FACILITATING AND ASSESSING GENETICS LEARNING WITH BIOLOGICA™

BioLogica™ activities mediate students' interactions with a multilevel model of transmission genetics and foster their development of mental models of inheritance, while collecting log files of students' actions and answers as they work. This paper presents findings at the classroom level and at the level of student performance on selected assessment tasks. We analyzed log file data to document the nature and extent of BioLogica use in demographically diverse classrooms across the country. Our analysis is based on log file data collected in 10 schools encompassing 58 classrooms with the most complete datasets. Learning gains (as determined by pre and post tests) varied across different class levels (honors, college prep, regular, AP) and different implementation types. Paired t-tests revealed a significant difference between the pre- and post-test scores for students at each of the four class levels. College Prep students earned the greatest gains (mean=8.27). In 38 of the 58 classrooms, post test means were significantly higher than pretest means ($p<.05$; 1-tailed). However, in 5 classrooms post test means were significantly lower. Over all 58 classes, the number of activities used by a class accounted for 8.7% of the variance in gains. An ANOVA reveals that classes with gains used significantly more activities on average (6.04) than classes with losses (4.67; $F = 4.67$, $p < .05$). In addition, we examined the log files of individual students and characterized student performance on selected tasks relevant to specific models of inheritance, reasoning and inquiry skills. We compared performance on four tasks with related items in the pre and post tests as well as overall gains, determining that one task was a significant predictor of learning gains. We discuss the implications of these analyses for informing more nuanced and timely assessments of student learning and inquiry skills.

Barbara C. Buckley, The Concord Consortium
Janice Gobert The Concord Consortium
Amie Mansfield, The Concord Consortium
Paul Horwitz, The Concord Consortium

This paper presents BioLogica™ data illustrating the affordances of Pedagogica for facilitating research into the development of models of inheritance among high school students and students' ability to reason with those models to solve familiar and novel problems. BioLogica is one of the three content areas of the Modeling Across the Curriculum project.

Project Background

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).

BioLogica™ Research

BioLogica is a hypermodel (Paul Horwitz & Burke, 2002; P. Horwitz & Christie, 1999; P. Horwitz & Tinker, 2001) designed to help high school students understand and be able to reason about transmission genetics. It consists of a series of 12 model-based learning activities based on the idea of progressive model-building (Raghavan & Glaser, 1995; White & Frederiksen, 1990) within a framework of model-based learning (Buckley, 2000; J. D. Gobert & Buckley, 2000). It is available for download from http://mac.concord.org. Scaffolding guides students as they interact with multilevel models in a variety of tasks. Some of the tasks serve as embedded assessments; others as performance assessments. Log files generated while the students work through the activities provide a trail of student actions and inputs which we use to characterize and assess students' models and reasoning as well as problem-solving strategies and inquiry skills.

During the 2004-2005 school year, our activities were used in demographically diverse schools across the US. We analyzed log file data from nearly 2000 students to document the nature and extent of BioLogica use in each classroom, determining which activities were used over what period of time and whether they were used in a block of time or distributed over weeks. Our analysis is based on data collected in 58 classrooms in 10 schools with the most complete datasets. This information was triangulated with information from surveys completed by teachers to characterize different implementation types. Learning gains (as determined by comparing scores on identical pre and post tests, 33 multiple choice items) were compared across classes for different implementation types and for different class levels (honors, college prep, regular, AP).

BioLogica log files also capture the chronological sequence of inquiry processes, i.e., genetic crosses made, tool use (how they use the chromosome tool to examine offspring genotypes) and what other information they seek. To investigate the development of students' genetics understanding we examined the log files of individual students and created metrics with which to characterize student performance on selected tasks distributed throughout the 12 BioLogica activities. Protocols for analyzing this data were developed through repeated cycles of validation and data reduction to ensure that the concise reports and summaries captured students' actions accurately. We coded tasks for their relevance to the specific model of inheritance (simple dominance, incomplete dominance, sex-linked, and polygenic), reasoning (cause-to-effect, effect-to-cause) and inquiry skills (e.g., generating and interpreting data). We compared performance across tasks in other activities related to the same model of inheritance as well as to items in the pre and post tests.

Results

Students' learning gains, as evidenced by pre- to post-score comparisons, varied by class level. On average, across member schools, the Honors students earned the highest pre-test score (mean= 18.59), while the College Prep group earned the greatest gain scores (mean=8.27). Regular students, the largest constituency (n=402), earned an average pre-test score of 15.33 and average gain of 3.58. Paired t-tests revealed a significant difference between the Biologica pre- and post-test scores for students at each of the four class levels.

Table I: Mean raw scores, standard deviations, and percentage scores for Biologica pre- and post-test by class level (total n items=33)

Class Level		Mean Raw Score (Mean Percentage)	Std. Deviation
College Prep (n=44)	Total Score Pre	10.82 (33%)	3.70
	Total Score Post	19.09 (58%)	7.65
	Gain (mean difference)	8.27(25%)*	9.41
Honors (n=262)	Total Score Pre	18.59 (56%)	5.98
	Total Score Post	23.24 (70%)	5.62
	Gain (mean difference)	4.65 (14%)*	5.28
Regular (n=402)	Total Score Pre	15.33 (47%)	5.22
	Total Score Post	18.91 (57%)	6.51
	Gain (mean difference)	3.58 (11%)*	5.47
Remedial (n=9)	Total Score Pre	12.22 (37%)	1.64
	Total Score Post	16.67 (51%)	3.24
	Gain (mean difference)	4.44 (14%)*	2.74

*statistically significant at the p <.05 level

In 48 of the 58 classrooms where at least 50% of the students had taken both pre and post tests, post test means were higher pre test means; 38 classrooms performed significantly better (p<.05; 1-tailed). However, in 10 of the 58 classrooms post test means were lower than pre test means, with 5 classrooms scoring significantly lower (p<.05; 1-tailed). See Appendix A for the table of results. Lower performance on post tests could arise from a number of causes:

- ☐ BioLogica activities may be too difficult and therefore frustrating for some students.
- ☐ The teacher did not monitor students' use and progress or include it as part of their grade.
- ☐ We confused them.

Comparing classrooms with significant gains to classrooms with significant losses, an ANOVA reveals that classes with gains used significantly more activities on average (6.04) than classes with losses (4.67; F=4.67, p <.05). Over all 58 classes, the number of activities used by a class accounted for just 8.7% of the variance in gains.

One possible reason for the relatively low variance in gains may relate to the ways in which teachers used BioLogica activities. From their classroom communiqués, we know that some teachers used BioLogica to introduce concepts, to review concepts or interwoven throughout their genetics curriculum. We cannot make large-scale comparisons across uses due to sparse data from teachers' classroom communiqués**.**

However, in one large school, two teachers undertook an experiment in which one teacher (Teacher W) integrated BioLogica throughout the genetics unit while the other (Teacher S) taught genetics as he was accustomed to doing and used BioLogica at the end of the year as a review. Both taught Honors classes. We hypothesized that the classes who used BioLogica as review would have higher pretest scores and lower gains than the classes who used BioLogica in an integrated manner. Statistical analysis supported these hypotheses. As shown in Table II, the students who used it as a review had a significantly higher mean score on the pre test than the students who took the pretest before genetics instruction.

Table II. Pretest Means by teacher

	teachervar	N	Mean	Std. Deviation	Std. Error Mean
Total_Scorepre	Teacher W	60	17.47	4.073	.526
	Teacher S	107	23.20	4.435	.429

t=8.2444, p<.001

From Table III, we see that the students who used BioLogica for review posted significant gains over their pretest scores, but had significantly lower gains than the students who used BioLogica integrated within the genetics unit.

Table III. Gains by teacher

	teachervar	N	Mean	Std. Deviation	Std. Error Mean
gain	Teacher W	60	6.2833	5.44337	.70274
	Teacher S	107	3.0561	4.58840	.44358

t=4.074, p < .001

<u>BioLogica log file analyses</u>

When we analyzed students' actions when undertaking selected tasks posed in BioLogica activities, we learned that we are able to categorize students' performances and instantiate the process in computer programs that do so for the large number of students in our study. This section illustrates using data drawn from students' use of the *Monohybrid* activity.

After students have worked with the models of meiosis and fertilization, the causal models at the cellular level of our multilevel model of transmission genetics, we introduce them to the models of inheritance through work with models at the pedigree level. Using pedigree models they can manipulate the genotypes of organisms at the allele level, breed organisms and observe the traits of their offspring. *Monohybrid* is the foundational instructional activity at this level. *Monohybrid* poses four tasks intended to help students integrate their models of meiosis and fertilization (developed in the first three activities) into a model of inheritance. The four tasks use a progressive modeling approach to foster students' abilities to use the representations and their models to reason about models of inheritance. 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. Repeated crosses do not provide new data with which to reason and increase the complexity of the visual display of data. 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 (B), successful but haphazard (C), unsuccessful but systematic (D) and unsuccessful and haphazard (E).

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.

<u>Overall Performance on Tasks</u>

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

Table IV. Summary statistics for cross task comparison

Task	Average time on task (minutes)	N students who did task	correct prediction (%students)	Punnett square on 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.

<u>Correlation of monohybrid performance with post test scores and gain</u>

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. 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 V 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 V. Summary of Adjusted R Squareds 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 VI.) 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 VI. 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 VII.

Table VII. 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.

Discussion and conclusions

The very obvious limitation of this study is the assumption that the behavior captured in a log file results from the cognition of one student, when in fact it could result from informal collaboration with the student at the next computer. This would need to be addressed if the tasks were to be used for high-stake assessments.

We have demonstrated that students do learn genetics with BioLogica and that we can analyze their problem-solving and inquiry behaviors through the use of log files. Some students are able to solve problems quickly using presumably precompiled models of inheritance. Others succeed at tasks within their zone of proximal development by reasoning with not-yet-compiled models of inheritance, some proceeding systematically, some haphazardly. Still others do not succeed, often working haphazardly or giving up very quickly. It is not surprising that those who reason systematically, regardless of whether they succeed at a particular task, are more successful on the post test, since it is indicative of content knowledge models and the ability to reason with them. It will be interesting to see if this finding recurs as we examine student performance on other tasks in BioLogica. If it does, tasks like those analyzed in *Monohybrid* offer potential as replacements for or supplements to fact-based tests and performance assessments. Even more importantly, they offer opportunities for timely formative assessments that students and teachers can use to monitor learning.

References

American Association for the Advancement of Science. (1993). *Benchmarks for Science Literacy*: 1993. New York Oxford: Oxford University Press.

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., 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, 13*(1), 23-41.

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.

Dede, C., & Lewis, M. (1995). *Assessment of emerging educational technologies that might assist and enhance school-to-work transitions*. Washington, DC: National Technical Information Service.

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

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.

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., & 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.

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., & 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., Buckley, B., Dede, C., Horwitz, P., Wilensky, U., & Levy, S. (2004). *Modeling Across the Curriculum (MAC): Technology, Pedagogy, Assessment, & Research*. Paper presented at the American Educational Research Association, San Diego, CA.

Horwitz, P. & Tinker, R. (2001). Pedagogica to the rescue: A short history of hypermodel. Concord Newsletter, 5(1), 1-4.

Horwitz, P. (1995). Linking Models to Data: Hypermodels for Science Education. *The High School Journal, 79*(2), 148 - 156.

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.

Levy, S.T., Novak, M., Wilensky, U. (2005). Connected Chemistry Curriculum 1.3 Evanston, IL. Center for Connected Learning and Computer Based Modeling, Northwestern University. ccl.northwestern.edu /curriculum/chemistry/.

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.

Shavelson, R. J., & Ruiz-Primo, M. A. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement, 36*(1), 61-71.

Shavelson, R. J., Li, M., Ruiz-Primo, M. A., & 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.

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.

Wilensky, U. (1999). NetLogo, Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL. http://ccl.northwestern.edu/netlogo.

APPENDIX A.
Pre-Post Test Gains by School and Class (total n items=33)

School	Class ID	N	Minimum Gain	Maximum Gain	Mean Gain	Std. Deviation
AMS	5174	16	-1.00	22.00	7.63*	6.21
	5175	20	-8.00	14.00	4.50*	5.39
	5176	18	-2.00	16.00	6.89*	5.19
	5177	4	1.00	11.00	6.00	4.08
	5178	2	8.00	9.00	8.50*	0.71
	5352	24	-8.00	15.00	4.17*	5.27
	5353	21	-3.00	8.00	1.95*	3.23
	5354	3	-1.00	7.00	2.33	4.16
	5355	25	-6.00	13.00	2.64*	5.15
	5356	5	-1.00	11.00	4.80	4.92
	5357	26	-4.00	11.00	3.38*	4.29
	5358	3	-4.00	6.00	0.33	5.13
FHS1	3724	3	1.00	23.00	11.33	11.06
	4135	10	-1.00	6.00	1.30	2.06
	4136	7	-9.00	18.00	6.57	8.32
	5499	14	-2.00	21.00	11.71*	6.59
	5500	13	.00	28.00	14.31*	8.17
	5903	21	-8.00	15.00	5.38*	5.45
FHS2	5904	24	-2.00	9.00	3.38*	3.20
	6627	9	1.00	9.00	4.44*	2.74
	3505	2	-3.00	2.00	-0.50	3.54
FPS	3506	2	-2.00	9.00	3.50	7.78
	3507	3	-8.00	13.00	2.33	10.50
	3511	6	1.00	18.00	10.50*	6.66
	3918	19	-6.00	12.00	4.26*	4.90
	4120	19	-5.00	7.00	1.00	3.71
	4799	10	-1.00	20.00	9.70*	6.36
LHS1	4800	11	1.00	17.00	6.09*	4.41
	4405	6	-5.00	12.00	4.00*	6.99
	4406	8	-17.00	5.00	-3.00	6.59
	4407	3	-3.00	9.00	4.67	6.66
LHS2	3263	9	-10.00	6.00	-1.00	5.32
	3264	19	-6.00	10.00	2.16*	3.67
	3265	16	-1.00	9.00	3.88*	3.34

*statistically significant at the p <.05 level continued on next page.

Table XV (cont.): Pre-Post Test Gains by School and Class (total *n* items=33)

School	Class ID	N	Minimum Gain	Maximum Gain	Mean Gain	Std. Deviation
NHHS	3266	15	.00	11.00	4.40*	3.11
	3270	16	-8.00	7.00	0.75	4.02
	3271	22	-3.00	14.00	4.18*	3.79
	3272	20	-5.00	9.00	2.80*	3.83
	3273	21	-4.00	14.00	3.71*	5.53
	3813	4	-3.00	5.00	0.00	3.46
	3814	7	-9.00	6.00	-4.00	5.63
	3815	5	-8.00	.00	-4.00*	3.08
PHS	3816	7	-11.00	3.00	-2.86	4.78
	3817	8	-12.00	-2.00	-7.00*	3.46
	3822	3	3.00	4.00	3.33*	0.58
	5597	15	-6.00	13.00	5.40*	5.95
	5598	18	-6.00	13.00	5.83*	4.64
	5599	16	-3.00	14.00	7.00*	5.09
SHS	5600	17	-1.00	12.00	4.88*	4.00
	5601	11	-2.00	14.00	4.64*	5.05
	5602	22	2.00	14.00	6.77*	3.88
	4154	3	2.00	16.00	9.33	7.02
	4155	17	-1.00	17.00	6.41*	5.08
	4156	11	-6.00	9.00	3.00*	4.31
WCE	4377	16	-5.00	11.00	3.81*	4.31
	4378	15	-5.00	20.00	5.60*	5.51
	4383	14	1.00	17.00	8.71*	5.24
	4938	10	-5.00	4.00	0.30	2.71

*statistically significant at the p <.05 level