser, 1995; Gobert & Clement, 1999) in which simpler models (e.g., static representations of structural information) provide conceptual leverage for more complex models (e.g., causal models) of scientific phenomena, these, in turn, support model-based reasoning. We support students' model-based reasoning using scaffolds designed by our group (Gobert & Buckley, 2003) and in accordance with model-based learning theory (Gobert & Buckley, 2000); in doing so, we also draw on literature on students' difficulties in learning with models (Lowe, 1993).

The inquiry skills in national standards (NSES, 1996; U.S. Dept of Education, 1993) match pedagogically with model-based teaching and learning, the theoretical framework underlying our research, learning activities, and assessment (Gobert & Buckley, 2000). The tenets of model-based learning are based on the presupposition that understanding requires the construction of mental models, and that all subsequent problem-solving or reasoning are done by means of manipulating or 'running' these mental models (Johnson-Laird, 1983). Model-based reasoning also involves the testing, and subsequent reinforcement, revision, or rejection of mental models (Buckley & Boulter, 2000). This represents authentic science thinking in that it is analogous to hypothesis development and testing among scientists (Clement, 1989). The reasoning processes of hypothesis generation from the model, testing that hypothesis, and interpreting the data are among the higher order inquiry skills that are difficult to teach and are the type of reasoning needed in inquiry (Raghavan et al, 1995; Penner et al, 1997; White et al, 2002; Gobert, 2000).

Our technology: What's under the hood?

Content Engines (BioLogica, Dynamica, Chemica, & Connected Chemistry [i])

These are implemented in Java, using model, view, controller architecture. They are Event-driven via action listeners. Actions may be initiated by user, via UI, or by model, via state change events.

Script Layer

Our authoring environment uses node-and-arc representation of script structure. Nodes contain executable code and screen layout specification. Arcs implement flow control and node initialization and cleanup functions. State saving is achieved by designating alternative start nodes.

Pedagogica™

Links scripts to engines at runtime, using Java's introspection capability. These include generic objects such as questions, graphs, etc. They implement logging functions, including encryption and backup of log files.

CC Portal

Implements web-based school-topic-teacher-class-student registration process. It parses XML in student log files and populates MySQL database. It archives and maintains all data and provides browser-based online access to reports for administrators, teachers, and students.

Data mining tools

Produce customized reports for researchers. The output is exported to third-party statistics programs.

Measuring Inquiry in situ

Inquiry is critical to science reform efforts as acknowledged by national standards (NSES, 1996; U.S. Dept of Education, 1993) but research on inquiry skills has been hampered by the difficulty and complexity of measuring inquiry, in particular, separating inquiry from its context. Since inquiry skills are developed in rich scientific contexts, their assessment needs to be conducted within the scientific domains and contexts in which they are embedded (Mislevy et al., 2002). In the past, two approaches to measuring inquiry have been used: short answer tests of specific skills, and hands-on performance assessments. The former can be incorporated into large-scale standardized assessments but have been criticized because it is unclear whether decontextualized knowledge of the various skills that make up inquiry are sufficient to allow students to undertake inquiry (Pellegrino, 2001). The second option, performance assessment, appears to be more authentic because it requires a greater integration of specific skills to solve real problems (Ruiz-Primo & Shavelson, 1996) however, these are seldom used in schools, due largely to the difficulty of administering reliable assessments and the resulting high cost.

Our project uses a different approach that offers both the validity of performance tests and the simplicity demanded by large-scale assessments: that is, computer-based assessments of inquiry that are embedded in instructional activities. Using our scripted models which are based on scientific laws, the learner can ask a question of such a model, develop a plan of action for using the model to answer the question, run an experiment to test the model, collect the data, analyze the data, and communicate findings to other users. The advantages of using such a system to conduct empirically rigorous research on learning and assessment of inquiry skills are many, as follows. *Data collection.* Because all activities are on-computer, we can effortlessly and accurately monitor and record every student response and action. *Control.* Because we have a great degree of control over the learning environment, we can simplify it to make the content more accessible and the experiments both rigorous and easy to conduct. This can save time in development and one can increase the complexity of the science concepts being assessed. *Reproducibility.* All aspects of the assessment can be exactly reproduced—the experiment, the scaffolding, and the hints. Furthermore, there are no uncontrolled clues for the learner, such as the tone of a human response or non-verbal clues. *Integration with instruction.* The same model and technical environment used for learning activities can be used for assessment. The assessment can be part of instruction, so that additional class time is not required. *Reach.* Because only a networked computer is needed, we can conduct this research anywhere and the resulting assessment tools can be used in any school, nationwide or worldwide.

Inquiry Hot Spots

As previously stated, our computer models are hypermodels (Horwitz & Christie, 1999) which are scripted using Pedagogica ™ (Horwitz & Burke, 2002), a technological infrastructure that logs all students' interactions with our models and responds to learners based on their input. We log all students' interactions with models and use data from inquiry "hot spots", i.e., tasks that require deep reasoning and contain multiple components of model-based inquiry (representation, acquisition, integration, reasoning, and reflection) to characterize students' inquiry strategies on that task. We use students' log files on multiple inquiry hot spots across three domains to address how students'

inquiry skills are developing both within and across domains. One measure of students' inquiry skills is how systematic students are in manipulating models to achieve a goal. Systematicity has been found to be a reliable measure of students' strategic learning and knowledge acquisition strategies (Gobert, 1994; Thorndyke & Stasz, 1980) and is a good measure with which to compare learners since it bears on their skill at estimating solutions (Paige & Simon, 1966). As we proceed with the project we will use students' data on inquiry hot spots and evaluate their relationship to both conceptual learning measurements, i.e., pre-post content tests and to measures of students' epistemologies of models and views of science since students' epistemologies of models have been found to influence science learning (Gobert & Discenna, 1997), thus, it is possible that students' epistemologies influence the manner in which students strategically manipulate models as well.

BioLogica log file analysis

The MAC group has been analyzing students' log files from a model-based inquiry perspective in all three domains. We present data from BioLogica here; data on students' inquiry is presented elsewhere for both Dynamica (Gobert et al, 2006) and Connected Chemistry (Levy & Wilensky, 2006). Here we describe students' performance on inquiry hot spots as they progress through BioLogica's ™12 activities. In the Monohybrid activity students have opportunities to demonstrate not only their models of inheritance (i.e., content knowledge), but also their model-based inquiry skills. For example, we can track how systematic (or not) students are at this task by logging the crosses they make, i.e., students who have made the correct cross to answer the question, but do not realize it, or whether they make repeated crosses. Monohybrid is the fourth core activity and is intended to help students integrate their models of meiosis and fertilization (developed in the first 3 activities) into a model of inheritance. After introducing the Pedigree and Punnett square representations as well as the concept of probability, four tasks use a progressive modeling approach to foster students' abilities to use the representations and their models to reason about models of inheritance. Both of these are essential model-based inquiry skills.

The first two tasks in the series guide students' investigations of the distribution of traits among offspring. Tasks 3 and 4 differ in that they provide little scaffolding and ask students to manipulate the model of the dragon genome to set up particular situations. In task 3 we ask students if it is possible for a pair of dragons to have only 2-legged offspring and then challenge them to make it happen. Task 4 asks students to again manipulate the model of the dragon genome such that a trait appears to skip a generation. Although the first step in task 4 is procedurally the same as task 3, it uses a different model of inheritance (i.e., simple dominance vs. the incomplete-dominance model for Legs) so this constitutes transfer.

We use Task 3 here to describe the type of data collected and how we analyze students' inquiry skills. Students use the Chromosome tool to inspect and alter the genome of the parents. They then cross the parents using the Cross tool and observe the 40 offspring randomly generated by the meiosis and fertilization model. We determine whether students are successful by checking that the parents have the necessary genotypes to produce only 2-legged offspring. Student performance is scored by computer, based on whether they made the correct prediction, whether they were successful, how many attempts they made, and whether they repeated any crosses (an indication of haphazard (as opposed to systematic) behavior. There are six categories ranging from A (correct on first try) to F (failed with only one attempt). In between are categories that distinguish successful and systematic, successful but haphazard, and similar categories for unsuccessful students.

Students' actions and answers are captured in xml files that are uploaded to our server. The xml log files have been validated at each stage of processing and data reduction. First, we created concise chronological reports of students' actions. We then created algorithms that produced summary records consisting of one record per log file. We then created and validated algorithms for producing one record per student. It is used in the statistical analyses that follow. For Task 3, the statistical records include fields that identify school, class, teacher, and student ID numbers, cumulative time spent on Task 3, students' final selections for the prediction and whether correct, whether they were successful, number of attempts, and the Task 3 category (A-F) described above. Similar data extractions were performed for the other three tasks, based on the affordances of the data collected.

Overall Performance on Tasks

Table 1 includes summary statistics across the four tasks. As the tasks became more difficult, students spent longer on the tasks. As expected, fewer students succeed at Task 4 than Task 3.

Table 1. Summary statistics for cross task comparison

Task	Average time on task (minutes)	N students who did task	correct prediction (%students)	Punnett square first try (%students)	Punnett square select (%students)	Punnett square predict (%students)	task success (%students)	task success on first try (%students)
1	1.6	639	76%					
2	2.7	647	45%	72%	68%	72%		
3	4.0	581	59%				90%	40%
4	5.8	528					49%	20%

In Task 3 we asked students if a pair of dragons could have only 2-legged offspring and then challenged them to make it happen. Looking at the data more closely, 59% of the students predicted that it could be done. 90% of the students were able to accomplish the task, with 40% able to do so on the first attempt. 5% of the students who predicted it could be done did not succeed in doing so.

Correlation of monohybrid performance with post test scores and gain

In order to identify tasks that best predict learning gains, we ran a series of analyses using a dataset of 649 students in ten member schools. Member schools were selected to obtain a demographically and geographically diverse population of students. The majority of the students (54.2%) were enrolled in 'regular' classes. The correlations listed below are significant at $p<.001$, 1-tailed.

- All four tasks are correlated. (R Squareds ranging from .493 to .845)
- Pre and post test scores are significantly and strongly correlated. (R Squared = .572,)
- Monohybrid subscores pre and post are significantly and strongly correlated. (R Squared = .535)
  Total Gains are negatively correlated with Pre test scores. (R Squared = -.326)
  Monohybrid Gains are negatively correlated with Mono Post test subscore. (R Squared = -.465,). We interpret the negative correlations as indicative of a ceiling effect.
- Performance on all 4 tasks correlates with pre and post test scores and gains, both overall and for monohybrid items, with the exception of Tasks 1 and 4 which are not significantly correlated with Monohybrid Gain. (R Squareds ranging from .075 to .450)

We then ran a series of regressions to determine how well performance on the Monohybrid tasks predicts outcomes and gains when pre test scores are used as a covariate. Mono Post and Mono Gain refer to those items on the post test specifically targeting monohybrid concepts. Table 2 summarizes the results. In each case, t-statistics of the regression coefficients for covariate (pretest) and independent variables (tasks) reveal that only the pretest and Task 3 are significant predictors of outcomes (1-tailed significance <.05).

Table 2. Summary of Adjusted R Squares for outcome variables

	Post test	Total gain	Mono Post	Mono Gain
Pre test	.325	.137	.281	.239
Plus Tasks 1-4	.399	.231	.346	.300
Change due to Tasks 1-4	.072	.094	.065	.061

It is interesting that the tasks make a greater contribution to the variance of post test scores and total gain than to the monohybrid post test score and gain. This suggests that students' performance on these tasks; Task 3 in particular, may be indicative of knowledge and skill components beyond those measured by the monohybrid items of the pre and post tests. We might infer that students who can reason from effect to cause, as required by Task 3, are better able to reason through the questions posed on the test.

We then examined the systematicity of these students' performance on Task 3. We eliminated A (successful on first try) and F (unsuccessful, only one try) categories since both involve just one trial. We compared the post test performance (See Table 3.) of the four remaining groups (n=192) with an ANCOVA using the pre-test as covariate. The ANCOVA indicates that the Pre-test covariate is significant, as is the four-category predictor variable (CS, CH, IS, IH) ($F=7.383$, $p < .001$). Together, the covariate and this variable account for 28.3% of the variance in the Post-test scores. Multiple comparisons using Tukey HSD shows that there are significant differences between the "CS" and "CH" groups as well as the "CS" and "IH" groups. The "CS" group has significantly higher post-test scores (co-varying on pre-test) than did either of the two other groups. The mean difference is significant at the .05 level.

Table 3. Post test mean scores (out of 33)

T3CATSYS4NUM	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
correct systematic (CS)	23.117	.643	21.847	24.388
correct haphazard (CH)	20.426	.788	18.870	21.981
incorrect systematic (IS)	15.306	3.146	9.093	21.518
incorrect haphazard (IH)	17.946	1.362	15.258	20.635

When we group and compare the systematic and haphazard students regardless of their success on Task 3, an ANCOVA indicates that the Pre-test covariate is significant, as is the two-category predictor variable (S, H) ($F=12.578$, $p < .001$). Together, the covariate and this variable account for 25.2% of the variance in the Post-test scores. The post test means of systematic and haphazard students are shown in Table 4.

Table 4. Post test mean scores of systematic and haphazard students.

T3CATSYS2NUM	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Haphazard	19.980	.740	18.518	21.442
Systematic	22.868	.651	21.582	24.154

We have shown that students' performances on Task 3 are a predictor of their later performances on the post test. It is not surprising that students who are successful on Task 3 and students who are systematic are also more likely to do well on the post test. We have also demonstrated our ability to analyze student behavior during inquiry in BioLogica™, which we will extend to other hotspots.

Discussion & Conclusions

In the MAC project activities are scaffolded to support progressive model-building and the development of modeling skills. It is our belief that scaffolding the development of inquiry skills is important because these skills can be then bootstrapped to better learn and more deeply understand science content. Our technology which logs all students' manipulations provides an authentic, time-efficient means of measuring inquiry during instruction, which represents progress in measuring inquiry skills. We are measuring inquiry in three domains, and thus are also able to address how modeling skills develop both within and across domains (Buckley & Gobert, 2006). Here, we provide a rich example of model-based inquiry from BioLogica and demonstrate its relationship to content learning.

This approach can inform instruction since our technology can identify students who are at risk, i.e., those who are failing to developing inquiry skills. Furthermore, since we are obtaining data from many students and from many diverse schools (we are currently running in 350+ schools in 20 countries and 41 states), we will develop a clear sense of when teachers *should* intervene. This is one of the goals of a proposed project (Horwitz, Gobert, & Koedinger, 2005). Logging students' inquiry skills is important to research on human cognition since it permits an excellent window into complex inquiry processes and how they develop. Lastly, we hope that our work can inform technology design, pedagogy, standards, and policy on technology-based inquiry learning.

Endnotes

(1) Levels of partnership determines the degree of support which schools obtain from us.

(2) Connected Chemistry was developed by Uri Wilensky, who is also a Co-PI on this project, & Sharona Levy as part of a sub-contract to Northwestern University.)

References

Buckley, B. C., & Boulter, C. J. (2000). Investigating the role of representations and expressed models in building mental models. In J. K. Gilbert & C. J. Boulter (Eds.), *Developing models in science education* (pp. 105-122). Dordrecht, Holland: Kluwer.

Buckley, B.C., & Gobert, J. (April, 2006). Investigating the affordances of technology for supporting and assessing inquiry in science learning. To be presented at the American Educational Research Association, San Francisco, CA.

Clement, J. (1989). Learning via model construction and criticism: Protocol evidence on sources of creativity in science. In J. A. Glover, R. R. Ronning & C. R. Reynolds (Eds.), *Handbook of creativity: Assessment, theory and research* (pp. 341-381). New York: Plenum Press.

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

Gobert, J. (2000). A typology of models for plate tectonics: Inferential power and barriers to understanding. *International Journal of Science Education, 22*(9), 937-977.

Gobert, J. D., & Buckley, B. C. (2000). Introduction to model-based teaching and learning in science education. *International Journal of Science Education, 22*(9), 891-894.57.

Gobert, J. D., & 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, 36*(1), 39-53.

Gobert, J., & Buckley, B. C. (2003). *Scaffolding model-based reasoning: Representations and cognitive affordances*. Concord, MA: The Concord Consortium.

Gobert, J., & Discenna, J. (1997). *The relationship between students' epistemologies and model-based reasoning.* Paper presented at the American Educational Research Association, Chicago.

Gobert, J., & Buckley, B. C., & Horwitz, P. (April, 2006). *Technology-enabled assessment of model-based learning and inquiry skills among high school biology and physics students*. To be presented at the American Educational Research Association, San Francisco, CA.

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

Horwitz, P., & Christie, M. (1999). Hypermodels: Embedding curriculum and assessment in computer-based manipulatives. *Journal of Education, 181*(2), 1-23.

Horwitz, P., Gobert, J., & Koedinger, K. (2005). *Using technology-based on-line physics inquiry assessments (UTOPIA).* A proposal submitted to the Institute for Educational Science, U.S. Dept. of Education.

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

Levy, S. & Wilensky, U. (April, 2006). *Emerging knowledge through an emergent perspective: High-school students' inquiry, exploration and learning in Connected Chemistry.* To be presented at the American Educational Research Association, San Francisco, CA.

Lowe, R. (1991). Expository Illustrations: A new challenge for reading instruction. *Australian Journal of Reading, 14*, 215-226.

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*. Washington, DC: National Academy Press.

Paige, J.M., and Simon, H. (1966). Cognitive processes in solving algebra word problems. In B. Kleinmuntz (Ed.), *Problem solving*. New York: Wiley.

Pellegrino, J. W. (2001). *Retinking and redesigning educational assessment: Preschool through postsecondary.* Denver CO: Education Commission of the States.

Penner, D. E., Giles, N. D., Lehrer, R., & Schauble, L. (1997). Building functional models: designing an elbow. *Journal of Research in Science Teaching, 34*(2), 125-143.

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

Raghavan, K., & Glaser, R. (1995). Model-based Analysis and Reasoning in Science: The MARS Curriculum. *Science Education, 79*(1), 37-61.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching, 33*(10), 1045-1063.

Schwarz, C. & White, B. (1999). *What do seventh grade students understand about scientific modeling from a model-oriented physics curriculum?* Presented at the National Association for Research in Science Teaching, March 28 - 31, Boston, MA.

Thorndyke, P., & Stasz, C. (1980). Individual differences in procedures for knowledge acquisition from maps. *Cognitive Psychology, 12*, 137-175.

U.S. Department of Education (1993). *Using Technology to Support Education Reform, Chapter III: Support for Student Learning Activities*, www.ed.gov/pubs/EdReformStudies/TechReforms.

White, B. Y., & Frederiksen, J. R. (1990). Causal Model Progressions as a Foundation for Intelligent Learning Environments. *Artificial Intelligence, 42*(1), 99-157.

White, B., Frederiksen, J., Frederiksen, T., Eslinger, E., Loper, S., & Collins, A. (2002). *Inquiry Island: Affordances of a multi-agent environment for scientific inquiry and reflective learning.* Paper presented at the Proceedings of the Fifth International Conference of the Learning Sciences (ICLS), Mahwah, NJ.