Model-based Teaching and Learning with Hypermodels: What do they learn? How do they learn? How do we know?

Barbara C. Buckley, Janice D. Gobert, & MaryAnn T. Christie
The Concord Consortium

Abstract

The research presented in this paper is part of a large-scale design study conducted in demographically diverse classrooms with software that is under development. BioLogica, a hypermodel environment for learning genetics, was used in multiple classes in eight high schools. BioLogica activities, data logging, and assessments were refined across this series of implementations. All students took a genetics content knowledge pre- and post-test and completed epistemological and experiential surveys. Traces of students' actions and responses to computer-based tasks were electronically collected (via a "log file" function) and systematically analyzed. An intensive three-day field test involving twenty-four middle school students served to refine methods and create narrative profiles of students' learning experiences, outcomes, and interactions with BioLogica. Since BioLogica activities, the instruments used to assess learning, and data logging capabilities changed over the course of the year, we report on two high school implementations and the field test as self-contained studies to document the changes and the outcomes at different phases of development.

Presented as part of the symposium Hypermodel Research in Theory and Practice.
(M. Christie, Chair), American Educational Research Association, New Orleans, April 2002.

This paper is based on research conducted as part of the Fostering Transfer project supported by the National Science Foundation under grant No. REC-0087579. Any opinions, findings, and conclusions expressed are those of the presenters and do not necessarily reflect the views of the National Science Foundation.

We are developing a computer-based learning environment that we hope will enable students in high school classrooms to build a deep understanding of core concepts in Mendelian genetics. The pedagogical challenges are numerous. What do we mean by deep understanding? How can we help them develop this deep understanding? How do we know when they've done so? Fortunately, we're part of a team that has been working on these challenges for several years with funding from NSF. This paper focuses on the learning that takes place when students use BioLogica, an interactive genetics curriculum, in their high school classrooms. It presents the model of learning we use, what this looks like in practice, how we determine the nature and extent of student learning when BioLogica is used in high school classrooms, and what we've learned about all of the above.

Our research is a large-scale design study conducted in demographically diverse classrooms with software that is under development. BioLogica was used in 8 high schools and in an intensive 3 –day field test with 24 middle school and high school students during their spring break. In this paper we report on two high school implementations where three teachers taught both experimental and control classes as two separate studies and on the intensive field test in a third study. We conclude with a discussion that draws from all three studies and points the way for future research and development.

Model-based learning

The theory we employ is an elaboration and extension of Model-Based Teaching and Learning (MBTL) set forth in a special issue of the International Journal of Science Education (Gobert & Buckley, 2000). The tenets of model-based learning are based on the presupposition that understanding requires the construction of mental models of the phenomena under study, and that all subsequent problem-solving, inferencing, or reasoning are done by means of manipulating or 'running' these mental models (Johnson-Laird, 1983). We view mental models as internal, cognitive representations used in reasoning of many kinds (Brewer, 1987; Rouse & Morris, 1986). Mental models, like prior knowledge, influence our perceptions of phenomena and our understanding of information. Interactions with phenomena and representations, in turn, influence our mental models (Gentner & Stevens, 1983; Johnson-Laird, 1983).

Before proceeding, it is important to define the different types of models we use in addition to the notion of models as mental representations of phenomena. Our starting point is Norman's (Norman, 1983) differentiation of models related to a target system or phenomenon that is to be represented or modeled. Norman distinguishes not only the learners' mental models but also the scientist's and designer's conceptual models of the system, as well as the researchers' conceptualization of the learner's mental models.

The hypermodels used in this project are another type of model that we add to the framework of models described by Norman (1983). Hypermodels are conceptual models realized as computer models embedded in interactive curricula. These are external representations (as opposed to mental models, which are internal representations) with which the learner interacts, and in doing so, constructs and/or revises his/her mental model. Representations are considered models only when they represent structural, dynamic, and/or causal aspects of the target model; that is, they are not just visualizations or diagrams of phenomena.

With these definitions in mind, 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 mental models. This is analogous to hypothesis development and testing seen among scientists (Clement, 1989). See Figure 1.

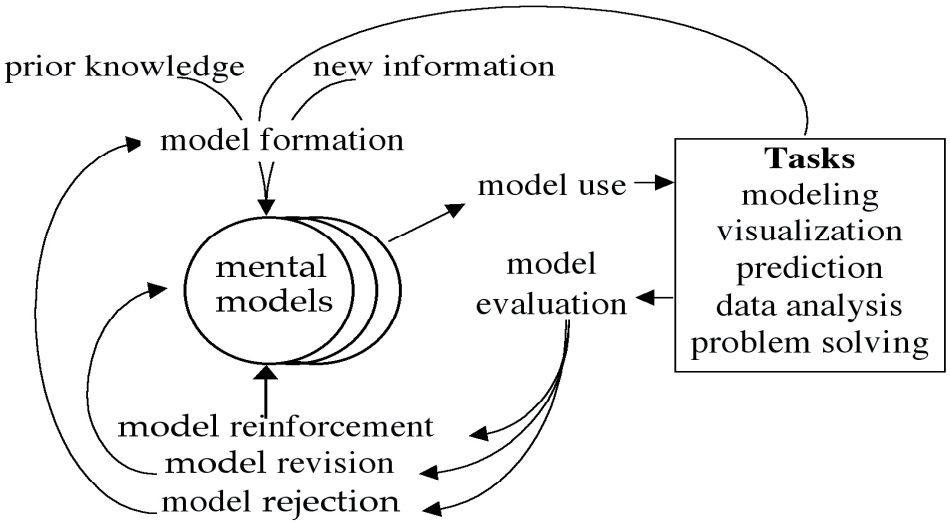

Figure 1. Model-based Learning Framework

In response to the demands of a task, a learner draws on prior knowledge and new information during model-formation to construct a mental model of some phenomenon (Buckley, 1992, 2000; Kozma, Jones, Wykoff, & Russell, 1992). The learner's prior knowledge may include a partial mental model of the phenomenon or naive models that are incompatible with the scientifically accepted model. If the learner's model is used successfully to accomplish the task, the model is reinforced (Clement, 1989) and may eventually become a precompiled, stable model (Vosniadou & Brewer, 1992) from which students are capable of making inferences, etc. However, if inconsistencies and/or deficiencies in the model are noted (Bransford, Sherwood, Vye, & Rieser, 1986), the learner may reject the model and form a new one, or revise the initial model (Chinn & Brewer, 1993; Clement, 1989; Schauble, Glaser, Raghavan, & Reiner, 1991). Model revision may involve modifying parts of the existing model or elaborating the model by adding to or combining existing models. Embedding a model in a larger model is an example of elaboration (deKleer & Brown, 1983; Monaghan & Clement, 1994; Stewart & Hafner, 1991). Thus, the mental model evolves through multiple recursions as it is made increasingly complex and, hopefully, more accurate (Johnson-Laird, 1983; White & Frederiksen, 1998).

Model-based Teaching with the BioLogica Hypermodel

Hypermodels are software environments that allow learners to interact with a manipulable model of some phenomenon in a domain (Horwitz, 1995, Horwitz and Christie, 1999). Hypermodels are controlled by short programs, or activity scripts, that mediate a learner's interactions with the model.

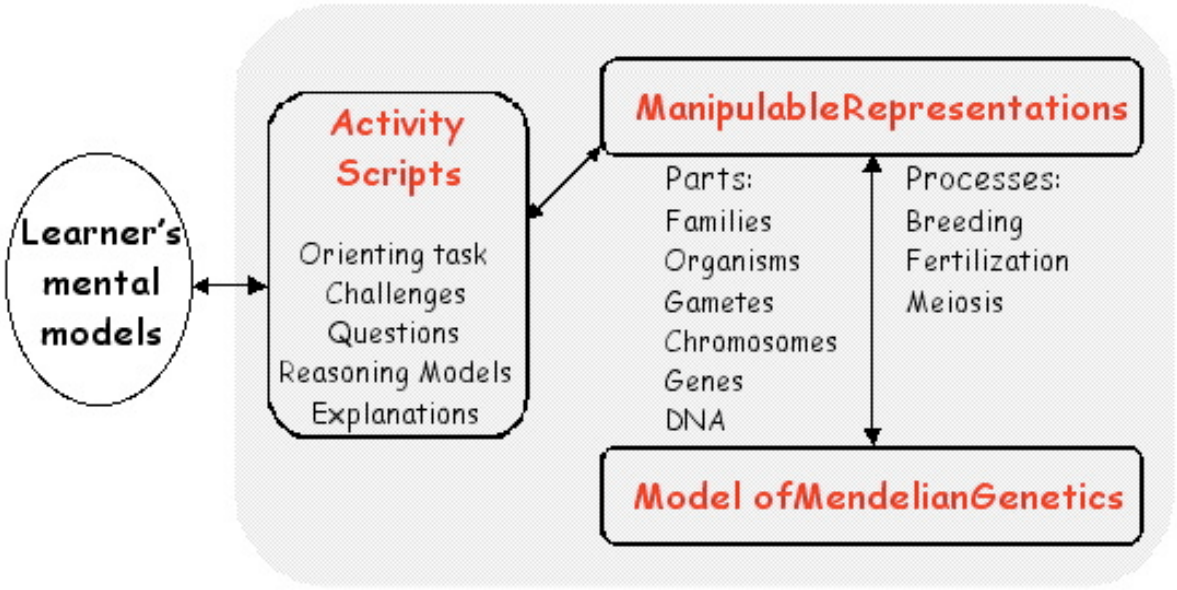

Figure 2. BioLogica Hypermodel

The core of the BioLogica hypermodel (See Figure 2) is a model of Mendelian genetics that represents the genetic mechanisms of inheritance as the parts and processes that take place at multiple levels of biological organization. The learner interacts with the model through manipulable representations appropriate to the task posed by the activity script. Because the manipulable representations operate on the core model, changing any one of them can affect each of the others, as appropriate. For example, altering an organism's genotype can affect not only that organism, but also any of its offspring that happen to inherit the altered gene. The activity scripts are used to orchestrate instructional activities built around a series of challenges. The pedagogical elements outlined in Table I scaffold the learner's interaction with the challenge, the manipulable representations and ultimately with the model of Mendelian genetics.

Table I. Pedagogical elements in activity scripts

Element	Description	Location
Advance organizer or orienting task	Briefly describes expected performance in terms of focus question, content and actions	Headlines in login screen Elaborated in introduction screen
Narratives	The narrative into which puzzles are woven	Begun in introduction and woven throughout the activity
Tasks & puzzles	Require cause-to-effect and effect-to-cause reasoning, within and between generations	The heart of all activities
Representations Assist	Links diagrams with vocabulary, highlights structures involved in genetics	Rollover exploration when new representation is introduced
Reasoning models	Walks learner through reasoning strategies	After initial exploration of concept and representations
Explanations	Summarizes concepts and models good explanations	After completion of task/puzzle
Embedded Questions	About representations, models, & reasoning About the pieces of the model central to the activity	Throughout activity
Feedback on actions & responses	Monitors responses and problem solving actions	More frequent initially then fades away before assessment questions
Parallel phenomena exposure	Scenarios and questions about genetics in others species and in humans	After in-depth exploration and experimentation with dragon genetics
Assessment Questions	General concepts Activity specific	End of activity
Reflective questions	What they knew What they learned What they wonder about now	End of activity

Each of the elements serves both pedagogical purposes and research purposes. For example, embedded questions are intended to facilitate a deeper engagement on the part of the learner by posing questions that require more attention to the details of the activity. They are also intended to elicit responses from which we can infer the state of the learner's mental models. Narratives and instructions model the kind of reasoning a geneticist might pursue in solving a challenge. Activity scripts also monitor a learner's actions so that constructive feedback can be given as appropriate. Explanations and solutions are provided only after the learner has had an opportunity to grapple with the challenge. Each activity also creates a log file that records students' actions and responses and can be processed to yield information useful to teachers, students and researchers.

Research Design and Methods

Our research is a large-scale design study conducted in demographically diverse classrooms with software that is under development. The activities, data logging, and assessments were refined across a series of implementations. Data were collected in multiple classes from one middle school and eight high schools. An intensive three-day field test involving twenty-four middle school students served to refine methods and create narrative profiles of students' learning experiences, outcomes, and interactions with BioLogica. From this massive amount of data, we

will focus on three studies conducted at various times in the development of BioLogica to examine both development issues and research findings. Since BioLogica activities, the instruments used to assess learning, and data logging capabilities changed over the course of the year, we report on two high school implementations and the field test as self-contained studies. To provide a frame for these studies we begin by describing the development of activities, assessments and data logging.

Development of Activities

Prior research with GenScope (Hickey, Kindfield, Horwitz, & Christie, 1991, 1999; Hickey & Kindfield, 1999; Hickey, Wolfe, & Kindfield, 1998b; Horwitz et al., 1998; Kindfield & Hickey, 1999) guided the initial development of BioLogica activities, which are built around a series of puzzles or challenges. BioLogica activities began to be used in classrooms in November 2000. Revision of these activities was informed by classroom observations in three schools, prior research by the authors (Buckley, 2000; Buckley & Boulter, 2000; Gobert, 2000), and by the model-based teaching and learning framework described earlier. During the first phase of testing (11/00 through 2/01) BioLogica was used in three high schools, the number of activities was increased from 8 to 12, the first data logging was implemented, and a Teacher's Guide was created. Table II summarizes the development of BioLogica activities and data logging during the second phase of testing and implementations in five schools. It shows the activities included in each version, which activities were added or revised (◊), important additions or modifications, and finally which sites used this version. Fixing bugs in the software was ongoing. This paper describes the implementations at sites "M" and "L" and the field test at "MV". See Appendix A for a more complete description of BioLogica activities.

Table II. Summary of BioLogica Activities Development

Release date	3/01	4/01	5/1/01
Activities	Introduction Rules Meiosis Inheritance Monohybrid Mutations Mutations Inheritance Horns Dilemma Dihybrid Sex Linkage Sex Linkage II Scales Plates GenScope	◊Introduction ◊Rules ◊Meiosis Inheritance Monohybrid Mutations Mutations Inheritance Horns Dilemma Dihybrid ◊Sex Linkage Scales Plates ◊Invisible Dragons GenScope	Introduction Rules Meiosis Inheritance Monohybrid Mutations Mutations Inheritance Horns Dilemma Dihybrid Sex Linkage Scales Plates Invisible Dragons GenScope
Significant changes	Changed sex determination to match humans	Scaffolds in Intro & Rules, data logging refinements	LogIn/LogOut time stamps
Implementation sites	M	MV (field test)	L, CC, GM

Development of pre and post tests

Like the activities, the pre and post tests evolved over the academic year 2000-2001. In the initial implementation of BioLogica at site "MR" the pre and post tests were identical. The instrument consisted of 70 NewWorm items originally developed for the GenScope project (Hickey, Wolfe, & Kindfield, 1998a). Identical pre and post tests were administered at site "F", using a new, shorter, version of NewWorm from Hickey and Kindfield with the addition of four multiple choice items consisting of published items from MCAS and NY Regents exams. We learned that most students and many teachers considered even the 35-item test too long. In subsequent implementations (see Table III below) items were deleted from the test if the item tested a concept (such as crossing-over) that was not addressed in any of the BioLogica activities used in that implementation. The pre test items were a subset of the post test items. Additional multiple choice items were included to better assess how students performed on standardized test items and a few multiple choice questions were included that addressed common naive conceptions. Because the NewWorm test assumes an understanding of genetics terminology, but doesn't test for that understanding independent of the problem solving required for most NewWorm items, we introduced open-ended conceptual questions beginning with implementations in April 2001.

Table III. Developmental progression of pre and post tests

Sites Type of items	MR 11/00	F 2/01	M Pre 3/01	M Post 3/01	MV Pre 4/01	MV Post 4/01	L Pre 5/01	L Post 5/01	CC Pre 5/01	CC Post 5/01
NewWorm	70	31	23	27	0	32	23	32	25	32
Multiple choice	0	4	3	7	9	11	0	13	11	11
Open-ended	0	0	0	0	8	17	0	17	9	17

We examined all items on the pre and post tests in three ways: (1) using the matrix of reasoning patterns identified by Hickey, Wolfe, and Kindfield (Hickey et al., 1998b), (2) using a model-based learning perspective (Buckley & Boulter, 1999, 2000), and (3) identifying the genetics concept tested. The reasoning matrix has two dimensions: between vs. within generations, and cause-toeffect vs. effect-to-cause. Hickey et al (1998b) demonstrated that questions involving cause-to-effect were easier than effect-to-cause and that reasoning within a generation was easier than reasoning across generations. Drawing on the model-based learning framework, we analyzed items using a method developed for analyzing representations (Buckley & Boulter, 1999, 2000). This analysis focused primarily on the models of parts (structures) and processes needed to answer the question. We have not yet analyzed the utility and reliability of these dimensions.

Development of data logging

First, what do we mean by data logging? When students use BioLogica activities, the activity scripts that guide and control students' interactions with the underlying genetic models generate log files. The specific entries in a given log file are generated either automatically or by design. The automatic entries are used to capture start and stop times for an activity or a given challenge within the activity and to capture learners' responses to questions posed by the activity script. The designed entries can be tailored to answer specific research questions and provide assessment data for teachers, students, and researchers. Embedded questions were added to

activities implemented in April 2001 and later. In addition, most activity scripts capture learners' actions, such as how many crosses were made or what alleles were changed, as they explore and experiment with the representations of the underlying genetic model. Thus, the content of log files varies not only by activity but also by user.

The use of data logging increased throughout the year. Table IV summarizes the state of data logging across several implementations. During the first use of BioLogica in November 2000 only 5 activities generated log files. Of the 54 logs generated by 13 students only 57% were usable; that is, the other 43% were empty, lacking even a time stamp. This is caused either by the user quitting the activity immediately after launch or by a program crash. By May 2001 all 13 activities generated log files eventually achieving a yield rate of 94%.

Table IV. Implementation of Data Logging

Implementation Site	MR 11/00	M 3/01	MV 4/01	L 5/01	C 5/01
Activities Logged (N)	5	10	12	13	13
Students (N)	13	45	24	120	34
Logs generated(N)	54	5715	438	2128	214
Usable Logs (N)	31		319	1566	201
Yield (%)	57		73	74	94

Implementation Studies

At "M" and "L" high schools, three teachers taught both control classes and experimental classes. At site "L" three additional teachers used BioLogica with a range of biology classes. These implementations are reported in Study 1 and Study 2. We used the data collected during the intensive field test conducted at site "MV" to develop methods for analyzing log files. That work is reported in Study 3.

Study 1: "M" High School

Two teachers "C" and "V" used BioLogica in their tenth grade biology classes in March 2001. Each teacher taught two sections for the same length of time. The control section received the teacher's usual lessons in genetics. The experimental section used BioLogica shortly after a 'sex-change' operation that changed the dragon's sex chromosomes to match the human pattern of XX for females and XY for males. This change in the software introduced many bugs that affected the learners' interactions with BioLogica. The number of logs generated is one measure of the technological problems encountered during this implementation. The 45 students who used BioLogica generated over 5000 logs — a monument to the teachers' and students' persistence and patience in using new software.

The pre test consisted of 23 NewWorm items, plus 3 multiple choice items about common misconceptions for a total of 38 points. The post test consisted of 27 NewWorm items and 7 multiple choice items (3 from pre test and 4 from standardized tests).

Study 1A: Teacher "C", *Survey of Biology* course

Analysis of Pre-test.

A t-test was computed in order to test for pre-test differences before the implementation in Teacher C's classes. There were no statistically significant differences found between the control and experimental groups on the pre-test (t = .994; p = .328; $\overline{X}_c$ = 18.7, $\overline{X}_e$ = 16. 5).

Analysis of Post-Test.

An Analysis of Variance (ANOVA) was computed in order to determine whether there were any statistically significant differences on the total post-test score between the experimental and control groups. For this ANOVA the dependent variable was the total post-test score, the independent variable was group (experimental versus control), and the pre-test score was used as a covariate. There was a statistically significant difference found between the control and experimental groups on the post-test with the control group scoring higher. See Table V.

Table V. Comparison of pre and post test means for Survey of Biology course.

Group	N	Pretest			Post test			F statistic	Significance (p value)
		Total number of points	Mean	Std. Dev.	Total number of points	Mean	Std. Dev.		
Control	14	38	18.7	7.2	59	28.4	9.4	14.164	.000*
Experimental	20	38	16.5	6.1	59	23.9	9.1		

*significant at p<.05 α

Analysis of Selected Post-Test Items.

Because this is a largely exploratory study, we conducted a series of ANOVAs in which we selected items from the post-test and pooled them according to the concept that they were assessing. The pooled items assessed students' understanding of: inheritance, dominance, sex linkage, genotype and phenotype, monohybrid, dihybid, and pedigree. For each of these analyses, the pre-test scores were used as a covariate to allow us a more accurate measure of post-test differences between the control and experimental groups.

There was a statistically significant difference found between the control group and the experimental group on three concepts: dominance, genotype and phenotype, and monohybrid. For dominance and monohybrid the control group scored higher, whereas for genotype and phenotype the experimental group scored higher. These results are summarized in Table VI. No statistically significant difference was found on inheritance, sex linkage, dihybrid, or pedigree. See Appendix B for a summary table of this data.

Table VI. Concepts with Statistically Significant Differences between Control and Experimental Groups

Dependent variable	Mean Control Experimental	F statistic	Significance (p value)
dominance	2.357 1.800	7.387	.002*
genotype/phenotype	6.607 7.025	3.594	.039*
monohybrid	9.607 7.575	13.850	.000*

*significant at $p < .05$

ANOVAs were computed in a series of analyses in which we selected items from the post-test and pooled them according to the type of reasoning that they were assessing. The pooled items assessed students on seven different types of reasoning skills: cause-to-effect within generations (cew), effect-to-cause within generations (ecw), cause-to-effect between generations (ceb), effect-to-cause between generations (ecb), Punnett square, structure, and process. For each of these analyses, the pre-test scores were used as a covariate to allow us a more accurate measure of post-test differences between the control and experimental groups.

There was a statistically significant difference found between the control group and the experimental group on three types of reasoning: cause-to-effect between generations (ceb), Punnett squares, and structure. It should be noted, however, that the Levene's Test of Equality of Error Variances (which tests whether the error variances of the two groups are equal, was significant for Punnett squares and structure. Thus, we must interpret the results regarding the two groups' performance on these scores with caution. Students in the control group scored higher on pooled items assessing their reasoning skills of cause-to-effect between generations as well as their reasoning skills on Punnett squares. Students in the experimental group scored higher on pooled items assessing their reasoning skills on structure. These results are summarized in Table VII. No statistically significant difference was found on effect-to-cause within generations (ecw) and effect-to-cause between generations (ecb), however, the control group means were higher than the experimental group means. No statistically significant difference was found on cause-to-effect within generations (cew). An ANOVA could not be performed on process as the mean and standard deviation was zero (0). See Appendix B for a summary table of the data for "M" high school, Survey of Biology course.

Table VII. Summary of Types of Reasoning with Statistically Significant Differences between Control and Experimental Groups.

Dependent variable	**Mean** **Control** **Experimental**	**Levene's Test**	**F statistic**	**Significance (p value)**
cause-to-effect between generations (ceb)	6.107 4.700	.793	12.607	.000*
Punnett square	9.571 7.100	.024*	13.689	.000*
structure	.071 .600	.000*	6.317	.005*

*significant at p<.05

Study 1B: Teacher "V", *College Biology* course

Analysis of Pre-test.
A t-test was computed in order to test for pre-test differences before the implementation in Teacher V's classes. There were no statistically significant differences found between the experimental and control groups on the pre-test (t = .893; p = .378; $\overline{X}_c = 19.1$, $\overline{X}_e = 17.2$).

Analysis of Total Post-Test.
An Analysis of Variance (ANOVA) was computed in order to determine whether there were any statistically significant differences on the total post-test score between the experimental and control groups. For this ANOVA the dependent variable used was total post-test score, the independent variable used was group (control versus experimental) and the pre-test score was used as a covariate. There was a statistically significant difference found between the control and experimental groups on the post-test with the experimental group scoring higher. See Table VIII.

Table VIII. Comparison of pre and post test means for College Biology course.

Group	N	Pretest			Post test			F statistic	Significance (p value)
		Total number of points	Mean	Std. Dev.	Total number of points	Mean	Std. Dev.		
Control	12	38	19.1	5.4	59	37.8	8.9	3.757	.034*
Experimental	25	38	17.2	6.3	59	38.4	7.1		

*significant at p<.05 α

Analysis of Selected Post-Test Items.
ANOVAs were computed in a series of analyses in which we selected items from the post-test and pooled them according to the concept that they were assessing. The pooled items assessed

students' understanding of: inheritance, dominance, sex linkage, genotype and phenotype, monohybrid, dihybid, and pedigree. For each of these analyses, the pre-test scores were used as a covariate to allow us a more accurate measure of post-test differences between the control and experimental groups.

There was a statistically significant difference found between the control group and the experimental group on two concepts: inheritance and pedigree. For inheritance, the control group scored higher, whereas for pedigree the experimental group scored higher. It should be noted, however, that the Levene's Test of Equality of Error Variances (which tests whether the error variances of the two groups are equal, was significant for inheritance. Thus, we must interpret the results regarding the two groups' performance on this score with caution. A summary of results on inheritance and pedigree can be found Table IX. No statistically significant difference was found at $p < .05\alpha$ on dominance, sex linkage, monohybrid and dihybrid, however, the experimental group means were higher than the control group means. No statistically significant difference was found on genotype and phenotype. See Appendix B for a summary table of this data.

Table IX. Summary of Concepts with Statistically Significant Differences between the Control Means and Experimental Means

Dependent variable	**Mean** **Control** **Experimental**	**Levene's Test**	**F statistic**	**Significance (p value)**
inheritance	2.417	.014*	4.268	.022*
	1.640			
pedigree	.500	.770	4.085	.026*
	.960			

*significant at $p < .05\alpha$

ANOVAs were computed in a series of analyses in which we selected items from the post-test and pooled them according to the type of reasoning that they were assessing. The pooled items assessed students on seven different types of reasoning skills: cause-to-effect within generations (cew), effect-to-cause within generations (ecw), cause-to-effect between generations (ceb), effect-to-cause between generations (ecb), Punnett square, structure, and process. For each of these analyses, the pre-test scores were used as a covariate to allow us a more accurate measure of post-test differences between the control and experimental groups.

There was a statistically significant difference between the control group means and the experimental group means on one type of reasoning: cause-to-effect within generations (cew). The control group means for this type of reasoning were higher than the experimental group means. It should be noted, however, that the Levene's Test of Equality of Error Variances (which tests whether the error variances of the two groups are equal, was significant for this type of reasoning. Thus, we must interpret the results regarding the two groups' performance on these scores with caution. These results are summarized in Table X. No statistically significant difference was found at $p = .05\alpha$ on effect-to-cause within generations (ecw) and Punnett squares; however, the control group means were higher than the experimental group means. No

statistically significant difference was found at p = .05α on cause-to-effect between generations (ceb), effect-to-cause between generations, and structure; however, the experimental group means were higher than the control group means. An ANOVA could not be performed on process as the mean and standard deviation was zero (0). See Appendix B for a summary table of the data for "M" High School, Teacher "V".

Table X. Summary of Types of Reasoning with Statistically Significant Differences between the Control Means and Experimental Means

Dependent variable	**Mean** **Control** **Experimental**	**Levene's Test**	**F statistic**	**Significance (p value)**
cause-to-effect within generations (cew)	5.833 5.120	.022	4.688	.016*

*significant at p < .05α

Study 2: "L" High School

During this implementation in May 2001, four teachers used BioLogica daily for one week with a range of classes. In addition, one of the teachers taught two control classes of ninth grade students matched with her two experimental classes.

BioLogica activities were more stable than during the "M" implementation, but still not totally bug free. There were two changes in activities compared to the previous implementation: *Sex-Linkage* and *Sex-Linkage II* were combined and revised into one activity and *Invisible Dragons*, a game that requires reasoning from effect-to-cause between generations, was added. The data logging generated accurate start and end times. Although we weren't able to match log files to individual classes, we were able to average across all the logs for this implementation. Table XI summarizes the logs collected at "L" high school for all classes.

Table XI. Logs collected at "L" high school.

Implementation Site	**L 5/01**
Activities Logged (N)	13
Students (N)	120
Logs generated(N)	2128
Valid Logs (N)	1566
Yield (%)	74
1) **Introduction logs (N)**	273
2) **Rules logs (N)**	266
3) **Meiosis logs (N)**	283
4) **Inheritance logs (N)**	130
5) **Monohybrid logs (N)**	204
6) **Mutations logs (N)**	98
7) **Mutations 2 logs (N)**	36
8) Horns Dilemma logs (N)	32
9) Dihybrid logs (N)	26
10) Sex-Linkage logs (N)	52
11) Scales logs (N)	35
12) Plates logs (N)	25
13) Invisible Dragons logs (N)	106

We could not determine with precision what proportion of students completed each activity. However, we can estimate which activities were completed and which were not completed by most of the students. We estimated that the first five activities were completed by all students. Similarly, it is likely that most of the students did not complete activities after Monohybrid. Even Invisible Dragons with 106 logs may be the result of a small number of students using the activity multiple times. See Study 3 for more detailed information on processing and analyzing log files.

We used this information in conjunction with item analysis of the post test to identify which items on the post test to exclude from analysis. We examined the mean score on each item across the entire data set from "L". We looked at each item that had a mean of less than 50% of the possible points and considered whether it tested a concept that most of the students had explored with BioLogica. If it tested understanding of a concept that was included in the first five activities, we did not exclude the item. Conversely, if the items with low mean scores were not covered in the first five activities we excluded the item from further analysis. By mistake, an older version of the pretest was administered that included only 23 NewWorm items. The post test consisted of 32 NewWorm items, 13 multiple choice items that covered misconceptions and standardized test items, and 17 open-end concept questions for a total of 106 points. We excluded 19 items, which resulted in a possible maximum score of 62 points. Therefore, the analysis reported in this study is based on what we term 'relevant' test items. We report first the study of the ninth grade classes, then the school wide results.

Study 2A: Ninth grade control and experimental classes

The two control classes received the teacher's usual introduction to genetics during the week that the two experimental classes used BioLogica. One of the experimental classes received an introductory lesson prior to using BioLogica but the other did not.

Analysis of Pre-test.

A t-test was computed in order to test for pre-test differences before the implementation in the four classes at "L" high school. There was a statistically significant difference found between the four groups on the pre-test (See Table XII). Post hoc contrasts yielded that the differences found were between the "red" group and the other three groups, "yellow, "green", and "blue".

Table XII. Comparison of pretest means for Ninth Grade classes.

Class	Group	N	Pretest			F statistic	Significance (p value)
			Total number of points	Mean	Std. Dev.		
blue	Control	16	40	12.3	2.9	4.14	.009*
yellow	Control	20	40	13.1	3.5		
green	Experimental	22	40	13.6	3.7		
red	Experimental	14	40	9.0	6.1		

*significant at $p<.05\,\alpha$

Analysis of Post-Test.

We pooled the two control groups and the two experimental groups since it was the same teacher and the same level of students in all four groups. The pre-test score was used as a covariate in order to take into account the pre-test differences. An Analysis of Variance (ANOVA) was computed in order to determine whether there were any statistically significant differences on the total post-test score between the experimental and control groups. For this ANOVA the dependent variable was total post-test score, the independent variable was group (control versus experimental), and the pre-test score was used as a covariate. There was a statistically

significant difference found between the control and experimental groups on the post-test with the experimental group scoring higher. See Table XIII.

Table XIII. Comparison of Control and Experimental Groups on Post test mean with Pretest covariate.

Group	N	Pretest		Post test			F statistic	Significance (p value)
		Total number of points	Mean	Total number of points	Mean	Std. Dev.		
control	36	40	12.7	62	25.7	11.0	6.269	.003*
experimental	36	40	11.8	62	31.1	10.6		

*significant at $p<.05\alpha$

Analysis of Selected Post-Test Items.
ANOVAs were computed in a series of analyses in which we selected items from the entire post-test and pooled them according to the concept that they were assessing. The pooled items assessed students' understanding of: inheritance, dominance, sex linkage, genotype and phenotype, monohybrid, dihybid, and pedigree. For each of these analyses, the pre-test scores were used as a covariate to allow us a more accurate measure of post-test differences between the control and experimental groups.

There was a statistically significant difference found between the control group and the experimental group on two concepts: monohybrid and dihybrid. For both concepts the experimental group scored higher than the control group. A summary of results on monohybrid and dihybrid can be found Table XIV. No statistically significant difference was found on inheritance, dominance, sex linkage, genotype and phenotype or pedigree. See Appendix C for a summary table of this data.

Table XIV. Concepts with Statistically Significant Differences between the Control and Experimental Groups.

Dependent variable	**Mean** **Control** **Experimental**	**F statistic**	**Significance (p value)**
monhybrid	7.444 9.167	5.798	.005*
dihybrid	1.250 1.694	4.261	.018*

*significant at $p < .05\alpha$

ANOVAs were computed in a series of analyses in which we selected items from the post-test and pooled them according to the type of reasoning that they were assessing. The pooled items assessed students on seven different types of reasoning skills: cause-to-effect within generations (cew), effect-to-cause within generations (ecw), cause-to-effect between generations (ceb), effect-to-cause between generations (ecb), Punnett square, structure, and process. For each of

these analyses, the pre-test scores were used as a covariate to allow us a more accurate measure of post-test differences between the control and experimental groups.

There was a statistically significant difference found between the control group and the experimental group on six of the seven types of reasoning: cause-to-effect within generations (cew), cause-to-effect between generations (ceb), effect-to-cause between generations (ecb), Punnett squares, structure and process. The experimental group means were higher than the control group means on the following types of reasoning: cause-to-effect within generations (cew), cause-to-effect between generations (ceb), Punnett squares, structure, and process. The control group means were higher than the experimental group means on effect-to-cause between generations (ecb). It should be noted, however, that the Levene's Test of Equality of Error Variances (which tests whether the error variances of the two groups are equal, was significant for cause-to-effect within generations, and effect-to-cause between generations. Thus, we must interpret the results regarding the two groups' performance on these scores with caution. These results, therefore, should be interpreted with caution. These results are summarized in Table XV. No statistically significant difference was found at $p = .05\alpha$ on effect-to-cause within generations. See Appendix C for a summary table of the data for "L" High School.

Table XV. Types of Reasoning with Statistically Significant Differences between Control and Experimental Groups

Dependent variable	**Mean** **Control** **Experimental**	**Levene's Test**	**F statistic**	**Significance (p value)**
cause-to-effect within generations (cew)	4.472 5.083	.019*	5.734	.005*
cause-to-effect between generations (ceb)	5.958 7.056	.904	5.951	.004*
effect-to-cause between generations (ecb)	5.792 5.694	.011*	3.572	.033*
Punnett square	5.250 5.653	.987	4.280	.018*
structure	2.917 4.139	.192	3.985	.023*
process	.667 1.194	.059	7.420	.001*

*significant at $p < .05$

Study 2B: Analysis of all class levels at "L"

Although the ninth grade was the only class level with controls, three other teachers used BioLogica with their biology students. Analysis of the pre and post test scores for those classes are compared in Table XVI.

Table XVI. Pre and post test scores for different class levels

Type of Class	N	Pretest		Post test			F statistic	Significance (p value)
		Total Number of points	Mean	Total number of points	Mean	Std. Dev.		
ninth grade	36	40	11.8	62	31.1	10.6	26.941	.000*+
business biology	6	40	8.8	62	17.1	5.5		
college prep biology	49	40	13.1	62	31.3	12.6		
honors biology	30	40	21.6	62	47.3	9.4		

*significant at $p<.05\alpha$

+Levene's Test for Equality of Error Variances is significant at $p<.05\alpha$

Post hoc contrasts (see Table XVII) reveal that the honors biology class significantly outperforms the other classes on the post test and that the business biology class significantly underperforms the other classes. It is interesting that the ninth graders did as well as the college prep biology.

Table XVII. Tukey HSD post hoc contrasts across class types.

Type of Class (mean)	Type of Class	Mean	Standard Deviation	Significance (p value)
ninth grade (31.1)	business biology	17.1	5.5	.025*
	college prep biology	31.3	12.6	1.000
	honors biology	47.3	9.4	.000*
business biology (17.1)	ninth grade	31.1	10.6	.025*
	college prep biology	31.3	12.6	.018*
	honors biology	47.3	9.4	.000*
college prep biology (31.3)	ninth grade	31.1	10.6	1.000
	business biology	17.1	5.5	.018*
	honors biology	47.3	9.4	.000*
honors biology (47.3)	ninth grade	31.1	10.6	.000*
	business biology	17.1	5.5	.000*
	college prep biology	31.3	12.6	.000*

*significant at $p<.05\alpha$

Study 3: Intensive 3-day field test at "MV"

The primary purpose of the intensive field test was to find bugs in the software. A secondary purpose was to investigate the usefulness of log files for helping us understand what and how students learn when using BioLogica. Log files were intended from the start as a research tool to help us (a) assess what learners understand, (b) how that understanding changes and (c) how they use BioLogica. We're trying to develop a fine-grained understanding of how different students learn with BioLogica. However, we also want to identify variables that can be quantified and used as covariates in analyzing pre and post test gains. In this study we looked for ways to design and process log files so that they provide the data needed in a format that can be analyzed with less effort.

We decided to use the data collected during the intensive field test for this study because it is a small population that completed nearly all of the BioLogica activities under relatively controlled circumstances. We reasoned that because it was not a classroom situation, learning gains could be attributed more fully to BioLogica use. Therefore, connections between log files and learning should be more direct and less influenced by other classroom variables.

In the intensive field test conducted at site "MV" in April 2001, 24 middle school and high school students were paid to use all the BioLogica activities over a 3-day period and to take the pre and post tests and surveys. Students used computers in two rooms. In one room there were 10 students working individually on iBooks. In the other room 14 students worked on iMacs; 10 students worked in 5 pairs and 4 students worked individually. They used an earlier version of the same set of activities used in the "L" high school implementation. The Invisible Dragons activity was used for the first time on the last day of the field test. The composition of the pre and post tests is shown in Table XVIII.

Table XVIII. Test composition for "MV"

Sites Type of items	MV Pre 4/01	MV Post 4/01
NewWorm	0	32
Multiple choice	9	11
Open-ended	8	17

A total of 387 log files were generated, of which 306 or 79% were usable files. Files that contain no date or data beyond the user name and the activity name are unusable. They are caused either by the user deciding not to run the activity or by the activity script crashing.

An excerpt from an unprocessed log file is shown in Table XIX. The full version of the log file can be found in Appendix D. The user's name has been changed for privacy reasons. Each log contains the following kinds of data: user name, date & time, question & answer, and actions (basically mouse clicks), automatically identified with XML tags such as <user> or <date>. The XML tags enable the logs to be processed for different purposes.

Table XIX. Selections from an unprocessed log file generated by the Introduction activity.

```
- <log>
  <user>julia</user>
- <question>
  <date>2001.04.17.21.46.01 04/17/01 | 21:46:01</date>
  Good job! You've created your first dragon. How would you describe it? Does anything surprise you
about this dragon? Type your description in the box below.
  <answer>this dragon has no arms or wings and also no fire.</answer>
  </question>

- <question>
  <date>2001.04.17.21.46.18 04/17/01 | 21:46:18</date>
  Quite a variety of dragons here! Dragons apparently come in many different colors. What OTHER
differences do you notice? Type your answer in the box below.
  <answer>There are many types of dragon . I pictured dragons to look like the frist male on my
list.</answer>
  </question>

- <question>
  <date>2001.04.17.21.57.40 04/17/01 | 21:57:40</date>
  What did you notice as you examined their chromosomes? Type your answer in the box below.
  <answer>You could have the same chromosomes but if you had the wrong mixer you could kill the
dragon.</answer>
  </question>

- <action>
  <date>2001.04.17.22.03.17 04/17/01 | 22:03:17</date>
  Matched comparison dragon after 28tries.
  </action>

- <question>
  <date>2001.04.17.22.20.19 04/17/01 | 22:20:19</date>
  Explain why the dragons you created in this activity look different from one another.
  <answer>they have diffrent types of chromosomes that make diffrent cells.</answer>
  </question>

  </log>
```

Analyzing log files for BioLogica use

The objective in analyzing log files was to identify and characterize variables we could use to quantify and compare the use of BioLogica across students. What kinds of data are useful in explaining the results of the pre and post tests? Given the variability across implementations, one of the first questions is: Which of the activities did the students complete? Related process questions are: How much time did learners spend on an activity? How mindful was their interaction with the activity? What challenges did they accomplish? How easily? Which questions did they attempt? Were their answers correct? How thoughtful were their answers? We cannot yet answer these questions with a reasonable degree of reliability or effort.

Even from the brief excerpt in Table XIX, it is clear how time-consuming it is to make sense of

unprocessed log files. Therefore, we chose to analyze the logs of a small subset of learners who represented different levels of prior knowledge and performance on the post test.

We began by analyzing the pre and post test scores to identify the subset of learners. We plotted the post test scores against the pre test scores (See Figure 3).

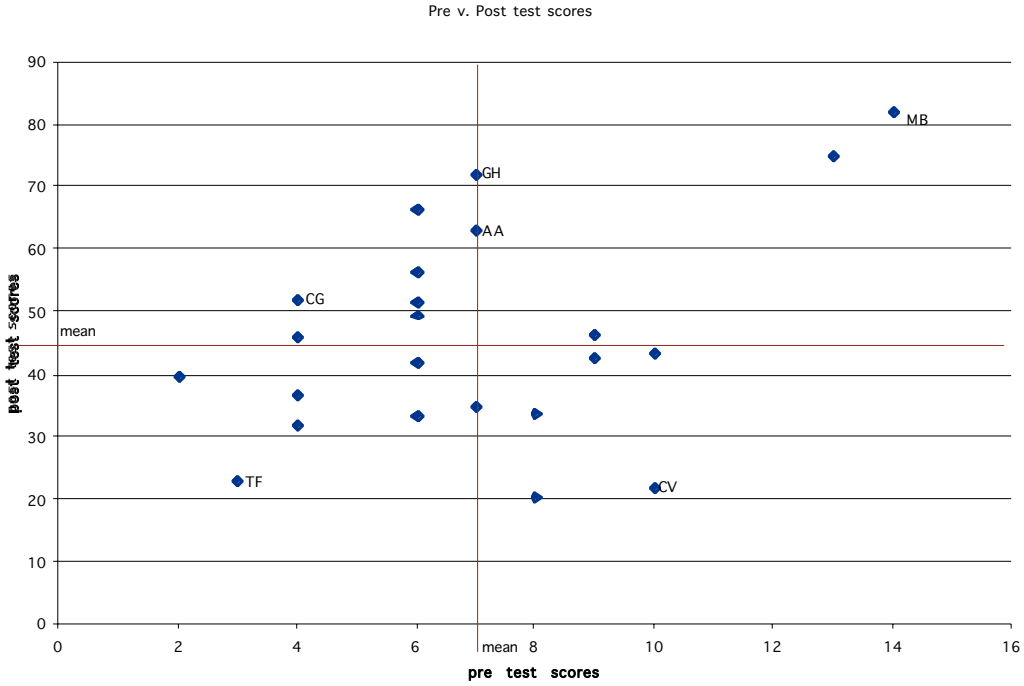

Figure 3. Scatter plot of post test scores against pretest scores for field test participants.

We chose to examine the log files for five students, labeled TF, CG, CV, GH, and MB. MB was chosen for high scores on both the pre test and the post test. GH was chosen to represent a successful student with an average pre test score. TF was chosen because he came in with a low pre test score and didn't perform very well on the post test. CG and CV were chosen as a result of observations made while they were working through BioLogica activities. CG is dyslexic and reads with difficulty, but worked diligently on activities. CV scored well on the pre test but was observed mindlessly clicking through, rather than reasoning through, an activity.

The next challenge was identifying the logs of these students. Students could type in any name or variation of that name when they logged in to use an activity. In this study we were able, with reasonable yield and effort, to connect the log files to individual learners.

To determine which activities the learner used, we created a computer program to strip out the XML tags and generate a file of tab-delimited text that could be imported into a database or spreadsheet, which could then be sorted. Table XX shows the activities used by student GH in the order in which they were used, when, and for how long. There were also four empty logs indicating bugs or aborted launches.

Table XX. Activities used by GH

Activity	date	time	length of time
Introduction	4/17/01	21:33:53	0:31
Rules	4/17/01	22:19:44	0:53
Rules	4/17/01	23:06:57	0:01
Meiosis	4/18/01	1:04:17	0:18
Inheritance	4/18/01	1:32:31	0:03
Monohybrid	4/18/01	1:37:40	0:14
Mutations	4/18/01	1:53:28	0:27
Mutations2	4/18/01	2:25:38	0:20
HornsDilemma	4/18/01	21:21:08	0:03
HornsDilemma	4/18/01	21:26:00	0:09
Dihybrid	4/18/01	21:44:23	0:17
Plates	4/18/01	23:17:18	0:05
Sex-Linkage	4/18/01	23:25:27	0:21
		Total	3:42

The initial approach to analyzing the log files was to read through the logs of one student at a time in order to develop a sense of that student's path through BioLogica, what challenges and questions were difficult or easy, and how engaged the student seemed to be. By the time we had read through the logs for all five students, we began to see patterns of data that might be relevant and quantifiable. In addition to the total-time, candidates for further analysis included the proportion of questions answered, the proportion of multiple choice questions answered correctly, how many attempts it took to solve a challenge, how much time it took to solve the challenge, the length of time each explanation was displayed. To date we have been able to investigate only a few of these possibilities.

We created another computer program to process the logs to calculate how many questions the user encountered and answered in an activity and how many words were in their answers. We learned that the data on number-of-questions-encountered and the number-of-words-per-answer were a better measure of the state of development of the activities than of the learner's engagement with the activity.

We continued to search for some measure of the quality of a learner's interaction with the activity. We knew from classroom observations that time spent with an activity does not automatically equate to time engaged with an activity. Similarly, interaction with the activity in terms of actions taken or number of mouse clicks does not necessarily equate to mindful interaction. We conceptualized this quality measure as an index of interaction and operationalized it as the length of time a user spends with an activity divided by the number of actions taken by the user. A low index of interaction indicates many actions in a short period of time, while a high index of interaction indicates a few actions in a long period of time. Crashes also affect the index of interaction. If a user has been through part of the activity before, the tendency is to quickly click through to reach the new portion of the activity.

We learned that our data logging is not yet systematic enough to produce meaningful numbers for index of interaction. The median index of interaction varied over the 11 logged activities from .08 to 2.0. This is primarily a reflection of the variation across activities in terms of what

gets logged, how many challenges and questions were posed, etc. For a given activity there does appear to be a range of productive interaction with outliers at both extremes. One extreme is the very low index of interaction calculated for CV who clicked mindlessly through the activities, playing BioLogica as if it were a game. The other extreme, a very high index of interaction, was calculated for CG, the dyslexic but diligent learner, who took a very long time, even with assistance, to read the instructions and information in BioLogica, but who could then reason through the activities very well. Because of the length of time required for each activity, this learner did not complete all the activities. Since learners may use different numbers of activities, an index of interaction averaged across all activities cannot be used for comparison across learners.

Analyzing log files for learner's models and understanding

We were able to make the log files easier to read with text harvests, like that shown in Table XXI. Text harvests grew out of requests from teachers for a way of assessing students' work with BioLogica. A report generator strips out all the XML tags and other entries leaving all the questions in an activity and the answers entered by the learner.

Table XXI. Text harvest from processed log file.

Introduction roberto <04/17/01 | 22:26:40>
Q: Quite a variety of dragons here! Dragons apparently come in many different colors. What OTHER differences do you notice? Type your answer in the box below.
A: Some dragons have wings, some have arms, legs, both, or none at all. Others have different shaped tails, some have horns or none, some breathe fire or not, and they have different body shapes.
Q: What did you notice as you examined their chromosomes? Type your answer in the box below.
A: The dominant alleles usually gave them the characteristics that they have.
Q: In particular, how do the chromosomes of male and female dragons differ? Type your answer in the box below.
A: The male and female chromosomes respond differently to the dominant allele.
Q: Why do you think you can't match both dragons to the comparison dragon? Type your answer in the box below.
A: I don't think the female dragon has the genes for the yellow color dragon.
Q: What do you think 'phenotype' means? Tell us in your own words and give us an example.
A: Phenotype is the physical characteristics of a person who gets it from the genes. An example would be a girl with wavy brown hair, big brown eyes, short stature, wing shaped eyebrows, heart shaped face, etc...
Q: What is the connection between genotype and phenotype? That is, how does genotype relate to phenotype? Tell us in your own words and give us an example.
A: They relate to each other because they involve chromosomes and alleles. The genes from the genotypes give the physical characteristics of someone, this is the phenotype. Ex. TT this genotype gives the phenotype tall genes to someone
Q: Explain why the dragons you created in this activity look different from one another.
A: They have different colors, shapes, sizes, legs, horns, tails, wings.
Q: Where are a dragon's genes located? Click all that apply.
A: chromosomes alleles DNA
Q: What did you know about phenotype and genotype before you started this activity?
A: Phenotype are the physical characteristics Genotype are the letters of the genes
Q: What do you think you learned while working with this introductory activity?
A: About the different male and female characteristics
Q: What question(s) do you currently have about phenotype and genotype?
A: none at the moment

Text harvests make it easier to assess one measure of a learner's models and understanding of genetics. They are much easier to read than log files and can be coded and scored just like pre

and post tests. However, they don't enable us to infer much about learners' learning or reasoning strategies. We need to consider the learner's performance on the challenges as well. Taken together, the pre and post tests, text harvests, and log files can be used to describe a learner's path through BioLogica. An example using the data generated by GH is included in Appendix E. The log files of different students can be compared for a given activity. In another paper we will present the results of Study 3 as case studies of individual students with cross case comparisons.

For this paper, the important findings of Study 3 relate to what we've learned about creating and analyzing log files. We present these as an annotated wish list of requirements for data logging.

- Consistent user names.
  We now use a log-in procedure that requires students to use pull down menus to select first, the class name, and then, their name from the class roster. This should greatly increase our yield of useful logs as well as the usefulness of those logs in understanding how BioLogica use influences genetics learning.
- Relevant variables such as school, class level, and gender
  These are also included when configuring the class roster for the log-in procedure. We may also decide to include variables such as ESL for non-native speakers or IEP for those on individualized education plans.
- Systematic logging of performance on challenges
  This will require analysis of existing activity scripts and the logs they generate to develop the fine-grained data needed and ways of post-processing that data for easier analysis.
- ID tags for questions embedded in scripts
  Each question embedded in a script should have an identifier that enables a person or program to pull all the responses to that question from the database.
- Precode questions and tasks
  Each question, action, and task could be associated with one or more data analysis codes to facilitate content and statistical analysis.
- Fine-grained time stamps and calculations of time spent working on pedagogical elements
  This also requires consistent use of pedagogical elements within scripts in order to calculate time spent on activities such as reading explanations or manipulating alleles.

The task of mining useful data from log files continues as we try to develop ways of generating logs that are more systematically related to research questions and variables and ways of analyzing logs in less labor intensive ways that are useful to teachers and students as well as researchers.

Discussion

In the three studies presented we have tried to illustrate how fruitful it can be to conduct classroom-based research with interactive curricula. We have also illustrated how difficult such studies can be. Conducting the kinds of studies presented in this paper requires the systematic development of a series of tasks that engage learners, scaffolding that helps them learn how to accomplish the tasks, ongoing assessments to help us understand what and how students are

learning, and how to use technology to support and facilitate all of the above. It's harder than we thought.

Study 1: "M" high school, March 2001

In Study 1 students in a Survey of Biology course and a College Biology course used a version of BioLogica without scaffolding or metacognitive prompts. In addition, there were considerable technical difficulties in getting the activities to run properly. Both of these may have contributed to the findings: there were significant differences between the control group and the experimental group in both classes; however, for the Survey of Biology course, the control group scored higher than the experimental group. The experimental group in the College Biology course outperformed the control group on the post test. These findings may indicate that the students in the Survey of Biology needed the kinds of scaffolding we put in later versions of BioLogica more than the students in the College Biology course. This is a tentative hypothesis we can put to the test in future research.

Study 2: "L" high school, May 2001

The ninth grade students in the experimental groups used a more stable version of BioLogica that included scaffolding in about half of the activities. We are therefore delighted to see that the experimental group outperformed the control group. When we examine the kinds of items on which they outperformed the control group, we see that it is primarily on the easier type of reasoning from cause-to-effect which is a good starting point. We don't know from the data logs how many of the activities students actually completed because of the difficulty of matching logs to individual students. The experimental groups also outperformed the control groups on the structure and process items. This is encouraging because parts and processes are what students manipulate in many BioLogica activities. They also are critical aspects of model-based learning (Buckley, 2000; Buckley & Boulter, 2000; Gobert, 2000; Gobert & Buckley, 2000). Their performance on Punnett squares is also encouraging because of the frequency with which Punnett squares are used as reasoning tools in BioLogica activities.

Although the improvement in the results between Studies 1 and 2 is encouraging, we don't yet feel that we have a complete model of model-based learning with hypermodels. The contrasting data between the different classes is most likely a result of student differences; however, there are other possibilities as well.

The results may indicate a need for teacher professional development in using technology in classrooms. In some work presently underway by our group, we are looking closely at teacher variables and classroom culture variables in order to determine what aspects of teachers' pedagogy needs to change in order to implement technology successfully in the classroom (Horwtiz et al, 2001 (MAC grant); (Christie, 2002b). Effecting change in teachers' pedagogical strategies with technology and affecting classroom culture are the two next areas we as a community need to address so that it can make full use of technology in classrooms. We look forward to being able to consider the data from log files to better understand how learners use BioLogica and how that influences what they learn.

This paper has focused on the cognitive aspects of model-based learning, but there are individual and classroom factors influencing whether and to what extent a learner may actually engage in model-based learning. The presumed motivation for engaging in model-based learning is a desire to understand. Motivation theorists conceptualize this desire as an intrinsic love of 'learning for learning's sake' (Ames, 1992; Nicholls, 1989). It is also described as intrinsic motivation (Deci & Ryan, 1985; Lepper & Chabay, 1985; Lepper & Malone, 1987), a learning orientation rather than a performance orientation (Dweck, 1986), or an intentional learning stance (Bereiter & Scardamalia, 1989). Extrinsic motivation arising from teacher assessment is also a factor in students' engagement in classroom activities whether computer-based or not.

We acknowledge that both intrinsic motivation and the extent to which extrinsic factors may influence motivation and learning differ not only by student but also by the context and culture in which the student learns (Christie, 2002a). Christie's work on students' perceptions of learning (Christie, 1999, 2001, 2002a) has shown promising indications that motivational constructs also vary within domain and between tasks, according to the modality, e.g., text-based versus computer-based representations. Thus, students' achievement cognitions (Christie, 2001, 2002a), their understanding of the nature of models (Gobert & Discenna, 1997; Gobert, Snyder, & Houghton, 2002) and of science (Schommer, 1993) may influence not only their motivation to engage in learning but also the strategies they employ in response to the task (Songer & Linn, 1991), and ultimately, the outcomes.

Lastly, we acknowledge that classroom culture plays a highly influential role in student learning (Ames, 1992; Ames & Ames, 1984; Perkins, 1995). More often than not the role of classroom culture in learning has been excluded from research in student content learning. We believe that understandings about classroom culture are critical to our work and to any research that is situated in the classroom (Christie, 2002a). We are integrating classroom culture into our MBTL framework from an educational anthropological perspective, which views culture as a set of shared beliefs held by and acted upon by participants in that culture (Spindler 1987; Spradley, 1972).

We have student surveys from the implementations presented in this paper that are intended to help us articulate and investigate the impact of learners' views of the nature of models and their views of science learning. It will be a first step in elaborating our model of model-based learning to include these other factors.

Future research

As we demonstrated in Study 3, particularly, there is much we still need to invent and to investigate in developing effective interactive curricula that support model-based learning with hypermodels. We continue to grapple with the tension between gathering as much data as possible about each learner's understanding and the priorities and time constraints of the classroom. We must also fine-tune the assessments and activities for better alignment both with each other and with the model-based teaching and learning framework.

There is considerable research begging to be done on assessment using interactive items. We're not talking about adaptive testing but rather about using interactive items to assess a learner's

problem solving or model-building skills in a hypermodel environment such as BioLogica. We question how well traditional test items assess the state of learners' models of phenomena and plan to invesitgate the affordances that the hypermodel environment brings to this question.

References

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*(3), 261-271.

Ames, C., & Ames, R. (1984). Systems of student and teacher motivation: Towards a qualitative definition. *Journal of Educational Psychology, 76*(4), 535-556.

Bereiter, C., & Scardamalia, M. (1989). Intentional learning as a goal of instruction. In L. B. Resnick (Ed.), *Knowing, learning and instruction: Essays in honor of Robert Glaser* (pp. 361-392). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bransford, J., Sherwood, R., Vye, N., & Rieser, J. (1986). Teaching Thinking and problem solving. *American Psychologist, 41*(10), 1078-1089.

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

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, 22*(9), 895-935.

Buckley, B. C., & Boulter, C. J. (1999). Analysis of Representations in Model-Based Teaching and Learning in Science. In R. Paton & I. Neilson (Eds.), *Visual Representations and Interpretation* (pp. 289-294). Liverpool, England: Springer.

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.

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63*(1), 1-49.

Christie, M. (1999, April 20). "*We understood it more 'cause we were doin' it ourself:*" *Students self-described connections between participation and learning*. Paper presented at the American Educational Research Association, Montreal, Canada.

Christie, M. (2001, April 26, 2001). *Portraits of academic beliefs*. Paper presented at the Poster presented at the annual meeting of the New England Educational Research Organization (NEERO), Portsmouth, NH.

Christie, M. (2002a, April 1-5, 2002). *The role of classroom culture and learning contexts in achievement beliefs*. Paper presented at the Paper to be presented at the annual meeting of the American Educational Research Association (AERA), New Orleans, LA.

Christie, M. (2002b). *The role of culture and context in achievement beliefs: An exploration*. Paper presented at the Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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.

Deci, E., & Ryan, R. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Academic Press.

deKleer, J., & Brown, J. S. (1983). Assumptions and ambiguities in mechanistic mental models. In A. L. Stevens & D. Gentner (Eds.), *Mental Models* (pp. 155-190). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*, 1040-1048.

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

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

Gobert, J., Snyder, J., & Houghton, C. (2002, April 1-5). *The influence of students' understanding of models on model-based reasoning*. Paper presented at the To be presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

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.

Hickey, D. T., Kindfield, A. C. H., Horwitz, P., & Christie, M. A. (1991). Integrating curriculum, instruction, assessment, and evaluation in a technology-supported genetics learning environment.

Hickey, D. T., Kindfield, A. C. H., Horwitz, P., & Christie, M. A. (1999). Advancing educational theory by enhancing practice in a technology-supported genetics learning environment. *Journal of Education, 181*(2), 25-55.

Hickey, D. T., & Kindfield, A. C. H. D. (1999). *Assessment-oriented scaffolding of student and teacher performance in a technology-supported genetics environment*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Hickey, D. T., Wolfe, E. W., & Kindfield, A. C. H. (1998a, April). *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, San Diego.

Hickey, D. T., Wolfe, E. W., & Kindfield, A. C. H. (1998b, April). *Assessing learning in a technology-supported genetics environment: Evidential and systemic validity issues*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego.

Horwitz, P., Schwartz, J., Kindfield, A. C. H., Yessis, L. M., Hickey, D. T., Heidenberg, A. J., & Wolfe, E. W. (1998). *Implementation and evaluation of GenScope™ learning environment: Issues, solutions, and results*. Paper presented at the Third International Conference of the Learning Sciences, Charlottesville, VA.

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

Kindfield, A. C. H., & Hickey, D. T. (1999). *Tools for scaffolding inquiry in the domain of introductory genetics*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Kozma, R., Jones, T., Wykoff, J., & Russell, J. (1992). *Multimedia, Multiple Representations, and Mental Models in Chemistry*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Lepper, M., & Chabay, R. W. (1985). Intrinsic motivation and instruction: Conflicting views of the role of motivational processes in computer-based education. *Educational Psychologist, 20*(4), 217-230.

Lepper, M., & Malone, T. (1987). Intrinsic motivation and instructional effectiveness in computer-based education. In R. E. Snow & M. J. Farr (Eds.), *Aptitude, Learning and Instruction III: Conative and Affective Process Analyses* (pp. 255-286). Hillsdale, NJ: LEA.

Monaghan, J., & Clement, J. (1994). *Factors affecting the efficacy of computer simulation for facilitating relative motion concept acquisition and visualization*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Nicholls, J. G. (1989). *The competitive ethos and democratic education*. Cambridge, MA: Harvard University Press.

Norman, D. A. (1983). Some observations on mental models. In A. L. Stevens & D. Gentner (Eds.), *Mental Models* (pp. 7-14). Hillsdale, NJ: Lawrence Erlbaum Associates.

Perkins, D. (1995). *Smart schools: Better thinking and learning for every child*. New York: The Free Press.

Rouse, w. B., & Morris, N. M. (1986). On Looking Into the Black Box: Prospects and Limits in the Search for Mental Models. *Psychological Bulletin, 100*(3), 349-363.

Schauble, L., Glaser, R., Raghavan, K., & Reiner, M. (1991). Causal models and experimentation strategies in scientific reasoning. *Journal of the Learning Sciences, 1*(2), 210-238.

Schommer, M. (1993). Epistemological development and academic performance among secondary students. *Journal of Educational Psychology, 85*(3), 406-411.

Songer, N. B., & Linn, M. C. (1991). How do students' views of science influence knowledge integration? *Journal of Research in Science Teaching, 28*(9), 761-784.

Spindler , G. D. (1987). *Doing the ethnography of schooling*. Prospect Heights, IL: Waveland Press, Inc.

Spradley, J. (1972). *Culture and cognition*: Chandler Publishing Company.

Stewart, J., & Hafner, B. (1991). Extending the conception of problem solving. *Science Education, 75*(1), 105-120.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology, 24*(4), 535-585.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*(1), 3-118.

Appendix A: Description of BioLogica Activities

(1) Introduction—What do dragons look like and why?

Introduction guides the user through BioLogica's representations of chromosomes, genes, and alleles and stresses the connection between genotype and phenotype. As learners use pulldown menus in the Chromosome View to change allele combinations, they see changes in the Organism View of BioLogica's dragons.

Goals:

Help user become familiar with the operation of the interface by using pull down menus in the Chromosome View to make changes in the appearance of the organism in the Organism View.
Enable user to link the representations in BioLogica with genetics concepts of traits, genotype, phenotype, chromosomes, genes and alleles.

Key Concepts:

Genotype determines phenotype.
Definition of traits, genotype, phenotype, chromosomes, genes and alleles

Levels/Views: Organism View, Chromosome View

(2) Rules--How do genes affect appearance?

The Rules activity is divided into three sub-activities that can be invoked independently via a menu selection. These are: 1) Traits, which deals with the four autosomal traits of dragons genetics – horns, wings, number of legs, and shape of tail; 2) Firebreathing, which introduces a sex-linked trait, and 3) Colors, which are polygenic and pleiotropic traits. one of the color genes contains a recessive lethal allele. Traits and Firebreathing are essential introductions to dominance, recessive, incomplete dominance, and sex-linkage. At this point in its development, BioLogica does not build on the Colors subactivity.

Goals

Enable user to explore the variable effects of changing alleles and allele combinations on appearance.
Provides opportunity for user to develop and use note-taking representations and search strategies.
Introduce concepts of dominance, recessive, incomplete dominance, and sex-linkage as well as homozygous and heterozygous.

Key Concepts

Genotype determines phenotype.
Particular allele combination produces particular trait.
Dominance/recessive/sex-linked inheritance

Levels/Views: Organism View, Chromosome View

(3) Meiosis: What do meiosis and fertilization have to do with making offspring?

Meiosis is also subdivided into three subactivities. Introduction to Meiosis, Meiosis Shuffles and Deals! and Designer Dragons. **Introduction to Meiosis** focuses on learning to use the interface and linking the representations with the concepts of gametes, meiosis and fertilization. It provides an introductory view into the process of gamete creation and the random distribution of the alleles and allows the student to inspect the alleles in each gamete and to choose combinations of gametes to fertilize. **Meiosis Shuffles and Deals!** links the representations with the names of the phases of meiosis found in textbooks. **Designer Dragons** offers students a series of challenges in the form of creating specific offspring by examining the chromosomes in the gametes from each parent and selecting those that will produce the desired phenotype in the offspring.

Goals

Introduce monohybrid inheritance stressing the equal contribution of each parent, the random shuffling of genes into gametes, and selection of gametes for fertlization.

Key Concepts

Meiosis
Fertilization
Chromosome segregation
Gamete selection

Levels/Views: Organism View, Meiosis View, Chromosome View

(4) Inheritance: What determines what the offspring look like?

Unlike the Meiosis activity where students select gametes to determine combinations of alleles that will be "passed down" to the offspring, the Inheritance activity uses the pedigree view, which emphasizes the randomness of this process in nature. It thus serves as an introduction to the role of probability in genetics. The Monohybrid activity is a useful follow-on to this one.

Goals

Examine the effect of random variation on the distribution of traits in a generation of offspring. Explore how the dominant & recessive properties of alleles determine proportion of offspring with trait.

Key Concepts

Monohybrid inheritance
Proportions

Levels/Views: Pedigree View, Organism View, Chromosome View

(5) Monohybrid: Do traits really skip generations?

Monohybrid introduces and uses Punnett Squares to help students understand how the combinatorics of meiosis and fertilization (an exhaustive count of all possible combination of parental genes) enables one to derive the probability that an offspring will possess a certain trait, and leads eventually to a prediction of the likely fraction of offspring with the trait in a sample of size n.

Goals

Examine the effect of random variation on the distribution of traits in a generation of offspring. Use Punnett Squares to understand how the dominant & recessive properties of alleles determine the proportion of offspring with a particular trait.

Key Concepts

Monohybrid inheritance
Punnett Squares
Combinatorics -> probability -> statistics

Levels/Views: Meiosis View, Pedigree View, Punnett Square, Chromosome View

(6)Mutations--What happens when you change the DNA?

This activity introduces and uses the BioLogica DNA View, which provides an expanded view of the red lines on chromosomes that represent genes. The activity introduces students to the DNA and base pairs that form particular alleles and challenges students to modify the DNA of a dragon and to observe the consequences.

Goals

Introduction to Molecular Level (DNA) of genetics
Make students aware of what mutations actually look like at the DNA level
Focus attention on some of the varieties of ways in which mutated alleles can result in alterations in phenotypes.

Levels/Views: Organism View, Chromosome View, DNA View

(7) Mutation Inheritance —How are mutations inherited?

This activity uses mutations to explore alternative modes of inheritance of traits that are controlled by genes with more than two alleles.

Goals

To support student investigations of some of the things that can happen when a gene comes in more than two varieties.

Key Concepts

Modes of inheritance, non-Mendelian genetics
Levels/Views: Pedigree View, Organism View, Chromosome View, DNA View

(8) Horns Dilemma--Can two horned parents have a hornless baby?

This activity focuses students' attention on the connection between parental genes and those of the offspring, in the context of posing a challenge that requires the student to alter a parental gene, making it heterozygous so that a homozygous recessive offspring can result.

Goals

Encourage multilevel thinking and confidence in model
Exercise effect-to-cause & cause-to-effect reasoning

Key Concepts

Inheritance of "hidden" traits

Levels/Views: Organism View, Chromosome View, Meiosis View

(9) Dihybrid

This activity explores what happens when you study the inheritance patterns for 2 traits at a time. It provides experience with dragons with traits on the same chromosome and with peas with traits on different chromosomes.

Key Concepts

Effect of independent assortment of chromosomes on the inheritance of two traits.

Levels/Views: Pedigree View, Organism View, Chromosome View

(10) Sex Linkage -- What difference does it make if a gene is on the X chromosome?

This activity begins with a review of how the X and Y chromosomes interact to produce male and female (introduced in the Rules activity). Once it is clear to the students that female dragons are XX and males are XY, the activity uses the firebreathing (recessive, X-linked) trait to help them learn how sex-linked traits are inherited.

Goals

Reinforce the XX = female, XY = male pattern.
Help students to understand how and why sex-linked inheritance differs from autosomal inheritance.

Key Concepts

Sex-linked inheritance

Levels/Views: Pedigree View, Organism View, Chromosome View

(11) Scales—What causes that Scaly Skin?

This is an advanced activity that guides the student through the process of investigating a trait – scaly skin – the gene for which has been hidden. Four questions are asked: Are scales genetically inherited? Are they recessive or dominant? Are they sex-linked or autosomal? And which chromosome are they on?

Goals

Learn to investigate how a new trait is inherited, using nothing but breeding information.

Key Concepts

Inference of mode of inheritance from statistics

Levels/Views: Pedigree View, Organism View, Chromosome View

(12) Plates

Plates is similar to scales, but with incompletely dominant traits.

Goals

The challenge is to figure out where the gene for this new trait is located. More practice with a problem that closely approximates the reasoning process of professional scientists.

Key Concepts

Investigating how new traits are inherited from statistics

Levels/Views: Pedigree View, Organism View, Chromosome View

(13) Invisible Dragons

In this activity we test the student's ability to solve a real genetics puzzle. They are presented with 2 invisible dragons and their task is to figure out the genetic makeup of this couple. They may make crosses, look at the chromosomes and even make a backcross, but all of these costs money. The players start out with $20000 in the bank. Each procedure costs money as does each wrong answer. Players make money by answering questions correctly.

The important part of the exercise is the development of strategies for getting the answers, something every geneticist has to learn. These are:

- Check the sex-linked traits first. You can usually tell what the Dad's genes are by the F1 generation.
- Do a backcross—cross a recessive offspring with either the father or mother. If the dominant trait shows up, you know that the parent has it.
- Cross 2 dominant offspring. If they produce at least one recessive child, you know that each of them carries a recessive gene.

GenScope

This is an open-ended interface to all the BioLogica functionality. It has no "story line" or monitoring functions. GenScope is essentially a BioLogica implementation of GenScope that runs on the Mac and the PC.

Appendix B: Summary Tables for Site "M" (Study 1)

Study 1A, Survey of Biology Course, Teacher "C"

Control = 14 students; Experimental = 20 students

Univariate Analysis of Variance (ANOVA) – using pretest as a covariate

Question subgroup	Dependent variable	Mean control experimental	F statistic	Significance (p value)	Number of items	Total number of points
All questions	posttest	28.357 23.850	14.164	.000*	23	59
concept	inheritance	1.786 .950	2.074	.143	3	3
	dominance	2.357 1.800	7.387	.002*	5	7
	sex-linkage	2.214 1.200	2.103	.139	4	7
	genotype/ phenotype	6.607 7.025	3.594	.039*	9	9
	monohybrid	9.607 7.575	13.850	.000*	6	17
	dihybrid	.286 .200	.049	.952	1	1
	pedigree	.500 .850	2.043	.147	2	4
type of reasoning	cew	4.357 4.650	1.251	.300	6	6
	ecw	2.750 2.625	2.264	.121	4	4
	ceb	6.107 4.700	12.607	.000*	6	11
	ecb	5.571 4.250	2.344	.113	12	19
	Punnett square	9.571 7.100	13.689	.000*•	4	16
	structure	.071 .600	6.317	.005*•	1	1
	process	.000 .000	N/A	N/A	1	1

* significant at $p < .05\alpha$

• Levene's Test for Equality of Error Variances is significant at $p < .05\alpha$

Study 1B, College Biology Course, Teacher "V"

Control = 12 students; Experimental = 25 students

Univariate Analysis of Variance (ANOVA) – using pretest as a covariate

Question subgroup	Dependent variable	Mean control experimental	F statistic	Significance (p value)	Number of items	Total number of points
All questions	posttest	37.833 38.400	3.757	.034*	23	59
concept	inheritance	2.417 1.640	4.268	.022*•	3	3
	dominance	3.667 3.760	.805	.456•	5	7
	sex-linkage	1.917 2.440	1.973	.155	4	7
	genotype/ phenotype	8.208 8.080	1.638	.209•	9	9
	monohybrid	13.583 14.100	2.806	.074+	6	17
	dihybrid	.167 .480	.740	.485	1	1
	pedigree	.500 .960	4.085	.026*	2	4
type of reasoning	cew	5.833 5.120	4.688	.016*•	6	6
	ecw	3.375 3.360	1.061	.357	4	4
	ceb	8.250 8.600	2.470	.100	6	11
	ecb	6.583 7.480	2.826	.073+	12	19
	Punnett square	13.250 13.140	1.815	.178	4	16
	structure	.417 .640	2.824	.073+	1	1
	process	.000 .000	N/A	N/A	1	1

* significant at $p < .05\alpha$

• Levene's Test for Equality of Error Variances is significant at $p < .05\alpha$

+ significant at $p < .05\alpha$

Appendix C: Summary Table for Site "L" (Study 2)

Ninth grade classes: Total number of students = 72
Control group = 36 students: Experimental group = 36 students

Summary Table of Univariate Analyses of Variance Using Pretest as a Covariate

Question subgroup	Dependent variable	Mean control experimental	F statistic	Significance (p value)	Number of items	Total number of points
All questions	posttest	30.000 34.431	5.355	.007*	52	106
concept	inheritance	5.778 6.472	1.870	.162	10	15
	dominance	3.875 4.722	2.524	.088+	8	13
	sex-linkage	1.000 1.139	1.143	.325	6	10
	genotype/ phenotype	8.208 9.028	1.527	.225	11	11
	monohybrid	7.444 9.167	5.798	.005*	8	19
	dihybrid	1.250 1.694	4.261	.018*	4	6
	pedigree	.778 .750	1.098	.339	3	6
type of reasoning	cew	4.472 5.083	5.734	.005*•	6	6
	ecw	2.569 2.750	2.630	.079+	4	4
	ceb	5.958 7.056	5.951	.004*	11	18
	ecb	5.792 5.694	3.572	.033*•	15	22
	Punnett square	5.250 5.653	4.280	.018*	4	16
	structure	2.917 4.139	3.985	.023*	7	10
	process	.667 1.194	7.420	.001*	3	6

* significant at $p < .05\,\alpha$
• Levene's Test for Equality of Error Variances is significant at $p < .05\,\alpha$
+ significant at $p < .10\,\alpha$

Appendix D: Unprocessed log file [annotations by Buckley in square brackets]

- <log>
<user>julia</user>
- <question>
<date>2001.04.17.21.46.01 04/17/01 | 21:46:01</date>
Good job! You've created your first dragon. How would you describe it? Does anything surprise you about this dragon? Type your description in the box below.
<answer>this dragon has no arms or wings and also no fire.</answer>
</question>

- <question>
<date>2001.04.17.21.46.18 04/17/01 | 21:46:18</date>
Quite a variety of dragons here! Dragons apparently come in many different colors. What OTHER differences do you notice? Type your answer in the box below.
<answer>There are many types of dragon . I pictured dragons to look like the frist male on my list.</answer>
</question>

- <action> [This should have a question tag even though it is a radio button multiple choice.]
<date>2001.04.17.21.50.16 04/17/01 | 21:50:16</date>
How many pairs of chromosomes do dragons have: 2 [This is an incorrect answer, but it was not tallied or monitored.]
</action>

- <action>
<date>2001.04.17.21.50.16 04/17/01 | 21:50:16</date>
Did you kill any dragons: No
</action>

- <question>
<date>2001.04.17.21.57.40 04/17/01 | 21:57:40</date>
What did you notice as you examined their chromosomes? Type your answer in the box below.
<answer>You could have the same chromosomes but if you had the wrong mixer you could kill the dragon.</answer>
</question>

- <question>
<date>2001.04.17.21.57.40 04/17/01 | 21:57:40</date>
In particular, how do the chromosomes of male and female dragons differ? Type your answer in the box below.
<answer>They could look the same but there chromosomes could be diffrent.</answer>
</question>

[The following is an example of an action that is a challenge task. The script counts the number of allele changes needed to change the genotype of one dragon to produce the phenotype of the comparison dragon.]

- <action>
<date>2001.04.17.22.03.17 04/17/01 | 22:03:17</date>
Matched comparison dragon after 28tries.
</action>

- <action>
<date>2001.04.17.22.04.36 04/17/01 | 22:04:36</date>
Student clicked YES, they think that the second dragon can be made to look like the comparison dragon.

</action>

- <action>
<date>2001.04.17.22.04.43 04/17/01 | 22:04:43</date>
Student clicked NO, they don't think that the second dragon can be made to look like the comparison dragon.
</action>

[looks like user changed their mind. This is also useful]

- <question>
<date>2001.04.17.22.06.18 04/17/01 | 22:06:18</date>
Why do you think you can't match them? Type your answer in the box.
<answer>they are two diffrent types.</answer>
</question>

- <action>
<date>2001.04.17.22.07.51 04/17/01 | 22:07:51</date>
Matched second comparison dragon after 14tries.
</action>

[this is the first of the open ended conceptual questions at the end of the activity.]

- <question>
<date>2001.04.17.22.16.41 04/17/01 | 22:16:41</date>
What do you think 'genotype' means? Tell us in your own words and give us an example.
<answer>It is the type in the diffrent chromosomes in the male and female.</answer>
</question>

- <question>
<date>2001.04.17.22.17.07 04/17/01 | 22:17:07</date>
What do you think 'phenotype' means? Tell us in your own words and give us an example.
<answer>i do not now</answer>
</question>

- <question>
<date>2001.04.17.22.18.24 04/17/01 | 22:18:24</date>
What is the connection between genotype and phenotype? That is, how does genotype relate to phenotype? Tell us in your own words and give us an example.
<answer>I do not now.</answer>
</question>

[This is the first of the questions that focus on the user's understanding of the activity.]

- <question>
<date>2001.04.17.22.20.19 04/17/01 | 22:20:19</date>
Explain why the dragons you created in this activity look different from one another.
<answer>they have diffrent types of chromosomes that make diffrent cells.</answer>
</question>

- <question>
<date>2001.04.17.22.22.00 04/17/01 | 22:22:00</date>
Phenotype determines genotype.
<answer>False</answer>
</question>

- <action>
<date>2001.04.17.22.22.00 04/17/01 | 22:22:00</date>
In order of decreasing size: dna = 2 dragons = 4 genes = 3 chromosomes = 1.
</action>

- <question>
<date>2001.04.17.22.22.00 04/17/01 | 22:22:00</date>
Where are a dragon's genes located? Click all that apply.
<answer>chromosomes DNA</answer>
</question>

- <question>
<date>2001.04.17.22.26.30 04/17/01 | 22:26:30</date>
Alleles are particular forms of: Click all that apply.
<answer>chromosomes genes</answer>
</question>

- <question>
<date>2001.04.17.22.26.30 04/17/01 | 22:26:30</date>
Particular combinations of alleles determine: Click all that apply.
<answer>traits chromosomes</answer>
</question>

- <question>
<date>2001.04.17.22.26.30 04/17/01 | 22:26:30</date>
Please explain your answer in the box below.
<answer>they said in the begining that all the things you see of the dragon are called traits. Im not shear that thous are the exsact words but thats what I remember.</answer>
</question>

[this is the first of the reflective questions]
- <question>
<date>2001.04.17.22.26.35 04/17/01 | 22:26:35</date>
What did you know about phenotype and genotype before you started this activity?
<answer />
</question>

- <question>
<date>2001.04.17.22.26.36 04/17/01 | 22:26:36</date>
What do you think you learned while working with this introductory activity?
<answer />
</question>

- <question>
<date>2001.04.17.22.26.40 04/17/01 | 22:26:40</date>
What question(s) do you currently have about phenotype and genotype?
<answer />
</question>

</log>

Appendix E: Learning path of GH.

Based on responses to open-ended questions on the pre test, GH began the study with a model of inheritance as something you receive from parents or grandparents and that it is a random process. "I guess it's when genes are passed down to you." The chromosome is "some sort of gene" and traits as "some quality of you". No responses were offered for fertilization, pedigree or sex-linked. This is a potentially good basis for beginning to use BioLogica in that GH recognizes genotypic entities (genes and chromosomes) as well as phenotypic entities(traits).

According to available logs, GH did not complete the Scales activity. GH seemed to have little difficulty completing most of the challenges in the activities — until encountering the pedigree view in Monohybrid. GH answered questions, could define dominant and recessive in terms of representations by upper case and lower case letters respectively, but sometimes had difficulty applying them to multiple choice questions embedded in the Rules activity. GH was able to reason with and manipulate the meiosis representation and make the connection between the representations of chromosomes and alleles in chromosome and meiosis views. GH could reason with Punnett squares but seemed to lack a strategy for making crosses and reasoning about the results in pedigree view. [Monohybrid seems to be a place where scaffolding is needed to help learners figure out and reason with this new representation.] Dihybrid and Plates logs provide further evidence of difficulties with pedigree use and reasoning. In Plates, for example, GH made just one cross, looked at chromosomes and concluded (incorrectly) that plates are recessive because "the quality didn't seem to travel through the genes often." The evidence for dominance seems to be frequency.

Responses to the questions on the post test suggest that there is some confusion about the physical relationship among chromosomes, genes, and alleles. The chromosome is now "the group of genes or alleles that give traits" and a *gene* is "a group of DNA". But *allele* is "a group of genes/DNA", which suggests that allele is not understood. Heterozygous, homozygous and autosomal alleles are understood, but dominant, recessive and sex-linked are not. Dominance seems to associated solely with the frequency with which it turns up in the offspring. For example, dominance is "the overpowering type of alleles that determine more in the offspring outcome" while recessive is "the 'weak' type of alleles that determine less in offspring". This is echoed in a later explanation "If there is complete dominance, the offspring should all have the trait." Sex linked is "when an allele is different in each of the sexes (on either the X or the Y chromosome)".

The mechanisms of inheritance (meiosis and fertilization) were defined as "the process in which DNA is taken from the parents and made into an offspring" and "when an egg from the female and sperm from the male is brought together to form an offspring", respectively. It is not clear whether this is part of a mental model of meiosis, or a text-based definition.

In the NewWorm portion of the post test, the learner was able to map from genotype to phenotype and vice versa, determine the sex of the NewWorm, and create and use Punnett squares effectively to reason about monohybrid inheritance. The facts that [1] only 1 of the 3 dihybrid questions was answered correctly without explanation and that [2] GH was able to identify the gametes and offspring in Punnett square but [3] did not associate them with the outcomes of meiosis and fertilization, suggests to me the lack of a mental model. If GH had a mental model of the chromosomes and the process of meiosis and fertilization, it would be possible to reason about dihybrid inheritance by visualizing the movement of alleles on the same or different chromosomes.

GH's ability to reason with the pedigree representation is unclear. GH seems to have forgotten or did not apply BioLogica activities in this explanation, "if having small nostrils was recessive, then the baby couldn't have them, because the parents don't. It then has to be dominant." This completely ignores the possibility that the parents could be heterozygous dominant in the trait, the focus of more than one activity. Answers to the multiple-choice questions were little changed from the pre test, just one point increase.