Tools for Scaffolding Inquiry in the Domain of Introductory Genetics[1]

Ann C. H. Kindfield[2]
Montclair State University

Daniel T. Hickey
Georgia State University

Assuming that scientific inquiry skills are ideally developed in particular domains, domain-specific tools are needed to support student inquiry. In addition to a valid model of the development of reasoning in the domain, other useful tools include technology-supported environments for conducting experiments and engaging in inquiry, and tools for assessing inquiry skills. Our research concerns two such tools in introductory genetics. These tools were developed within a three-year effort to design and implement a system for evaluating student learning in *GenScope*™ (Horwitz & Christie, in press; Horwitz, Neumann, & Schwartz, 1996), an open-ended exploratory software tool that students can use to investigate a variety of phenomena in genetics.[3] All work described here was carried out in close collaboration with Paul Horwitz and colleagues now at the Concord Consortium.[4] One tool that we developed, the *NewWorm Assessment*, was developed around a sophisticated model of reasoning in the domain, which included several different dimensions with each dimension varying according to the complexity of reasoning involved. The NewWorm was designed to scaffold student performance across increasingly complex problems as defined by the reasoning dimensions. Our theoretical analysis of the complexity within dimensions was supported by extensive validity inquiry conducted in a small number of the classrooms where the NewWorm instrument was being used to assess learning in GenScope and comparison environments (Hickey, Wolfe, & Kindfield, in press) and through multifaceted Rasch analysis of the performance of over 500 GenScope and comparison students on the NewWorm Assessment (Hickey & Kindfield, 1999; Hickey, Kindfield, Wolfe, & Heidenberg, 1999; Kindfield, Hickey, & Yessis, 1999). Midway through the project we developed a second tool, the *Dragon Investigations*, a set of supplementary activities designed to bridge a potential gap between world of GenScope dragons (the primary organism of interest in the software tool) and the world of the NewWorm (the novel organism of interest in the assessment). In essence the Dragon Investigations encouraged students to organize the reasoning skills they were practicing within GenScope around the dimensions of reasoning characteristic of the domain and targeted by the NewWorm Assessment, thus scaffolding student performance from computer-based tool to paper-and-pencil assessment.

In this paper, we will address the design and performance of the NewWorm Assessment and Dragon Investigations in completed work and consider how these currently paper-and-pencil tools might be incorporated into the new "scriptable" *BioLogica*™, the next generation of GenScope (Horwitz, 1999) to provide more immediate scaffolding to support student construction of sophisticated reasoning/inquiry skills in genetics.

[1] Paper presented as part of the symposium *Thinking Like a Scientist: Issues in Teaching and Learning Scientific Inquiry* at the Annual Meeting of the American Educational Research Association, Montreal, April 19-23, 1999. This research was supported by the National Science Foundation Applications of Advanced Technology Program Grant RED-95-5348. The authors acknowledge the substantial contributions of Edward W. Wolfe, Alex Heidenberg, Brian Davis, Kirsten Mixter, and Krista Herron to this effort, and thank the many administrators, teachers, and students who made this research possible.

[2] Address correspondence to Ann Kindfield, 325 Zion Road, Neshanic Station, NJ 08853-3011 (ann.kindfield@montclair.edu)

[3] For more information on the GenScope program, including software downloads, visit http://genscope.concord.org/

[4] The Concord Consortium is a nonprofit educational research and development organization in Concord, MA that focuses on developing and disseminating innovative instructional technology. For more information visit http://www.concord.org/.

Tool Design

The NewWorm Assessment

For the purpose of evaluating student learning as a result of interacting with the GenScope Learning Environment, we developed an assessment instrument called the *NewWorm Assessment*. The NewWorm Assessment uses a fabricated species, NewWorms, to systematically explore student understanding of introductory genetics concepts and reasoning. Constraints on the design of the NewWorm Assessment were the need to (a) use a paper-and-pencil format, (b) satisfy both (proximal) research and (ultimate) dissemination goals, (c) assess multi-level reasoning, (d) compare GenScope and non-GenScope users, and (e) assess a broad range of student populations. The NewWorm Assessment addresses these constraints by using a species whose genetics mimics that of GenScope dragons, but is novel and understandable to both GenScope and non-GenScope students, and by using questions that progressively shift from simple to complex forms of reasoning all in a paper-and-pencil format.

All NewWorm items can be classified along two primary dimensions: (1) Domain-general Reasoning Type (cause–to-effect, effect-to-cause, and process reasoning) and (2) Domain-specific Reasoning Type (within-generations and between-generations). In general, reasoning within generations is easier than reasoning between generations and reasoning from causes to effects (from genotypes to phenotypes[5]) is easier than reasoning from effects to causes (from phenotypes to genotypes) (Stewart, 1988; Stewart & Hafner, 1994) which in turn is easier that reasoning about processes. Reasoning about processes can be further divided into reasoning about process inputs and outputs versus reasoning about process events with the former generally being easier than the latter (Kindfield, 1994). Typical introductory genetics instruction focuses on within- and between-generation, cause-to-effect reasoning which often can be accomplished through pattern matching and application of algorithms with little to no understanding of the relationship between Mendelian inheritance (the transmission of traits from parent to offspring) and the underlying process of meiosis (one of two processes that determines *how* those traits are transmitted). To the extent that process reasoning is dealt with in typical introductory instruction, it is almost exclusively confined to reasoning about inputs and outputs (Kindfield, 1994). GenScope was designed to support the development of reasoning in all of these categories and thus all categories were represented in the NewWorm Assessment.[6] Table 1 displays each item reasoning type category along with brief descriptions of example problems from the NewWorm Assessment.

TABLE 1 HERE

In addition to these primary dimensions of reasoning, items can also be distinguished according to one, two, or three secondary reasoning dimensions—the particular genetics involved, the explicitness of provided information, and/or the type of information used/sought as elaborated in Table 2. Within these dimensions we also anticipated a hierarchy of difficulty such that for example autosomal problems would

[5] An organism's genotype for a particular characteristic is the organism's genetic make-up (e.g., TT vs. Tt vs. tt) for that characteristic and its phenotype is its observable appearance for the characteristic (e.g., tall vs. short).

[6] In the NewWorm Assessment, the processes of interest were meiosis and fertilization, both of which typically contribute to generational change and thus fall into between-generation domain-specific reasoning. Within-generation processes like transcription and translation were not dealt with in the GenScope curriculum or the NewWorm Assessment.

be easier than X-linked problems, problems dealing with monohybrid crosses would be easier than those dealing with dihybrid crosses, problems with explicitly provided information would be easier than those for which certain information was implicitly provided, and categorical problems would be easier than those requiring probabilistic reasoning which in turn would be easier than those requiring short answers.

TABLE 2 HERE

The sense in which the NewWorm Assessment itself scaffolded student performance across increasingly complex problems lies in the order of problem presentation. The pretest version of NewWorm began with what theoretically were the simplest items, that is, within-generation problems—cause-to-effect followed by effect-to-cause. The remainder of the pretest consisted of between-generation items ranging from monohybrid, cause-to-effect problems to input/output-reasoning process problems. Table 3 displays the order of problem types on the pretest. In addition to the pretest items, the posttest also included more difficult items in order to capture expected increases in domain reasoning following instruction. These between-generation items range from effect-to-cause, X-linked, monohybrid problems to event-reasoning process problems as delineated in Table 4. The posttest was administered in two parts, A and B, with most of the pretest items appearing in Part A and the new, more difficult items appearing in Part B.

TABLES 3 & 4 HERE

Over the course of multiple semesters of data collection, the NewWorm Assessment was revised on several occasions in order to better elicit the desired reasoning types. In total, four different versions of NewWorm were used with some common items across all four versions. Versions 0, 1, and 2 were used in the GenScope ’97 implementations and Version 3 was used in the GenScope ’98 and ’99 implementations (see Hickey, Kindfield, Wolfe, & Heidenberg, 1999 for implementation details). Version 0 varies most markedly from the remaining versions while Version 3 is virtually identical to its precursor, Version 2, varying only in minor wording changes and reordering of individual questions within some item types. All four versions were based on the same underlying design parameters.

The Dragon Investigations

About midway through our collaboration with the GenScope design team, we all realized a need for curricular enhancements that would encourage the development of domain reasoning skills that we believed to be among the appropriate outcomes of instruction and thus targeted on the NewWorm Assessment. The GenScope design team made substantial revisions to the software and continued developing and refining curricular activities, most of which consisted of 1-3 page “puzzle” exercises that typical students could complete within a single class period. Table 5 lists most of activities that were provided to the implementation teachers during the study, and teachers were strongly encouraged to develop their own activities and share them with others.[7] Meanwhile, in keeping with contemporary perspectives on assessment and instruction (e.g., Frederiksen & Collins, 1989; Paris & Ayers, 1994;

[7] A number of 40-hour teacher development workshops were held during the course of the study. Several of the participants in the implementation research described here were recruited from or otherwise participated in these workshops.

Wiggins, 1993; Wolfe, Bixby, Glenn, & Gardner, 1993) we began developing ways to help students learn the domain reasoning skills that we were simultaneously developing an assessment system to document. The ultimate outcome of the curricular part of our effort was a set of paper-and-pencil activities known as the *Dragon Investigations* that used the familiar GenScope dragons to scaffold domain reasoning. Each activity focused on a particular aspect of domain reasoning, and included both a student worksheet and a teacher version that consisted of an answer key and a detailed explanation of the domain reasoning covered. Our goal was to create very focused class discussions around difficult concepts by building on the teacher's and students' shared, simplified understanding of the domain as represented by the dragon genome and the genetics embedded in GenScope. The ultimate set of 11 activities was carefully sequenced across increasingly complex aspects of inheritance and increasingly expert kinds of domain reasoning following the same theoretical notions described for the NewWorm Assessment.

TABLE 5 HERE

We also arranged the Dragon Investigations and a subset of GenScope computer activities into six curricular units organized around domain reasoning concepts. As shown in Table 6, the six units were *Introduction, Basic Inheritance, DNA & Meiotic Events and Inheritance, Two-gene Inheritance, Alignment and Crossover, and Reasoning about Inheritance.* Each unit included a statement of the overall learning goal, a description of the relevant readings and activities from conventional biology texts and curricula, and a description of activities and learning goals for each of 2-5 GenScope computer activities and 1-3 Dragon Investigations. A package including a teacher guide and a packet of student worksheets was reproduced and distributed to implementation teachers partway through the main ('98) implementation year. Some of the activities in Table 5 were excluded from the package because they were either redundant or divergent relative to the domain reasoning concepts represented by the NewWorm assessment. Thus this revision represented at least some degree of "narrowing" of the curriculum to focus on the learning outcomes that we were attempting to capture with our assessment practice. Keep in mind however that those learning outcomes represent a broad range of domain reasoning skills as indicated in Tables 3 and 4.

As with the NewWorm Assessment, the Dragon Investigations alone or as embedded in the six-unit curriculum "package" scaffolded student performance across increasingly complex problems through their organizational sequence which followed the same sequential structure as the NewWorm as indicated by the unit structure delineated in Table 6. The Dragon Investigations provided a second level of scaffolding by familiarizing students with the "look and feel" of the NewWorm Assessment.

TABLE 6 HERE

Tool Performance

The NewWorm Assessment

Scoring

Through the 1996-1997 school year, a research assistant at Educational Testing Service scored completed assessments; subsequent assessments were scored by graduate research assistants at Georgia State University. Of the 87 individual items on the (posttest) assessment, 24 of them required some sort of interpretation in order to score. On 14 of the 24 items, scores were dichotomous (right or wrong) and the other 10 items were given either no, partial, or full credit. Interrater reliability on these 24 items averaged .86. Several of the most difficult items had very low reliability because only a handful of students were able to answer them correctly. Additional reliability data were provided by the scaling procedure and are described below.

Scaling

Student scores were analyzed using multi-faceted Rasch scaling (Linacre, 1989). This latent-trait modeling procedure locates each assessment item and each individual's score on a single linear scale. This provides an estimate of the relative difficulty of each item and the relative proficiency represented by each student's assessment performance on a common metric. This method also yields data about the precision (i.e., standard error) and reliability of the entire scale as well the degree to which each individual's and each item's pattern of scores fit the expectation of the latent-trait model. These latent-trait analyses were used to provide formative information as the instrument was developed and refined. That is, scores were scaled with Versions 0, 1, and 2 of the instrument, and the analyses were used to identify items that required revision.

A potential problem exists because we collected data at different times using different versions of the instrument. In addition we added more difficult items to the posttest version of the as described earlier. Because some items were unchanged across pretest and posttest and across versions, it was possible to analyze all of the scores together, thereby placing every item and every individual on the same linear continuum using standard equating procedures.

Item scale scores and validity data

The first stage of the analysis of student performance (Hickey et al., 1999) was scaling the posttest scores to derive difficulty indices for each item. These difficulty indices were then used to validate our theoretically derived assumptions about the primary and secondary dimensions of reasoning described earlier, as represented by the relative difficulty of the different items.

Primary dimensions of reasoning. Turning first to the primary dimensions of reasoning, Figure 1 shows the relative difficulty of the different items that were included to assess domain-general (cause-to-effect, effect-to cause, and process) reasoning and domain-specific (within-generation and between-generations) reasoning. The differences displayed in Figure 1 explicitly validate our assumptions about the different primary dimensions of reasoning. Specifically, reasoning within-generations is easier than between-generations, and cause-to-effect reasoning is easier than effect-to-cause, which in turn is easier than process reasoning.

FIGURE 1 HERE

The information captured in Figure 1 is useful for interpreting learning gains (see Hickey et al., 1999 for student performance data). Consider, for example, that the difference between the algorithmic cause-to-effect reasoning and the more expert effect-to-cause reasoning is roughly two logits of the six-logit range of ability in this sample. Similarly, the difference between within-generation reasoning and between-generation reasoning is roughly 1.5 logits. One particular advantage of this approach for communicating learning outcomes is the ability to exemplify specific levels of proficiency by referencing items with a difficulty equal (in logits) to that proficiency. In other words, a student's standing in the scale at posttest can be interpreted as an indication of the types of reasoning that the student is capable of post instruction and a change in student standing pre to post from –2 to 0 logits in this sample would indicate a qualitative shift from within-generation, cause-to-effect reasoning ability to between-generation, effect-to-cause reasoning ability.

Secondary dimensions of reasoning. Inasmuch as our expectations regarding the primary dimensions of reasoning were borne out as displayed in Figure 1, we must be careful to consider the types of items being compared according to secondary reasoning dimensions. This is because items may be more or less difficult in relation to one another for different underlying reasons. For example, the 14 cause-to-effect, within-generation items represented by one of the closed squares in Figure 1 include:

- six autosomal simple dominance items, only four of which are explicit (see Table 2),
- four autosomal incomplete dominance items, only two of which are explicit

- two X-linked, simple dominance items, both of which are explicit, and
- two items dealing with the chromosomal basis of sex determination.

The 12 effect-to-cause, within-generation items represented by the other closed square in Figure 1 include:

- six autosomal simple dominance items, only four of which are explicit,
- four autosomal incomplete dominance items, only two of which are explicit, and
- two X-linked, simple dominance items, both of which are explicit.

Thus dominance relationships and the explicitness of provided information could be confounding the Figure 1 comparison between cause-to-effect and effect-to-cause reasoning within generations. The downside of refining our analyses is that as we classify items in terms of primary and secondary reasoning dimensions, the number of items in each category gets smaller. The upside is a "truer" comparison along the dimensions of interest.

Since the number of possible comparisons is quite large and not all comparisons are as interesting as others, we will illustrate the refined analysis with two sets of comparisons. First, Figure 2 shows the comparison between domain-general reasoning type and chromosome type while "controlling" for domain-specific reasoning type, dominance relationship, and explicitness of information provided. Each item considered in this comparison is a within-generation, simple dominance, explicit item. The number in parentheses with each chromosome type represents the number of items of that secondary reasoning type in the comparison. What Figure 2 makes salient is that (a) when only domain-general reasoning is being compared, cause-to-effect (C-to-E) reasoning is easier than effect-to-cause (E-to-C) reasoning (by about the same margin as displayed in Figure 1) and (b) at least in the context of within-generation, explicit reasoning about simple dominance, chromosome type is irrelevant to difficulty.

FIGURE 2 HERE

Figure 3 similarly shows the comparison between domain-general reasoning type and chromosome type while "controlling" for domain-specific reasoning type, dominance relationship, and explicitness of information provided. In this case however, each item considered is a *between*-generation, simple dominance, explicit item. In addition, the type of information sought varied from categorical (Cat) to probabilistic (Prob) to short answer explanation (ShAns). One important difference between the derivation of Figures 2 and 3 is that all relevant within-generation items were considered whereas in Figure 3 only a subset of the potentially relevant between-generations items were considered because some of the entire set included additional potentially confounding secondary dimensions such as diagram interpretation. Figure 3 illustrates that (a) again when only domain-general reasoning is being compared, cause-to-effect reasoning is generally easier than effect-to-cause reasoning but (b) in contrast to within-generation reasoning, chromosome type does impact difficulty at least in the context of cause-to-effect reasoning. In addition, within effect-to-cause reasoning, providing adequate explanations for categorical responses is substantially more difficult than making adequate categorical responses.

FIGURE 3 HERE

Further comparisons that will explore relative difficulties of other secondary reasoning dimensions across primary reasoning dimensions are currently underway. In addition, other data supporting the substantive and structural validity of our assessment practice (following Messick, 1994; 1995; Shepard,

1993) were derived from think-aloud protocols and retrospective interviews collected during the second year of the project (reported in Hickey, Wolfe, & Kindfield, in press).

Item fit indices. Item fit indices show how well the relative difficulty of the various items conformed to the expectations of the Rasch model. Values of the standardized fit indices approximate a standard normal curve when data follow these expectations. That is, we would expect 95% of the items to fall within ± 2. Items with fit indices outside of this range contain more error variance than is explained by the Rasch model. Of the 87 items, 60 items (68%) had a standardized infit statistic within ± 2 (and 69 items, or 79%, within ± 3.0 SD). Note that these percents are higher than those based on a standard normal curve, indicating slight misfit in some of the item responses. Additional analyses aimed at understanding the reasons behind misfit items are also currently underway.

Item separation. The Rasch modeling confirmed that the different items on the NewWorm represented a broad range of proficiency. The separation index for the items (a measure of the spread of the estimates relative to their precision) was 14. According to Fisher (1996) this means that the precision of our assessments allows us to differentiate between 20 statistically distinct strata of item difficulties.[8] This is supported by the fact that the chi-square test of the null hypothesis that all item difficulties are equal is statistically significant [χ^2 (86) = 14,565, $p < .005$].

The Dragon Investigations

GenScope was implemented in a variety of classrooms over the course of three academic years, 96-97, 97-98, and 98-99. These academic years correspond to the so-called ’97, ’98 and ’99 implementations (Hickey & Kindfield, 1999; Hickey et al., 1999). These classrooms varied according to the population of students served, ease of access to computers, and teacher familiarity with genetics and/or computer technology among other things. Since the Dragon Investigations were designed in part in response to our experiences with the earliest implementations, the ’97 implementations necessarily did not utilize them. The Dragon Investigations were utilized to varying degrees in the ’98 and ’99 implementation classrooms. Only the ’99 implementation was carried out in such a way as to systematically explore the role of the Dragon Investigations in student performance. Thus we will focus on this particular implementation here.

The ’99 implementation was carried out in three classrooms at a suburban/rural school that served relatively advantaged students. This implementation was carefully designed to address three unresolved issues from the previous implementations. We will focus on one of these issues here, namely the specific impact of the Dragon Investigation activities on NewWorm performance. To some degree, we expected the Dragon Investigations to compromise the evidential validity of the NewWorm assessment. We expected the familiarity with the NewWorm item format that was provided by the Dragon Investigations to provide some advantage for the GenScope students relative to the comparison students.

Three classrooms at this implementation site served a single pool of technical track (i.e., non university-bound) and learning disabled students. The course was called ABC Biology and roughly half of the students in all three classrooms were identified as having learning or behavioral disabilities. Ms. H taught two of the three classes and implemented the GenScope curriculum in both of her classrooms. Ms. H was a first-year teacher, and had participated in the GenScope research (primarily scoring assessments

[8] Strata = [(separation index + 1) * 4]/3

and evaluating curricular activities) during the previous year while she was a science education graduate student. Ms. F, an experienced general biology teacher, taught the third class which served as a non-GenScope comparison classroom. Ms. F was provided with a detailed summary of the reasoning concepts assessed in the GenScope curriculum and the NewWorm assessment. She was encouraged to do her very best using the methods that she normally used (lecture/worksheets/textbook/discussion) to help her students develop the targeted domain reasoning skills during roughly the same number of class periods as the GenScope classrooms.

In order to address the impact of the Dragon Investigations on NewWorm performance, Ms. H's first period class completed 15 GenScope computer activities (and no Dragon Investigations) over the roughly 25 class periods devoted to genetics. In contrast, her second period class completed only 10 of the GenScope computer activities, but completed 6 Dragon Investigations as in-class activities in lieu of the related computer activities. Thus, one group of students had roughly one third of their computer-based activities replaced by paper-and-pencil activities designed to teach specific aspects of domain reasoning and to familiarize students with the format of the NewWorm Assessment.

Figure 4 shows the reasoning gains in the three classrooms in the '99 implementation. The gain in Ms. F's comparison classroom (triangles) was a modest 0.83 logits; in contrast the gain in Ms. H's GenScope classroom that did not use the Dragon Investigations (squares) was an impressive 2.14 logits. Most impressively, the gain in the GenScope classroom that used the Dragon Investigations (circles) was 2.67 logits, the largest of any classroom in the three years of the study. The gains in both of the GenScope classrooms were significantly larger than the gain in the comparison [$F(1,37) = 9.10$, $p = .005$, and $F(1,38) = 14.02$, $p < .001$, respectively]. The differences in the gains in the two GenScope classrooms was not statistically significant [$F(1,37) = 2.24$, $p = .143$].

FIGURE 4 HERE

Given Ms. H's knowledge of the GenScope curriculum and her substantial education in the area, and our own continued refinements to the GenScope curriculum, this was one of the most successful implementations undertaken. While Ms. H had become very familiar with the content covered in GenScope and targeted in the NewWorm Assessment as a graduate research assistant the previous year, this was also her first year as a teacher. Fortunately, we also had about as valid a comparison population as is established in classroom-based instructional research. Given the validity of the comparison pairing and our close observation of the implementation, these results provide conclusive evidence that the GenScope learning environment is substantially more effective than the typical conventional learning environment that it would replace—at least in terms of the sort of domain reasoning skills assessed with the NewWorm.

Specifically with regard to the relative impact of the Dragon Investigations, these findings support our conclusion that these activities presented a small and very acceptable degree of compromise to the NewWorm's evidential validity. The (non-significantly) smaller gain in the GenScope classroom that did not complete the Dragon Investigations suggests that these activities do provide some help with the NewWorm, but that this help is limited to familiarity with item formats. We would expect to see a much larger difference in the two classrooms if the Dragon Investigations had more fundamentally compromised performance on the NewWorm (by reducing the problem complexity to the degree that they could be solved more algorithmically). Because the organism, genotypes, and phenotypes of the two instruments (and the format of some of the problems) is entirely different, it appears that the Dragon Investigations had precisely the desired effect—developing transferable domain reasoning skills.

Inquiry in the Context of GenScope

Thus far the focus of the work presented here has been on *reasoning* in the domain of genetics. So what of *inquiry* per se? As stated initially, we assume that scientific inquiry skills are ideally developed in particular domains. These skills are described differently by different people (e.g., Jungck & Calley, 1985; Peterson & Jungck, 1988; White and Frederiksen, 1997) but generally include the ability to formulate/ask questions and to seek answers to those questions through a variety of steps/means. The ability to conduct reasonable inquiry in the domain of interest both supports and is supported by the ability to reason in the domain—to understand the objects of inquiry, to formulate and pursue *relevant* questions, etc. In genetics, reasoning from effects to causes and about processes, two of the targeted domain-general reasoning types, are part and parcel of legitimate inquiry activity in the domain. From one perspective then, we can view the GenScope work to date as supporting and assessing students' development of "foundational" reasoning skills that will ultimately contribute to students' ability to conduct reasoned and reasonable domain-specific inquiry.

This perspective, as well as the bulk of our current practice, however largely ignores the potential of student inquiry contributing to the development of domain-specific reasoning skills. Inasmuch as the computer-based "puzzles", the Dragon Investigations, and the NewWorm Assessment support student exploration of very targeted questions, they do not systematically encourage students to formulate their own authentic questions about genetics nor do they elicit the kind of sustained inquiry that is more characteristic of scientific investigation. Ideally, we think that both aspects of the mutual influence between domain-specific reasoning and inquiry need to be addressed and believe that tools such as GenScope are most powerful when they are used to support learning as students engage in inquiry that has meaning to them (e.g., How did I end up with brown eyes when both of my parent have blue eyes? What's the meaning of this genetic counseling report?). To this end, we are currently planning a collaboration with Paul Horwitz and colleagues at The Concord Consortium and Peter Kindfield and selected teachers in Community School District Two, Manhattan to use GenScope, and the next generation of GenScope called *BioLogica*, to support the development of genetics reasoning and inquiry skills among students in project-based classrooms.

BioLogica*,* software that is currently being developed at The Concord Consortium (Horwitz, 1999), embodies additional aspects of inheritance relative to the broader life science curriculum and adds powerful scripting capabilities. This will allow for at least three substantial changes to student, teacher, and researcher use of the BioLogica (GenScope) environment: (1) existing computer-based "puzzles" and paper-and-pencil activities as well as new investigations will become scripted "interactivities" that can more readily encourage the development of reasoning and inquiry skills including the exploration of authentic questions and sustained inquiry, (2) student work within the environment will be scaffolded by the scripting "engine" in order to provide immediate support for the development of the reasoning and inquiry skills targeted in (1), and (3) student interactions with the environment will be captured and processed by the scripting "engine" for the purposes of immediate or delayed diagnosis and feedback. The last of these will provide much richer information as to the processes and products of student learning since it will include intimate views of student work as it unfolds. And all three together will inform the continued refinement of the environment by students, teachers, and researchers alike.

Conclusions

Our original goal was to create an assessment instrument that could reasonably, systematically, and reliably capture a range of reasoning ability in the domain of introductory genetics while scaffolding student performance. A second goal that developed in the context of working toward the first was to create student activities that could scaffold student performance on the assessment without compromising its validity. The results presented here demonstrate that in large part the theoretical underpinnings of the NewWorm Assessment have been validated. Further, the supplemental Dragon Investigations appear to function as we had hoped. Inasmuch as additional analyses are in order for further refinement of the assessment instrument, the supplemental activities, and the underlying model on which they are based, the NewWorm Assessment and Dragon Investigations represent part of an innovative package for structuring and assessing student reasoning in introductory genetics. This reasoning is one side of the reasoning/inquiry “coin” that we hope to understand better through further work with BioLogica.

References

Fisher, W. P. (1996). Reliability and separation. *Rasch Measurement Transactions*, *9*, 472.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Hickey, D. T., & Kindfield, A. C. H. (1999, April). 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.

Hickey, D. T., Kindfield, A. C. H., Wolfe, E. W., & Heidenberg, A. (1999, March). *GenScope™ evaluation design and learning outcomes.* Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Boston, MA.

Hickey, D. T., Wolfe, E. W., & Kindfield, A. C. H. (in press). Assessing learning in a technology-supported genetics environment: Evidential and systemic validity issues. *Educational Assessment.*

Horwitz, P. (1999, April). *Embedding curriculum and assessment in manipulable models*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal.

Horwitz, P. & Christie, M. (in press). Computer-based manipulatives for teaching scientific reasoning: An example. M.J. Jacobson & R.B. Kozma, (Eds.), *Learning the sciences of the Twenty-first century: Theory, research, and the design of advanced technology learning environments*. Hillsdale, NJ: Lawrence Erlbaum & Associates.

Horwitz, P., Neumann, E., & Schwartz, J. (1996). Teaching science at multiple levels: The GenScope program. *Communications of the ACM, 39*(8), 127-131.

Kindfield, A. C. H. (1994). Understanding a basic biological process: Expert and novice models of meiosis. *Science Education*. *78*, 255-283.

Jungck, J. R., & Calley, J. N. (1985). Strategic simulations and post-Socratic pedagogy: Constructing computer software to develop long-term inference through experimental inquiry. *The American Biology Teacher, 47(1)*, 11-15

Linacre, J. M. (1989). *Many-faceted Rasch measurement.* Chicago, IL: Mesa Press.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23(2),* 13-23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Paris, S. G., & Ayers, L. R. (1994). *Becoming reflective teachers and learners with authentic assessment.* Washington, DC: American Psychological Association.

Peterson, N. S., & Jungck, J. R. (1988). Problem-posing, problem-solving, and persuasion in biology education. *Academic Computing, 2*(6), 14-17, 48-50.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 404-450.

Stewart, J. (1988). Potential learning outcomes from solving genetics problems: A typology of problems. *Science Education*. *72*, 237-254.

Stewart, J., & Hafner, R. (1994). Research on problem solving: Genetics. In D. Gabel (Ed.) *Handbook of research on science teaching and learning* (pp. 284-300). New York: Macmillan.

White, B. Y., & Frederiksen, J. R. (1997). *The ThinkerTools Inquiry Project: Making Scientific Inquiry*

Accessible to Students. (Available from Center for Performance Assessment, Educational Testing Service, MailStop 18-E, Rosedale Road, Princeton, NJ 08541-0001)

Wiggins, G. (1993). Assessment: Authenticity, context, & validity. *Phi Delta Kappan, 75*, 200-214.

Wolfe, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17,* 31-74.

Table 1: Primary dimensions of reasoning represented by items in the NewWorm assessment.

		Domain-General Dimension of Reasoning (Novice ↔ Expert)		
		Cause-to-effect	Effect-to-cause	Process Reasoning
Domain-Specific Dimension of Reasoning (simple ↔ complex)	Between-generations	**Monohybrid inheritance I**: given genotypes of two parents, predict genotypes and phenotypes of offspring	**Monohybrid Inheritance II**: given phenotypes of a population of offspring, determine the underlying genetics of a novel characteristic	**Punnett Squares** (input/output reasoning): describe Punnett Squares in terms of ploidy; **Meiosis-The Process** (event reasoning): given genetic make-up of an organism and the products of a single meiosis, describe the meiotic events that resulted in this set of products
	Within-generations	**Genotype to Phenotype Mapping**: given genotypes and info about NewWorm genetics, predict phenotypes	**Phenotype to Genotype Mapping**: given phenotypes and info about NewWorm genetics, predict genotypes	none (see footnote 6)

Table 2: Secondary dimensions of reasoning represented by items in the NewWorm Assessment.

Particular Genetics Involved

Chromosome Type: autosomal vs. X-linked[9]

Dominance Relationships: simple vs. incomplete dominance

# of Genes of Interest: monohybrid vs. dihybrid crosses

Physical Relationship Between Genes: unlinked vs. linked

Explicitness of Information Provided

Explicit if genotype/mapping provided
e.g., Flat body = BB or Bb and Round body = bb

Implicit if genotype inferred from dominance relationship
e.g., mouth can be oval or split and oval allele is dominant to split allele

Type of Information Used/Sought

Categorical: involving a limited set of non-probabilistic options
e.g., yes/no/maybe; simple/incomplete

Probabilistic: involving probabilistic reasoning

Short answer: involving brief explanations

Diagram Interpretation/Generation: involving diagrammatic reasoning
e.g., with pedigrees; chromosomes

Definitiveness: is solution definitive or indeterminate
e.g., distribution of offspring phenotypes resulting from a dihybrid cross involving unlinked genes vs. a dihybrid cross involving linked genes for which the linkage distance is unknown

[9] Neither GenScope nor NewWorm explored Y-linked inheritance.

Table 3. Sequence of problem types and associated primary and secondary reasoning dimensions on the NewWorm pretest.

Problem Type	**Primary Reasoning Dimensions**	**Secondary Reasoning Dimensions**
Genotype-Phenotype Mapping	within generations; cause-to-effect	autosomal & X-linked; simple & incomplete; explicit & implicit; categorical
Phenotype-Genotype Mapping	within generations; effect-to-cause	autosomal & X-linked; simple & incomplete; explicit & implicit; categorical
Monohybrid Inheritance I	between generations; cause-to-effect	autosomal & X-linked; simple & incomplete; explicit; categorical & probabilistic; diagram interpretation
Dihybrid Inheritance	between generations; cause-to-effect	autosomal; simple & incomplete; unlinked & linked; explicit; categorical & probabilistic; definitive & indeterminate
Monohybrid Inheritance II	between generations; effect-to-cause	autosomal; simple; categorical & short answer
Pedigree I	between generations; effect-to-cause	autosomal or X-linked; simple; monohybrid; categorical; diagram interpretation
Meiosis: Gametes	between generations; process: input/output	categorical; diagram interpretation
Punnett Squares	between generations; process: input/output	categorical; diagram interpretation
Probability	between generations; cause-to-effect	autosomal; simple; monohybrid; explicit; probabilistic; short answer; diagram generation

Table 4. Sequence of problem types and associated primary and secondary reasoning dimensions on the NewWorm posttest (A = Posttest Part A and B = Posttest Part B).

Problem Type	**Primary Reasoning Dimensions**	**Secondary Reasoning Dimensions**
Genotype-Phenotype Mapping (A)	within generations; cause-to-effect	autosomal & X-linked; simple & incomplete; explicit & implicit; categorical
Phenotype-Genotype Mapping (A)	within generations; effect-to-cause	autosomal & X-linked; simple & incomplete; explicit & implicit; categorical
Monohybrid Inheritance I (A)	between generations; cause-to-effect	autosomal & X-linked; simple & incomplete; explicit; categorical & probabilistic; diagram interpretation
Dihybrid Inheritance (A)	between generations; cause-to-effect	autosomal; simple & incomplete; unlinked & linked; explicit; categorical & probabilistic; definitive & indeterminate
Pedigree I (A)	between generations; effect-to-cause	autosomal or X-linked; simple; monohybrid; categorical; diagram interpretation
Meiosis: Gametes (A)	between generations; process: input/output	categorical; diagram interpretation
Punnett Squares (A)	between generations; process: input/output	categorical; diagram interpretation
Monohybrid Inheritance II (B)	between generations; effect-to-cause	autosomal & X-linked; simple; categorical & short answer
Pedigree II (B)	between generations; effect-to-cause	autosomal & X-linked; simple; categorical & short answer; diagram interpretation; definitive & indeterminate
Meiosis: The Process (B)	between generations; process: events	diagram interpretation & generation
Probability (B)	between generations; cause-to-effect & process: events	autosomal; simple; monohybrid & dihybrid; explicit; probabilistic; short answer; diagram generation

Table 5. Activities included in initial GenScope curriculum.

Scavenger Hunt
To Kill a Dragon (homework)
Create-a-Dragon
Guinea Pig Problem (homework)
Horns Predictions
Rules of Inheritance (homework)
Popsicle Sticks-Mitosis
Importance of Mitosis (homework)
Popsicle Sticks-Meiosis
Importance of Meiosis (homework)
Legs (quiz)
Incomplete Dominance (homework)
Sex Determination (quiz)
Color Puzzles
Peter, Paul, & Mary
The Dragon Genome
Fire Breathing
Rules for Inheritance
Sex Linkage (quiz)
Challenge Problem: Color
A Dragon Mystery: Scales
Another Dragon Mystery: Plates
Dragon Puzzle I
Dragon Puzzle II
Sickle Cell Puzzles
Crossover: Humans
Sickle Cell Anemia (quiz)
Crossover: Blood Problems
Methemoglobin
Fruit Fly Problems (quiz)
Hitchhiker's Thumb
Vocabulary List-Dragons (homework)
Human Traits
MCET tape: Is Chrissie going to get Huntington's Disease?
MCET tape: Cystic Fibrosis
Labrador Retriever Colors
"Jeopardy" Test review
Computer Problems Review

Table 6. Activities included in the revised GenScope curriculum.

Unit	Activities
Unit One: Introduction to Genetics and GenScope	Scavenger Hunt Exploration[1] Meiosis/Chromosome Window[2]
Unit Two: Basic Inheritance	Taking Data[1] Making Predictions[1] From Genotypes to Phenotypes[2] Fire Breathing[1] From Parent to Offspring I[2] Cystic Fibrosis[1] Exploration: Human Species[1]
Unit Three: DNA & Meiotic Events and Inheritance	Mutations[1] Making Babies[1] From Parent to Offspring II[2] Blood Type Activity[1]
Unit Four: Two-Gene Inheritance	Sickle Cell[1] Bronze & Gold[1] Dihybrid Inheritance I[2] Labrador Colors[1]
Unit Five: Alignment and Crossover	Making Babies II[1] Alignment and Crossover during Meiosis[2] Crossover[1] Dihybrid Inheritance II[2] From Chromosomes to Gametes[2]
Unit Six: Reasoning about Inheritance	Hitchhiker's Thumb[1] From Pedigree to Mode of Inheritance I[2] From Pedigree to Mode of Inheritance II[2] Mystery Traits[1] From Offspring to Mode of Inheritance [2]

[1] GenScope Computer Activity.
[2] Paper-and-Pencil Dragon Investigation.

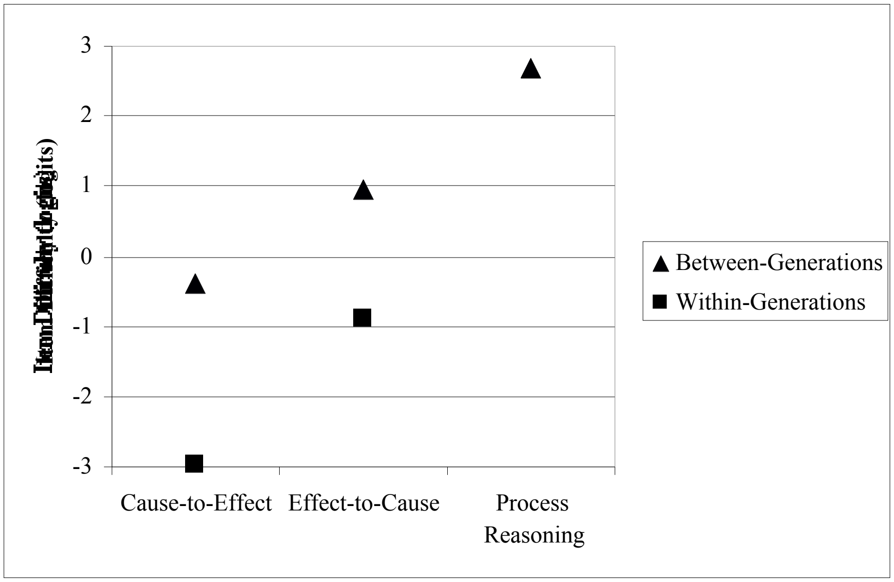
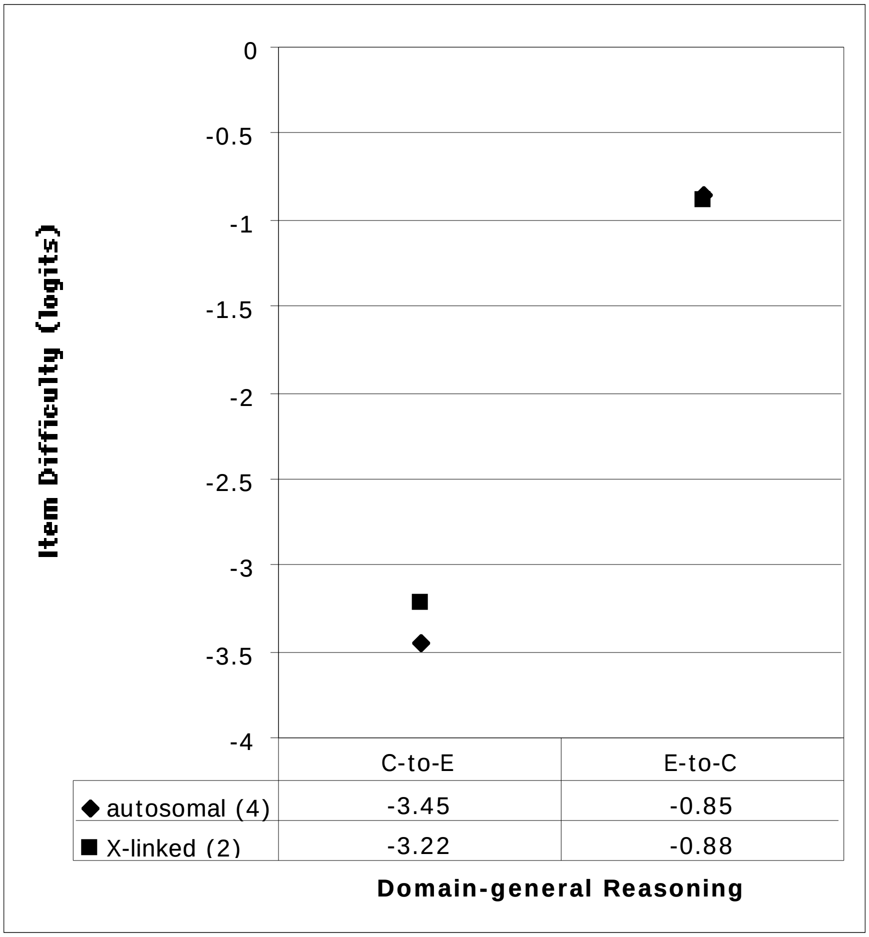
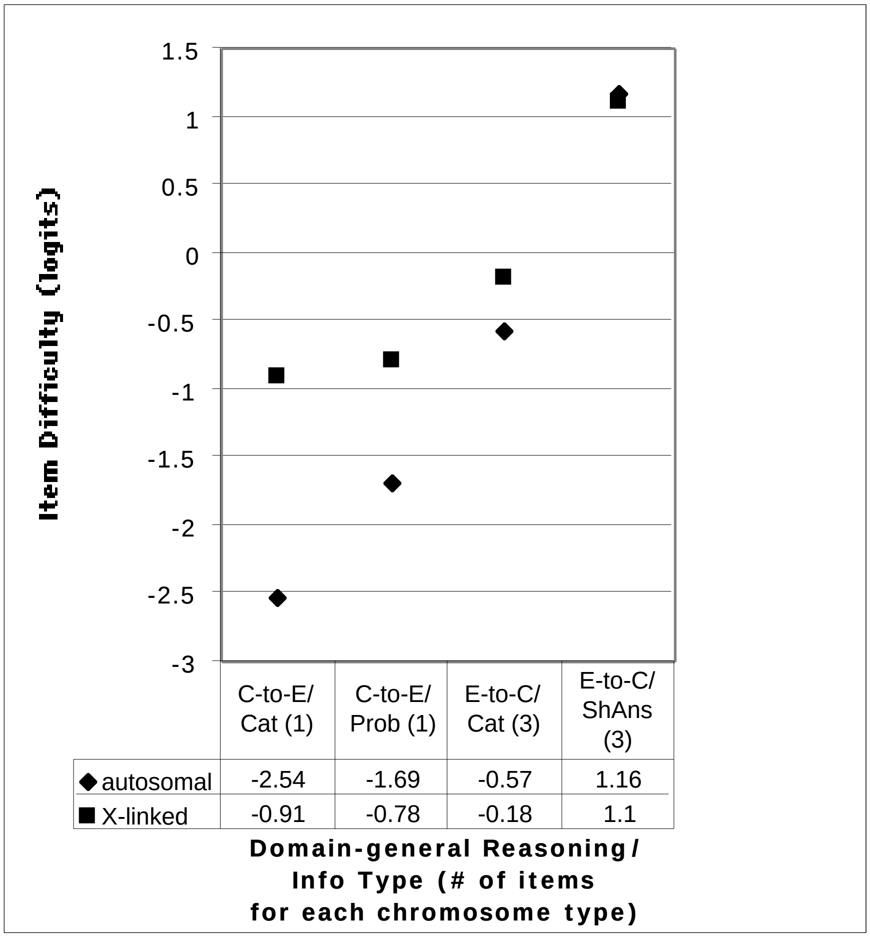


Figure 1: Relative difficulty of clusters of different items by primary reasoning dimensions.



	C-to-E	E-to-C
◆ autosomal (4)	-3.45	-0.85
■ X-linked (2)	-3.22	-0.88

Figure 2. Relative difficulty of within-generation, explicit reasoning about simple dominance.



	C-to-E/ Cat (1)	C-to-E/ Prob (1)	E-to-C/ Cat (3)	E-to-C/ ShAns (3)
◆ autosomal	-2.54	-1.69	-0.57	1.16
■ X-linked	-0.91	-0.78	-0.18	1.1

Figure 3. Relative difficulty of between-generation, explicit reasoning about simple dominance.

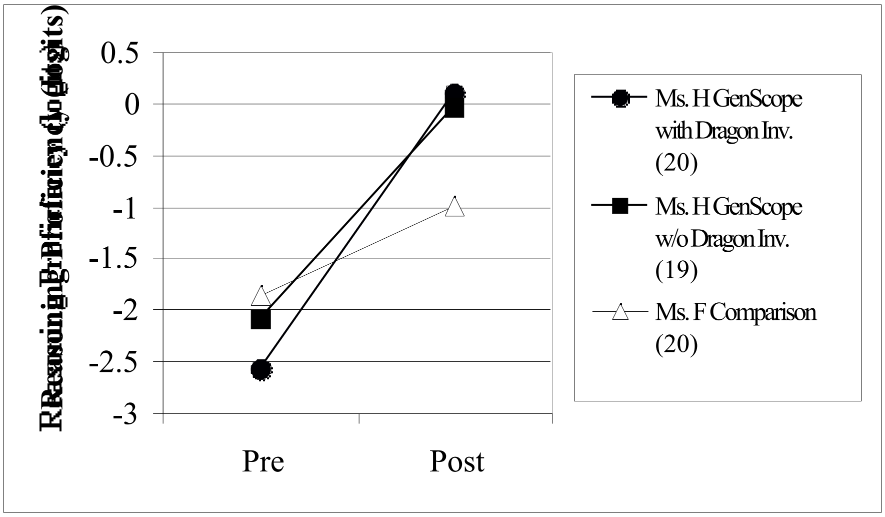


Figure 4. Reasoning gains in general biology classrooms during the '99 implementation.