Assessment-Oriented Scaffolding of Student and Teacher Performance in a Technology-Supported Genetics Environment

Daniel T. Hickey
Georgia State University

Ann C. H. Kindfield
Montclair State University

Author Notes

Presentation to the Annual Meeting of the American Educational Research Association, Montreal, April, 1999. This research was supported by the National Science Foundation Applications of Advanced Technology Program Grant RED-95-5348, and by a postdoctoral fellowship from the Center for Performance Assessment at Educational Testing Service. We gratefully acknowledge the 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. Daniel T. Hickey, Department of Educational Psychology and Special Education, Georgia State University, Atlanta, GA 30303, dhickey@gsu.edu.

Assessment-Oriented Scaffolding of Student and Teacher Performance in a Technology-Supported Genetics Environment

April, 1999

Abstract

GenScope™ is an open-ended exploratory software tool that students can use to investigate a variety of phenomena in genetics. This paper describes one aspect of a three-year collaboration between the developers and ourselves to implement, evaluate, and refine the software and associated curriculum in 40 secondary science classrooms. In response to initially disappointing learning outcomes in GenScope classrooms, we developed a set of off-computer activities known as *Dragon Investigations.* These activities consist of student worksheets and teacher aides that use the familiar GenScope organism, genome, and representation to scaffold the reasoning skills we were assessing with our *NewWorm* performance assessment. A subset of results for the larger implementation and evaluation research showed how these activities had a particularly large positive impact on reasoning gains. Reflecting and supporting new perspectives on assessment and instruction, these results document the small, acceptable compromise to evidential validity that these activities presented for our assessment practice. A new "scriptable" version of GenScope known as *BioLogica*™ should allow these and other sorts of effective learning activities and assessments to be seamlessly incorporated into the software environment and allow more comprehensive content learning and scientific inquiry skills to be learned and assessed.

With the support of the National Science Foundation, Paul Horwitz and colleagues, now at the Concord Consortium, have been developing and refining the *GenScope*™ software, developing curricular activities, and implementing those activities in classrooms (Horwitz & Christie, in press; Horwitz, Neumann, & Schwartz, 1996). [1] This software was designed for use by secondary life-science teachers and features state-of-the-art animation and dynamic interactivity. The software uses fanciful species such as dragons as well as real organisms, and dynamically links the entire range of biological organization, from the molecular level to the population level and ultimately encompassing evolution. As shown in Figure 1, the software allows students to observe and control many of the biological phenomena relevant to introductory genetics. [2]

In key respects, this application of educational computing is consistent with the policy recommendations for K-12 educational technology issued by the President's Committee of Advisors on Science and Technology (PCAST, 1997). This report urged educators to refocus the use of classroom technology away from teaching about technology itself, and towards achieving the broader goals of current educational reform efforts. Reflecting the views of many educators and researchers, the report expressed cautious optimism that contemporary constructivist models of teaching and learning will help ensure that investment in educational technology yields worthwhile returns in learning outcomes. The GenScope software is one of the most noteworthy examples of multimedia software that is consistent with these new perspectives. It is expected that employing this software in secondary life science classrooms can be shown to better help students develop the sort of domain reasoning skills called for in current science education standards (e.g., National Research Council, 1996).

This paper describes one aspect of a three-year collaboration between ourselves and GenScope's developers to implement, evaluate, and refine the software and associated curriculum—ultimately in 40 secondary science classrooms. Collectively, our central challenge was designing and deploying an assessment system that yielded the sort of evidence that policy makers are demanding, while also reflecting contemporary perspectives on instruction and assessment. Consider that, on one hand, the PCAST report recommended broadening the research on constructivist applications of technology beyond formative questions and interpretive methods. The report calls for more research that exposes

> well-explicated falsifiable hypotheses…to potential refutation through the execution of well-designed, carefully controlled experiments having sufficient statistical power to distinguish

[1] 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/.

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

> genuine effects of a relatively modest size from differences that can easily be explained as chance occurrences (p. 94).

Reflecting the immaturity of constructivist practices and strident opposition from some quarters, the commission's report advocates that such research "be conducted under conditions more typical of actual classrooms, using ordinary teachers, and without access to unusual financial or other resources, for example, or to special outside support from university researchers".

In order to carry out the sort of research advanced in the PCAST report, we needed to develop an assessment system that captured the full range of reasoning skills that GenScope ostensibly affords, while using a paper-and-pencil format that could equitably and efficiently assess learning in both GenScope and comparison classrooms. We believe that our overall evaluation and the evidence of GenScope's effectiveness that it yielded represents a worthwhile effort to carry out such research. Our evaluation and the overall learning outcomes will be considered here, but is presented in more detail in Hickey, Wolfe, Kindfield, and Heidenberg, (1998) and Hickey, Kindfield, Wolfe, and Heidenberg, (1999).

On the other hand, we were cognizant of contemporary assessment perspectives that emphasize how rigorous assessment practice can undermine students' learning of the skills and concepts being assessed. In addition to heightened awareness of the potentially negative consequences of assessment practices, contemporary perspectives highlight the ways that assessment can enhance learning. Frederiksen and Collins (1989) coalesced ideas about potentially positive consequences of assessment practices by introducing the notion of *systemic validity:*

> A systemically valid test is one that induces in the educational system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure. Evidence for systemic validity would be an improvement in those skills after the test has been in place within the educational system for a period of time (p 27).

Frederiksen and Collins characterized the aspects of a systemically valid assessment system, including the *components* of the system (a representative set of tasks, a definition of the primary traits for each subprocess, a library of exemplars, and a training system for scoring tests), *standards* for judging the assessments (directness, scope, reliability, and transparency) and *methods for fostering self-improvement* (practice in self-assessment, repeated testing, performance feedback, and multiple levels of success).

If an intended consequence of an assessment is increasing student understanding, that assessment should maximize student performance as much as possible, starting at whatever level of performance students are capable of:

> To make tests truly enabling we must do more than just couch test tasks in more authentic performance contexts. We must construct tests that assess whether students are learning how to learn, *given what they know*. Instead of testing whether students have learned how to read, we should test their ability to read to learn; instead of finding out whether they "know" formulas, we should find out whether they can use formulas to find other formulas. (Wiggins ,1993, p. 214, emphasis added).

In our opinion, these and other contemporary assessment theorists advance a lofty, but worthy benchmark for evaluating assessment practices. However, this benchmark is further elevated when combined with efforts to derive valid evidence of a program's effectiveness. To this end, we developed a set of learning activities known as the *Dragon Investigations* that use the familiar GenScope dragons to scaffold precisely the domain reasoning skills that we were attempting to assess. The following paper elaborates on our effort to balance concerns with evidential and systemic validity (described initially in Hickey, Wolfe, & Kindfield, in press), and presents evidence of our initial success.

Method

Assessment Design

Our initial challenge was devising a valid measure for assessing student proficiency in introductory genetics. Working within the constraints described above, our first year's effort yielded a pair of performance assessments known as the *NewFly* and the *NewWorm*. Both use a species whose genetics mimics that of GenScope dragons, but is novel and understandable to both GenScope and non-GenScope students. Each presents a coherent set of questions around one organism, progressively shifting from simple to complex forms of reasoning. The items were carefully sequenced to scaffold student performance across increasingly complex problems. The NewFly is somewhat more difficult, and was only used during the first round of implementation; therefore the following discussion and most of the results concern the NewWorm.

As shown in Figure 2, the initial NewWorm problems are simple, designed to be solvable by most secondary students prior to any actual instruction in genetics. These initial problems introduce the students to the organism, genome, and assessment environment. It was expected that students' success on the initial problems would provide motivation and understanding that would scaffold performance on the more difficult subsequent problems. Like the items in Figure 2, most of items called for categorical, single-word responses (or selection from multiple verbal or diagrammatic choices). However, the items

assessing more complex reasoning also asked students to explain why the categorical response was correct. Figure 3 shows an example of one such problem.

Table 1 shows how the various aspects of reasoning in introductory genetics 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 (i.e., not involving inheritance) is easier than reasoning between generations; reasoning from causes to effects (e.g., from genotypes to phenotypes[3]) is easier than reasoning from effects to causes (from phenotypes to genotypes), which in turn is easier than reasoning about processes (Stewart, 1988; Stewart & Hafner, 1994). 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). GenScope was designed to support the development of reasoning in all of these categories and thus all categories were represented in the NewWorm Assessment.[4] In addition to these primary dimensions, most items can also be distinguished according to the particular genetics involved (e.g., autosomal vs. X-linked inheritance), the explicitness of provided information, and/or type of information used/sought (i.e., categorical, probabilistic, diagrammatic, short answer, definitive vs. indeterminate).

Thus, all of the items shown in Figure 2 assess within-generation, cause-to-effect reasoning, and ask for a categorical response. The first five items in Figure 2 involve explicit information while Item 6 involves implicit information; Item 5 concerns a more complex sex-linked characteristic, while the others involve autosomes (not sex chromosomes). In contrast, the problem shown in Figure 3 is an example of between-generation effect-to-cause reasoning and asks for both categorical and short answer explanatory responses.

Data Collection and Analyses.

Over the course of our project, the NewWorm was used to document gains in genetics reasoning in 27 secondary science classes; the NewFly was used in 13 others. This included general science, general/technical life science, college prep life science, and honors life science classrooms, and serving as either GenScope or comparison classes. These classes came from eight schools in two different major metropolitan areas and were recruited for the study in various ways. This represented a diverse range of classrooms as well as a diverse group of introductory genetics learning environments. The specifics of the various populations and instructional environments will be described in the context of the results in

[3] *Genotype* refers to the genetic makeup of a particular characteristic (e.g., TT vs. Tt vs. tt), while *phenotype* refers to the observable aspects of that characteristic (plants that are tall vs. short).

the next section. Following is a description of the methods used to score and scale the completed assessments and a summary of the methods and results of efforts to support evidential validity.

Reliability

Initially, 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 assessment, 24 of them required some sort of interpretation in order to score. On 14 of the 24 items, scores were dichotomous (no credit or credit) and the other 10 items were given either no, partial, or full credit. When two graduate students scored assessments from five different classrooms, inter-rater reliability on these 24 item 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 overall score on a single linear scale. This provides an estimate of the each item's relative difficulty and each student's relative proficiency on a common metric. This method also yields data about the precision (i.e., standard error) and reliability of the entire scale as well as 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 after the second revision of the instrument, and the results were used to identify items that required revision.

We collected data at different times using different revisions of the instrument. In addition the posttest version of the instrument included more difficult items in order to capture the expected increases in domain reasoning following instruction. 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.

Structural validity

The first stage of the present analysis 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 different dimensions of domain reasoning, as represented by the relative difficulty of the different sets of items. Figure 4 shows the relative difficulty of the sets of items that address the various aspects of domain reasoning described above. The differences displayed in Figure 4 validate our assumptions about domain reasoning. Specifically reasoning within-generations is easier than between-generations, and

---

[4] Within-generation processes like transcription and translation were not dealt with in the GenScope software and are generally not included in secondary science classes.

cause-to-effect reasoning is easier than effect-to-cause, which in turn is easier than process reasoning. Other results not presented here further validated our assumptions about more specific dimensions of reasoning (e.g., categorical versus probabilistic reasoning about inheritance) and various aspects of the domain (e.g., autosomal versus sex-linked traits). (see Kindfield, Hickey & Yessis, 1999). 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).

The information captured in Figure 4 is useful for considering the proficiency gains that we observed. 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 proficiency in our sample. Similarly, within-generation reasoning is roughly 1.5 logits more difficult than between-generation reasoning. One obvious advantage of this approach for understanding and communicating outcomes is the ability to exemplify specific levels of proficiency by referencing items with a difficulty equal (in logits) to that level.

Generalizability

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). Because the percents are higher than those based on a standard normal curve, there is evidence of misfit in some of the item responses.

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 that the precision of our assessments allows us to differentiate between 20 statistically distinct strata of item difficulties.[5] 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, $p < .005$. Similarly, the scaling results confirmed that there was a broad range of proficiency represented by our sample. The separation index of 4.1 indicates that there are 6.8 statistically distinct strata of proficiency. [χ^2 (566) = 8,305, $p < .005$.] In addition, the analyses indicate that our students measures were highly reliable (latent-trait coefficient α equivalent = .94)

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

External Validity

In order to further validate the assessment content and the range of reasoning skills captured, the NewWorm was administered to students and faculty members in a university biology department. This included 2 each of from non-science major, freshman, junior, senior, and graduate biology courses and faculty. Figure 5 shows the mean score for each pair, along with the pretest and posttest scores of the four different groups of high school students in our sample. The increasingly more expert college students and faculty perform increasingly well.[6] Somewhat surprisingly, posttest performance for some groups of high school students reached the same level as the college undergraduates we sampled. Within the high school students, we see the expected declining proficiency across the honors, college preparatory, general technical, and general science students. Overall gains from pretest to posttest in the four groups of high school students ranged from .75 logits in the college prep students to 1.4 logits in the general/technical students. It is worthwhile to consider these gains in light of overall range of the scale and in light of the more specific aspects of reasoning represented by the corresponding assessment items that were summarized in Figure 4. For example, a gain of 1.4 logit is slightly less than one-quarter of the 6 logit range of the scale.

Results

Pilot Implementation

The GenScope development team had designed a variety of paper-and-pencil curricular activities when the present research was initiated and have continued refining them and developing new ones. Most of these were 1-3 page "puzzle" exercises that typical students could complete within a single class period. Roughly 40 of these activities were ultimately developed and made available to teachers who were implementing GenScope.[7]

Approximately 20 of the initial curricular activities were incorporated into two consecutive pilot implementations carried out by GenScope project personnel in general/technical biology classes in a suburban school. Students spent approximately 25 class periods working on the activities in collaborative

---

[6] The only exception is that the senior biology majors scored below the junior biology majors; however, the juniors had just completed an upper-division genetics course while the seniors had taken this course the previous year.

[7] Readers should note that additional curricular activities are being developed within the context of the *BioLogica* software that is currently being developed at the Concord Consortium. This software embodies additional aspects of inheritance relative to the broader life science curriculum and adds powerful scripting capabilities that allow activities to be incorporated into the software environment. Additional curriculum development efforts are also currently underway at Educational Development Center, a

pairs, with the assistance of the life science teachers and the 2-3 project personnel who where in the classroom every day. At the end of the instruction, students completed the NewFly assessment (NewWorm's precursor).

Results

Examination of these students' posttests revealed disappointing posttest performance (these pilot results were never scaled). Students had clearly developed the skills needed to navigate the GenScope environment and appeared to understand the relationship between the various windows and the difference between male and female chromosomes. However, they seemingly ended up with little "domain knowledge." For example, in the two classrooms in the second pilot implementation, only 20 of 44 students were able to solve cause-to-effect/between-generation problems involving autosomal inheritance (akin to, given the information in Figure 2, *Would an offspring of NewWorm 1 and NewWorm 2 have a round body? definitely yes, maybe, definitely no*) More critically none of the students could solve such problems involving the X-linked characteristic (akin to, *Would a male offspring of NewWorm 1 and NewWorm 2 have a pointed tail?*). Furthermore, these students were utterly baffled by the effect-to-cause problems like the one shown in Figure 3. We concluded that few students developed even the basic cause-to-effect between-generation reasoning ability represented by simple Punnett Square problems. Researchers like Stewart and Hafner (1994) have long argued that such problems (that are typically the extent of secondary genetics instruction) can be solved algorithmically with little or no actual knowledge of genetics.

The disappointing outcomes in light of the students' reportedly thoughtful engagement in the GenScope environment presented the issue of *transfer*. That is, were the students extracting relevant domain understanding in the GenScope environment that simply did not transfer to the (substantially different) assessment environment? After all, there were numerous transformations between the two environments, including media, organism, genome, characteristics, and social context. To test transfer, we had also administered the 44 students in the second set of pilot classrooms several short "zero-transfer" measures. These items used GenScope screen captures to assess domain reasoning in the more familiar GenScope context. The format of these items dramatically reduced the number and nature of transformations between the GenScope environment and the assessment environment.[8] This allowed us to compare each student's performance on items that used the familiar GenScope dragons, genome, and representation with corresponding NewWorm items.

---

nonprofit educational research and development organization in Newton, MA. For more information visit http://www.edc.org.

Performance on the zero-transfer measures supported our initial conclusion that students were not extracting the underlying genetics concepts from GenScope activities. We found that 22 students could solve the "zero transfer" effect-to-cause between-generation autosomal inheritance problems; this included 19 of the 20 students who had solved the corresponding NewFly assessment item. And again, none of the students could solve zero-transfer inheritance problems involving the X-linked characteristic. If students were learning the underlying domain concepts in GenScope, we should have seen greatly improved performance on the "zero transfer" items.

Conclusion and response

We concluded that whatever students were learning in the activities during the pilot implementation, they were extracting and learning few of the "invariant" aspects of the GenScope environment assumed to represent meaningful domain understanding. In response to these findings, GenScope's developers made substantial enhancements as part of their ongoing software development. In particular, new dragon characteristics were added that prevented students from viewing the characteristic's underlying genome (necessitating effect-to-cause reasoning). Additionally, the various curricular activities were refined and additional activities were developed that focused on more specific aspects of domain reasoning.

First Round of Formal Implementations.

The first formal round of GenScope implementations (where proficiency was assessed before and after instruction in both GenScope and comparison classrooms) employed software and curricula that reflected the enhancements made after the pilot implementations.

General Science Implementation at School 1.

One of the first formal implementations was conducted in a large urban school district that served disadvantaged students who by any standard were at substantial risk for school failure. The GenScope implementations in this district were carried out in the context of an (independent) district-wide initiative to incorporate genetics into the 9th grade general (or "Unified") science.

At the time of the study School 1 was under state receivership due to persistently poor performance. One of the science teachers at the school, Mr. H, was an early participant in the GenScope project. He had modest knowledge of biology and little knowledge of introductory genetics. In addition to the participation described here, Mr. H was employed as an educator/advisor to the GenScope development team. GenScope personnel were responsible for the scope and sequence of the curricula in

[8] Of course, there were still a number of transformations between the two environments, such as media and administration context.

the GenScope class, and were present in the classroom and managing the instruction during the roughly 25 consecutive class periods when genetics was covered.[9] During this time, students completed approximately 20 GenScope activities working in collaborative pairs in the computer lab. Most of the students had very little experience with computers and very little prior life science instruction.

Figure 6 shows that the GenScope students in Mr. H's class had minimal domain reasoning skills at pretest (-2.02, the solid square). After approximately 25 class periods devoted to the GenScope curriculum, the reasoning skills of the 11 students who completed both administrations of the NewWorm assessment increased 1.42 logits. Examination of individual scores revealed that most of these students were able to solve most of the between-generation cause-to-effect problems—consistent with a mean posttest proficiency of –0.6 logits.

As shown in Figure 6, the mean posttest score in the GenScope classroom was higher than mean posttest score for students in Mr. H's other two general science class (-1.26, the open square). These two classes had a similar group of students, and Mr. H used textbook, worksheets, and lectures to teach introductory genetics. The difference in posttest means just reached statistical significance [F (1,50) = 3.99, p = .051]. However, that lack of pretest data leaves pre-instructional differences unexplained, and instruction in the GenScope class was managed and delivered by individuals who were relatively more prepared to teach introductory genetics.

College Prep Biology and General Biology Implementations at School 2

The other set of implementations in the first formal round were conducted at a suburban school that served relatively advantaged students. Students in six classrooms completed the NewFly assessment before and after genetics instruction.[10, 11] This included two general/technical biology classrooms taught by the same teacher, where student participated in approximately 25 periods of GenScope instruction implemented by GenScope project staff. This also included two college prep biology comparison classes taught by one teacher, and a college prep biology class and a general/technical biology class taught by another teacher. As detailed in Hickey, Wolfe, & Kindfield (in press), the overall proficiency gain in the GenScope students was slightly smaller than in the pair of college prep comparison classrooms, and substantially larger than the gain in the other teachers' two classes. However, this apparent advantage for the GenScope curriculum over the latter two classes did not reach statistical significance [F (1,60) = 3.68,

[9] This included a Ph.D. physicist who was responsible for developing the GenScope software and a former high-school biology teacher with a Ph.D. in biology.

[10] Because there are no identical items on the two instrument, it is impossible to directly equate scores from the two sets of implementations.

$p = .060$]. Additionally, we suspected that that the GenScope students at School 2 may have devoted more class periods to introductory genetics than the students in the comparison classrooms, but we were unable to verify this.

Conclusions and Response

While the gains observed in the first round of formal implementations were better than the pilot implementations, they were still disappointingly modest. Because we did not have detailed information about the curriculum in the comparison classrooms and because the relative advantage in the GenScope classrooms was small, we were left with little evidence that the GenScope environment yielded better learning outcomes than the conventional curriculum it was designed to supplement or replace. We concluded that further enhancements to the classroom learning environment—specifically the curricular activities—were needed in order for students to realize the gains that would indicate achievement of the more sophisticated reasoning skills that we were targeting.

The impetus for our curricular enhancements came partly from extensive validity inquiry carried out alongside our assessment activity during the first round of implementations (detailed in Hickey, Wolfe, and Kindfield, in press). This investigation concluded that there was strong evidence supporting our assessment system's *evidential* validity—that is the *content*, *structural*, *substantive*, *generalizability*, and *external* aspects of validity in Messick's (1996) framework. In contrast, we concluded that there was very little evidence of *systemic* validity. In other words, it did not appear that our assessment practice had directly positive consequences for the learners. While there was some evidence of learning via the scaffolding providing within the assessment itself, and some curricular enhancements were made in response to our assessment findings, it was clear that our assessment practice did not help students develop the reasoning skills it was designed to assess.

In keeping with the perspectives on assessment and instruction described above, we began looking for ways to use our assessment practice to help students develop the domain reasoning skills that we were assessing. The ultimate outcome of our effort was a set of paper-and-pencil activities known as *Dragon Investigations*. These activities use the familiar GenScope dragons to scaffold the specific aspects of domain reasoning we were targeting in the NewWorm. Each Dragon Investigation activity focused on a particular aspect of domain reasoning, and consisted of a student worksheet and a teacher version that included both an answer key and a detailed explanation of the domain reasoning covered. Our goal was to create very focused class discussions around each aspect of reasoning by building on the

---

[11] Additionally, comparison data was collected in a comparison general biology classroom with the NewWorm instrument at School 2. These results will be presented with the other general biology results in the next section.

teacher's and students' shared, simplified understanding of the domain as represented by the dragon genome and the various GenScope windows. Within and across the ultimate set of 11 Dragon Investigations, we carefully sequenced increasingly complex aspects of inheritance and increasingly expert kinds of domain reasoning.

As shown in Table 2, we also organized the Dragon Investigations and a subset of GenScope computer activities into six curricular units organized around domain reasoning concepts. 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 for use by teachers in the second round of implementations. Some of the GenScope computer activities used in the previous implementation 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.

Second Round of Formal Implementations.

In the second round of implementations, some, but not all, of the teachers used the revised curriculum package including the Dragon Investigations. In terms of the issues in the present discussion, the results from the general science classrooms and the general/technical biology classrooms are most informative.

General Science Classrooms.

The second round included three teachers' general science classrooms from a single urban school district. This included Mr. H (who participated in the first round described above) and another teacher at his school, Ms. L, who used GenScope with her special education students. A third teacher, Ms. Q, was at a different school in the same district, and her classes ended up serving as comparisons.

Mr. H "on his own". By this round, Mr. H had completed a 40-hour GenScope teacher-training workshop and took over the instruction in introductory genetics in both of his general science classes. As the instruction was getting underway, Mr. H was given the revised GenScope curriculum described above. Mr. H used the revised GenScope curriculum to cover introductory genetics in both of his classes, but only one of them had access to the computer lab needed to complete the GenScope computer activities. Partly because of this, Mr. H relied very heavily on the Dragon Investigation activities in both classes. During the 36 days he reported devoting to introductory genetics, Mr. H worked through the GenScope curriculum guide with both of his classes. On the days when the students in one classroom

independently completed the GenScope computer activities, Mr. H would either go over the activity with the other class and provide the relevant information on the chalkboard, or would use other worksheets, text readings, or lectures, to target the same domain concepts.

Figure 6 shows that the gains in both of Mr. H's classes (the triangles) were substantial, and similar to the gain in his Round 1 GenScope class. Most notable is that the gain in the GenScope class that had access to the computers (the closed triangles) was nearly identical to the gain in the class that did not (the open triangles).

Ms. L's special education classes. Another teacher at School 1, Ms. L, initiated a similar implementation in her three general science classes. Like Mr. H, Ms. L had little experience teaching genetics. She described herself as "primarily a special education teacher" who learned "minimal" genetics content in the context of single non-science major biology course in college. Ms. L's classes served students who had been identified as either learning disabled or behaviorally disabled, and other students who were not identified as special needs but were unable to keep up or get along in the other general science classes. In all three classes Ms. L closely followed the new curriculum package and reported relying very heavily on both the student worksheets and teacher versions of the Dragon Investigations. Like Mr. H, Ms. L's access to the computer lab was limited, and just one of her classes was able to use it for two or three periods a week. As in Mr. H's classes, on the days when those students completed the GenScope computer activities, the students in the other two classrooms often completed the GenScope computer activity worksheets as a whole-class activity. Ms. L described this process as very methodically going over the key points on each of the activities, sometimes making diagrams on the chalkboard to illustrate what would be taking place on the computer.

As shown in Figure 6, we again see that the gain in the class that had access to the computers (the closed circle) is identical to the gain in the two classes that did not have access to the computers (the open circle). (The gain in the two non-computer classrooms was very similar.) Given the very low pretest scores, these gains were a quite substantial 2.3 logits, one of the largest gain we saw. However, the differences in the gains between Mr. H's two classrooms and Ms. L's three classrooms was not statistically significant, $F\,(1,49) = 2.80$, $p = .101$.

Ms. Q's "derailed GenScope" comparison classes. We compared the GenScope classrooms at School 1 to a pair of classrooms in the same urban school district in a school that served a somewhat less at-risk student population (but with a substantial proportion of non-native English speakers). Ms. Q was an experienced biology teacher who described herself as "very comfortable" teaching genetics. She had participated in the GenScope teacher-training workshop and then elected to implement the curriculum and participate in the evaluation in her two 9th grade general science classrooms. However, difficulties with the software and computer lab access led her to entirely abandon the GenScope curriculum on the second

day. She reported spending an additional 18 class periods covering genetics using a mix of lectures, demonstrations, textbook assignments, and worksheets. As shown in Figure 6, (the asterisks) these students showed relatively smaller gains in domain reasoning, about 0.40 logits. This gain was significantly less than the gain across the five 1998 GenScope classes at School 1, [F (1,82) = 8.75, p = .004].

While Ms. Q did report assigning a grade to student performance on the assessment, it seems possible that the students may not have taken the posttest as seriously as they might have if they had continued with the GenScope implementation. Ms. Q gave the answer "not very seriously" to the survey question, *How seriously did your students take the GenScope assessment?* In contrast, the other two general science teachers, like most other teachers in the study, reported that their students took the assessment "seriously". However, the posttest scores for Ms Q's classroom were accurately represented by the Rasch model.[12] This suggests that her students did, in fact, make a concerted effort on the posttest.

Summary of findings in general science classrooms. These results showed that the GenScope learning environment supported substantial domain reasoning gains in a wide range of disadvantaged students in inner-city classrooms. These results also provide some evidence that GenScope provides a more effective environment for developing domain reasoning skills, compared to the more conventional environment it was designed to supplement or replace.

Perhaps the strongest conclusion from these results concerns the surprising finding that classes that used the GenScope curriculum without actually completing the GenScope computer activities showed the same reasoning gains as the same teacher's other classes that spent a substantial proportion of their time on the computer activities. Both Mr. H and Ms. L reported initially that the reason they limited computer access to a single class each was because of difficulties getting access to the lab. Subsequent investigation revealed that students in both of the classrooms that accessed the lab still encountered numerous difficulties carrying out the GenScope activities. The teachers reported problems such as scheduling changes, hardware and software problems, and confusion and problems with some of the activities. In summary, it appeared that on the days when one class went to the computer lab to independently figure out complex concepts under challenging conditions, their other class(es) participated in a teacher-managed whole-class activity that targeted the same concepts—primarily using the Dragon Investigations.

While the gains in the general science classrooms using the GenScope curricula were substantial, it should be noted that postttest proficiency still left many students below the level of reasoning

[12] Only five of the students had standardized fit statistics larger than ± 2.0 (the largest was –2.54); the means of the point-biserial correlations (a general measure of fit) for these students was .61, higher than any other classrooms.

associated with even the simplest effect-to-cause reasoning. Of the 137 general science students in this round, only 71 had posttest scores higher than –1.0 (the level associated with within-generation effect-to-cause reasoning[13]). Furthermore, only ten of the students had posttest scores above 0.0, and only one student had a posttest score above +0.5.

General Biology Classrooms

The remaining results presented here concern instruction in general (i.e. not college prep, also refereed to as "ABC" or "technical") biology classrooms. In addition to the general biology classes completing the NewWorm before and after instruction in the second round, one comparison general biology class participated during the first round as well.

"Programmed instruction" comparison. During the first round, Ms. B at School 2 (a suburban school serving relatively advantaged students) used a locally developed programmed instruction unit to cover genetics in her general biology classroom. Ms. B supplemented class lecture and discussion about genetics with the self-paced programmed instruction module that had been developed by a previous teacher at the school. Unfortunately, we were unable to obtain additional information about the curriculum or how many class periods were devoted to it. As shown in Figure 7, Ms. B's students (the open squares) showed a gain of .95 logits from pretest to posttest.

Second round GenScope implementations. During the second round implementations, another teacher at School 2, Mr. R implemented the GenScope curriculum in his two general biology classrooms. Mr. R was the science program director at the school and described himself as "above average" in his knowledge of genetics, his comfort teaching it, and his comfort integrating computer software into his curriculum. Prior the implementation Mr. R had taken his teaching sabbatical to work with the GenScope development team. Roughly half of the students in each classroom were on individual educational programs for learning or behavioral disabilities, and he characterized both classrooms as "challenging". During the fifteen 90-minute class periods across five weeks that were devoted to genetics, Mr. R reported that his students completed seven GenScope computer activities as whole-class activities, and completed another seven GenScope computer activities independently. All 11 Dragon Investigations were assigned as homework, but students were not graded on these activities and "at least half" of the students did not complete them. However, Mr. R reviewed the Dragon Investigations with the class in a fairly intensive review prior to posttesting. As shown by the triangles in Figure 7, the mean gains between Mr. R's two classrooms differed significantly [1.51 vs. 1.03 logits, $F(1,30) = 4.51$, $p = .042$].

[13] Domain reasoning at this level, essentially entails selecting the possible genotypes (i.e., *BB, Bb, bb, B-* or *b-*) that would yield a given the phenotype of a given characteristic (e.g., *flat* vs. *round body*). (Refer to Figure 2 and Figure 4.)

While the gain in Mr. R's two classes together was not significantly larger than the gain in Ms. B's comparison classroom, F (1,48) = 2.40, p = .128, the difference in the gains between Mr. R's class showing the larger gain and Ms. B's comparison class did reach statistical significance, F (1,32) = 6.16, p = .019.

The second GenScope implementation in a general biology classroom was in two of the five GenScope classrooms taught by Ms. M at School 3. School 3 served a suburban/industrial community and included a wide range of students. Ms. M is a self-described "genetics fanatic" who had independently incorporated the GenScope software into her biology curriculum the previous year after obtaining an early version of the software from the Internet. After participating in (and helping facilitate) a 40-hour teacher training workshop, she implemented the various GenScope computer activities in all five of her biology classes during the second round of implementations. This included the two general science classes reported here. Reflecting her strong orientation toward genetics, Ms. M used genetics topics to organize the broader life science content, and so reported devoting more days to genetics. In contrast to other teachers in the second round, Ms. M. did not follow the revised GenScope curricula; rather she fit in the various computer activities and Dragon Investigations alongside her other activities as appropriate.

During the 35 general biology class periods during which genetics was covered, Ms. M reported including 12 GenScope computer activities and 2 Dragon Investigations. Together these 21 students showed a modest gain of only .68 logits. However, as shown in Figure 7, (circles) the differences in gains in these two classrooms differed substantially, with one class gaining 1.23 logits and the other gaining only .09 logits, [although the small numbers and substantial variance precluded a statistically significant difference in gains, F (1,20) = 3.28, p = .086]. Further examination revealed that scores for 4 of the 10 students in the latter class actually declined. These students gave incorrect answers to problems on the posttest that were answered correctly on the pretest, and the fit statistics for these students suggests a pattern of responses that is entirely inconsistent with the underlying latent-trait model.[14] This raises obvious concerns about the conditions under which the posttest was completed.

Summary of findings in general biology classrooms. Overall, these results show that GenScope was an effective environment for learning introductory genetics when implemented in general biology classrooms. Relative to the comparison programmed instruction unit at School 2, the GenScope environment was significantly more effective in some classrooms but not others. Regarding the impact of the differences in the curriculum, the use of the Dragon Investigations was rather unsystematic and poorly documented; therefor the findings are rather inconclusive.

---

[14] The standardized infit statistics for these four students ranged between ±2.4 and ± 3.78

Relative to the general science students, the posttest proficiency in the general biology students was certainly higher. Of the 78 general biology students who completed the posttest in this round, 64 had posttest scores above –1.0. Still, however, only 8 of the students had posttest scores at or above +1.0, the level associated with between-generation effect-to-cause reasoning. As such, we still had not shown that the GenScope environment could develop what we considered meaningful domain reasoning skills in a substantial portion of non-college bound students.

Conclusions after the Second Round of Implementation

At what was supposed to have been the end of the implementations in the project, we were all still disappointed with the learning outcomes in the GenScope classrooms. The posttest proficiencies in the college prep biology and honors biology classrooms (reported in Hickey, Kindfield, Wolfe, & Heidenberg, 1999) were substantially higher, but the *gains* were substantially smaller than the gains in the general science and general biology classrooms reported here. We were not entirely surprised, because virtually every GenScope teacher had reported difficulties with computer access and problems with software and hardware. The teachers reported that a lot of time was wasted traveling to and from the computer lab to complete the GenScope computer activities, and entire class periods were "lost" when things did not work out correctly or when students misunderstood the problem.

In addition to limiting the learning gains in the GenScope classrooms, the difficulties with computer access and the curricular activities biased the results in favor of the non-computer GenScope classes and the non-GenScope comparison classrooms. This is because teachers in these classes ended up with additional time to focus on the reasoning skills targeted in the assessment. In particular, in the general science classrooms in the second round, it was clear that the students who did not complete the GenScope computer activities benefited from GenScope's curricular activities and structure, but without the difficulties and wasted time associated with the computer activities.

An additional issue clouded our interpretation of the results. While there was substantial evidence that the Dragon Investigations helped students do well on the NewWorm, we had not ruled out the possibility that they had done well because we had compromised the validity of the test. Specifically, at this point the data did not support the argument that the Dragon Investigations were primarily facilitating the development of the reasoning skills needed to do well on the NewWorm (as we expected). This left open the possibility that the Dragon Investigation fundamentally compromised the NewWorm by reducing the complex problems to simple algorithms. Reflecting the perspective on assessment and instruction discussed previously, we assumed that the Dragon Investigations sacrificed a small degree of NewWorm's evidential validity in exchange for maximized learning. Specifically, we expected the

Dragon Investigations to familiarize GenScope students with the NewWorm item format, providing them a small advantage relative to the comparison students.

We elected to conduct a follow-up study the subsequent school year to systematically examine the issues that were left unresolved at the end of the second round of implementations.

Follow-up Implementation at School 8

The follow-up study was conducted in three general biology classrooms at a suburban/rural school that served relatively advantaged students. Three classrooms at School 8 served a single pool of technical track (i.e., non university-bound) 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 classes and implemented the GenScope curriculum in both of them. Ms. H was a first-year teacher, and had participated in the GenScope research (primarily scoring assessments and evaluating curricular activities) during the previous year while she was a science education graduate student.

Addressing the first issue concerning problems with the computer activities, the GenScope activities were further refined and debugged and the students completed these activities using 10 laptop computers in Ms. H's biology lab/classroom. Addressing the second issue concerning the carryover effects of the GenScope curriculum and the associated lack of valid implementation/comparison pairs, a very experienced biology teacher was recruited to provide an "ideal" comparison class. Ms. F, who taught general biology to the same population of students, 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 classroom.

Addressing the third issue concerning the impact of the Dragon Investigations, the first class completed 15 GenScope computer activities (and no Dragon Investigations) over the roughly 25 class periods devoted to genetics. In contrast, the second period completed only 10 of the GenScope computer activities, but completed 6 Dragon Investigations as in-class activities in lieu of the computer activities. Thus, one group of students had roughly one-third of their computer-based activities replaced by the paper-and-pencil Dragon Investigations. Regular observations and daily videotaping confirmed that Ms. H was sufficiently versed with the curriculum to ensure that students covered the domain reasoning skills using the GenScope computer activities. Therefore, if the Dragon Investigations were compromising the NewWorm by leading students to solve the posttest problems using simple algorithms, the students in the second period would do substantially better than the students in the first period. Student solution

processes on the NewWorm posttest were further validated using retrospective think-aloud interviews with four students from each of the GenScope classrooms. This entailed having the first author review the completed posttest with students, asking them to explain their reasoning behind each answer. Students were prompted to explain their reasoning for the items they scored correctly, and were provided encouragement and hints on the items they answered incorrectly or elected to not answer at all in an effort to see if they could be prompted to answer them correctly (see Hickey, Wolfe, & Kindfield, in press, for more details on this method).

Regular observations and daily videotaping revealed that essentially none of the difficulties with the GenScope computer activities were encountered in either classroom. During the computer activities Ms. H (and a teacher's aide provided to support the learning disabled students) wandered among the students to answer questions and keep students on-task. Typically Ms. H would provide a brief introduction to the computer activity and would later call the class' attention during or following the activity to review or clarify a particular point. It is important to note however that these were by no means "ideal" classroom environments. Reflecting the number of behaviorally disabled students and the overall modest proficiency of these students, the videotapes reveal a good deal of "horsing around" during the computer activities and fairly extended stretches of off-task activity. During most periods, Ms. H and the aide were required to devote a substantial amount of their attention to maintaining order.

Figure 8 shows the mean reasoning gains in the three classes in School 8. 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 notably, the gain in the GenScope classroom that used the Dragon Investigations (circles) was 2.67 logits, the largest of any classroom in the study. The gains in both of the GenScope classrooms were significantly larger than the gain in the comparison class [$F(1,37) = 9.10$, $p = .005$, and $F(1,38) = 14.02$, $p < .001$, respectively]. The differences in the gains in the two GenScope classrooms was not statistically significant [$F(1,37) = 2.24$, $p = .143$].

Regarding the relative impact of the Dragon Investigations, these findings support our conclusion that these activities presented a small, acceptable degree of compromise to our evaluation'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. We assume that we would have seen a much larger relative gain in the second class if the Dragon Investigations had fundamentally compromised performance on the NewWorm (by reducing the problem complexity of the problems to the degree that they could be solved more algorithmically). The retrospective interviews revealed that the students from both classes used the expected reasoning processes to solve the problems they answered correctly, and there was no evidence that students in the

second period were more able to solve problems for which they could not also explain the domain concepts used to solve them. In summary, because the organism, genotypes, and phenotypes of the two instruments (and the format of some of the problems) are entirely different, it appears that the Dragon Investigations had precisely the desired effect—developing transferable domain reasoning skills.

Regarding the learning outcomes, this was clearly our most successful implementation to date. It appears that the further enhancements to the curriculum and software, implemented on computers in the classes under the constant guidance of a knowledgeable teacher yielded the degree of learning outcomes that we had all been seeking since the start of the project. Of the 39 students in Ms. H's two classes, 10 students (25%) had posttest scores at or above 1.0, the level corresponding to effect-to-cause between-generation reasoning. Given the modest academic proficiency overall and the rather low pretest scores, this seems like a substantial accomplishment. While Ms. H had become very familiar with the content covered in GenScope and targeted in the NewWorm working as a graduate research assistant the previous year, this was also her first year as a teacher. It seems reasonable that other life science teachers could achieve these same gains, given a genetics background associated with undergraduate biology coursework, the kind of workshop instruction typically provided to GenScope teachers, and the curricular package developed during the second round. 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 environments—at least in terms of the sort of domain reasoning skills assessed with the NewWorm.

Conclusions

The first conclusion from these findings concerns the Dragon Investigations. Traditional program evaluators likely consider our development of instructional activities that "teach to the test" as downright heretical. However, new perspectives on assessment argue that preserving evidential validity at all costs is inappropriate and unethical. These results show that it is possible to sacrifice a small, knowable, degree of evidential validity in exchange for dramatic increases in consequential validity. After all, the ultimate goal of all educational assessment and evaluation is enhancing learning. These results support theorists like Frederiksen and Collins (1989) who argue that assessment practices are simply too valuable for organizing curricula, scaffolding learning, and motivating students to continue preserving evidential validity's sanctity at all costs. The follow-up study provides one example of how we can take advantage of the powerful affordances of assessment practice for learning while still providing the degree of evidential validity called for in current policy documents such as the PCAST (1997) report. These results

provide both a justification and a framework to further develop linked instructional and assessment activities within the new *BioLogica* software. When paired with the validity inquiry described in Hickey, Wolfe, & Kindfield (in press) and used in the type of implementation studies described here, we should be able to continue developing "systemically valid" assessment practices that simultaneously maximize learning and preserve evidential validity.

Another conclusion can be drawn by considering the results of the follow-up GenScope implementation in light of the results from the previous year's implementation. That the outcomes in these GenScope classes were more positive than those in the other GenScope classes shows that providing computer access *in the classroom* enhances the teacher's ability to use the technology-supported curriculum to support meaningful learning. These findings provide additional support to the many arguments for placing of computers in content area classrooms rather than computer labs, and lend additional credence to the conclusion of the PCAST (1997) report that priority should be given to using computers to teach subject area content rather than teaching about computers themselves. Fortunately, the number of computers in content area classrooms appears to be increasing. Additionally, it seems that educators are recognizing that the security risks and relative fragility posed by laptop computers may well be outweighed by their flexibility, small footprint, and minimal power consumption and heat generation. These are certainly positive developments in the deployment of computer technology in the service of classroom learning.

The broader conclusion from this research concerns the value of the GenScope learning environment for developing students' ability to reason about introductory genetics. The results from the follow-up study at School 8 provide convincing evidence that the GenScope environment is more effective than the typical approaches to genetics instruction at developing domain-reasoning proficiency. In light of the dimensions of domain reasoning described earlier, the gains in the follow-up GenScope classrooms represent roughly the difference between within-generation and between-generation reasoning, or the difference between cause-to-effect and effect-to-cause reasoning. It seems reasonable to interpret these gains as fundamental, qualitative change in students' ability to reason in the domain of introductory genetics. We interpret these increases as the sort of increases in domain reasoning skill whose absence in typical biology classrooms has long been lamented by science education researchers like Stewart and Hafner (1994). We expect that subsequent implementations where the sorts of curricular activities described here (and more) are incorporated into the computer environment, and is accessed using computers in life science classrooms, should yield even more dramatic learning outcomes.

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., Wolfe, E. W., & Heidenberg, A. (1998). Implementation and evaluation of the GenScope™ learning environment: Issues, solutions, and results [Learning outcomes: Section 4]. In M.Guzdial, J. Kolodner, A. Bruckman, & A. Ram (Eds.), *Proceedings of the Third Annual International Conference of the Learning Sciences* (pp. 6-10*)*. Charlottesville, VA: Association for the Advancement of Computers in Education.

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. & 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.

Kindfield, A. C. H., Hickey, D. T., & Yessis, L. M. (1999, March). *Assessing Student Understanding of Genetics: The NewWorm© Assessment.* Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Boston, MA.

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.

National Research Council. (1996). *National Science Education Standard*. Washington, DC: National Academy Press.

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

President's Committee of Advisors on Science and Technology, Panel on Educational Technology (PCAST) (1997, March). *Report to the president on the use of technology to strengthen K-12 education in the United States.* Author.

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.

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

Table 2. Activities in 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[1]
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”.

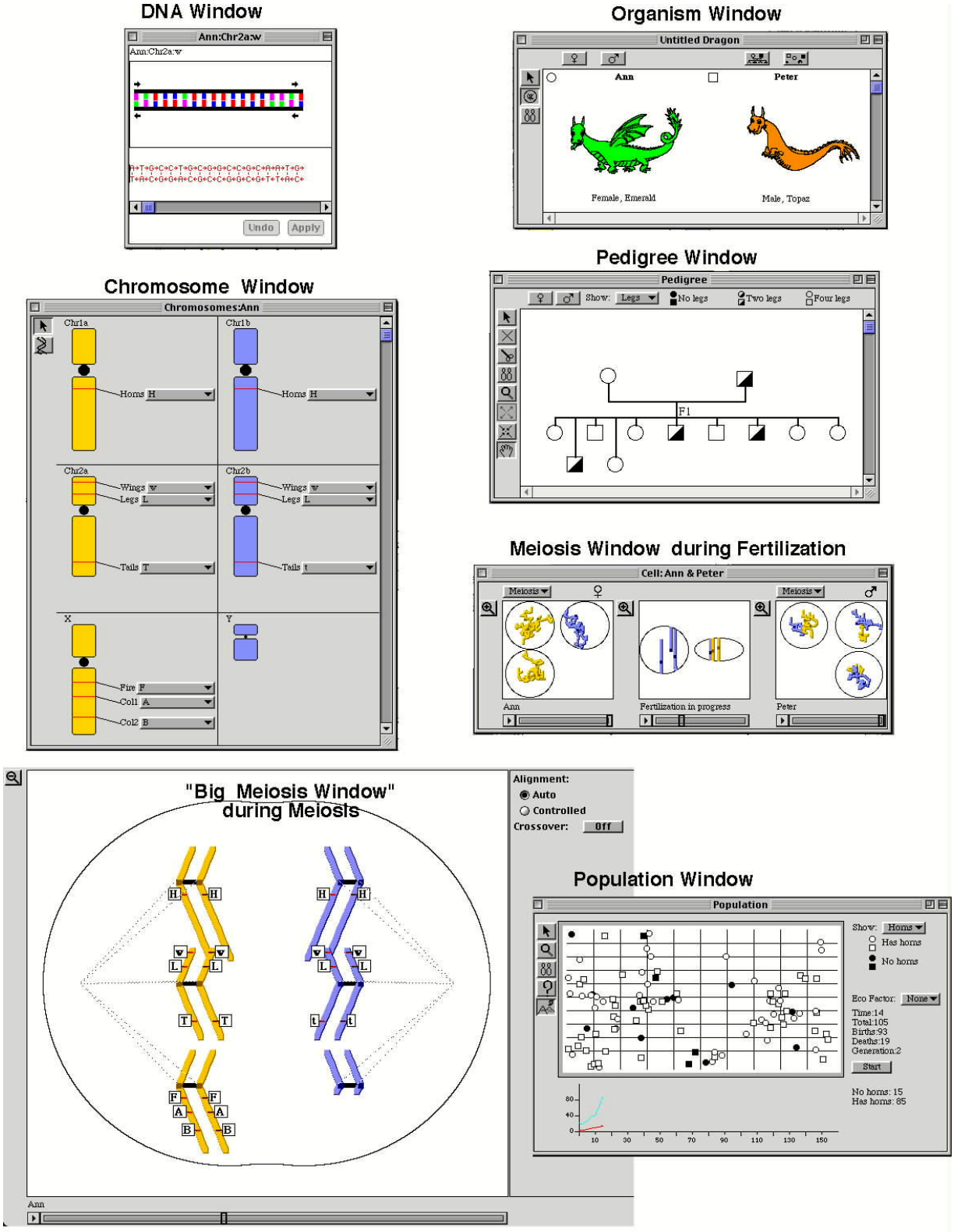


Figure 1. Examples of Screens in the *GenScope* Software.

NewWorm Genetics

Body: Flat: **BB** or **Bb** Round: **bb**

Mouth: Oval: **??** Slit: **??**

Head: Broad: **??** Medium: **??** Narrow: **??**

Rings: No Rings: **RR** or **Rr** Rings: **rr**

Color: Green: **CC** Brown: **Cc** Black: **cc**

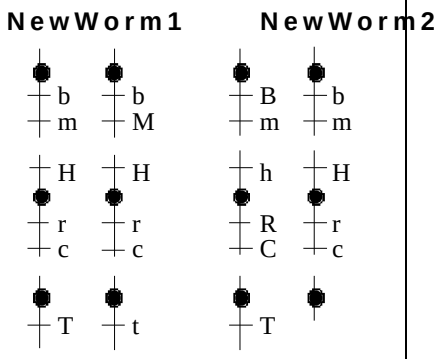
Tail (Male): Pointed: **TT** or **Tt** Blunt: **tt**

Tail (Female): Pointed: **T–** Blunt: **t–**

(The Tail gene is on the **X** chromosome.)
(The **–** [dash] stands for the **Y** chromosome.)

Sex: Males: **XX** Females: **XY**

Two NewWorm Genotypes



Determine phenotypes (traits) from NewWorm1 and NewWorm2's genotypes:

		NewWorm1		**NewWorm2**
What body shape?	1a.	_____	1b.	_____
Does it have rings?	2a.	_____	2b.	_____
What color?	3a.	_____	3b.	_____
What kind of tail?	4a.	_____	4b.	_____
Male or female?	5a.	_____	5b.	_____

If the allele for **oval mouth** (**M**) is dominant to the allele for **slit mouth** (**m**):

What kind of mouth?	6a.	_____	6b.	_____

Figure 2: Example *NewWorm* items assessing cause-to-effect, within-generation reasoning.

Another inherited characteristic in the NewWorm is Eyelids. Both NewWorm1 and NewWorm2 have clear eyelids. However when you mate them and produce 100 offspring, you find:

- 74 (51 males and 23 females) have clear eyelids
- 26 (0 males and 26 females) have cloudy eyelids

Remember: Males are XX and females are XY.

1. There are two alleles for Eyelids. Is the relationship between the two alleles simple dominance or incomplete dominance? Answer: _____

1a. What is it about the **offspring** that indicates simple or incomplete dominance?

2. **If** one of the Eyelids alleles is dominant, which one is it (clear, cloudy, OR neither)?
Answer: _____

2a. What is it about the **offspring data** that shows you which, if any, allele is dominant?

3. Is the gene for Eyelids autosomal or X-linked? Answer: _____

3a. What is it about the **offspring data** that indicates whether the gene is autosomal or X-linked?

Figure 3. Example *NewWorm* item assessing effect-to-cause, between-generation reasoning.

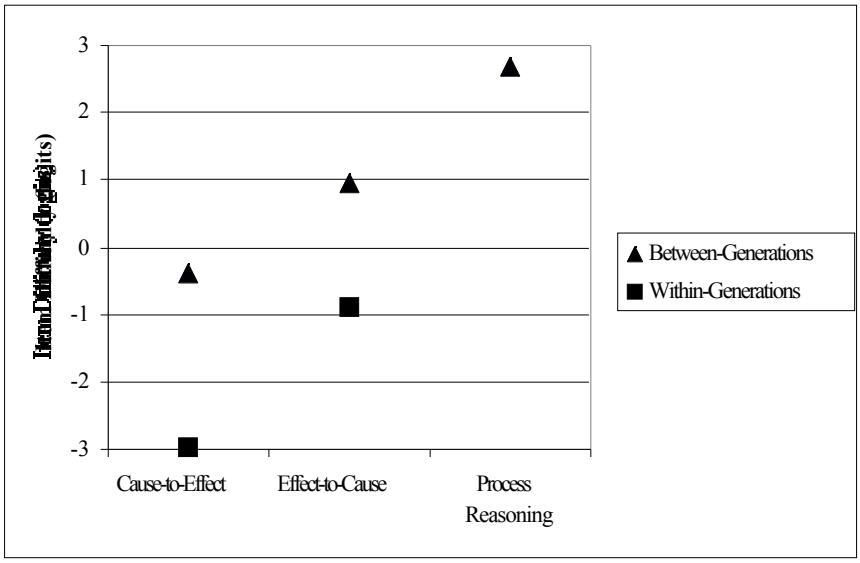


Figure 4: Relative difficulty of clusters of *NewWorm* items, by reasoning type.

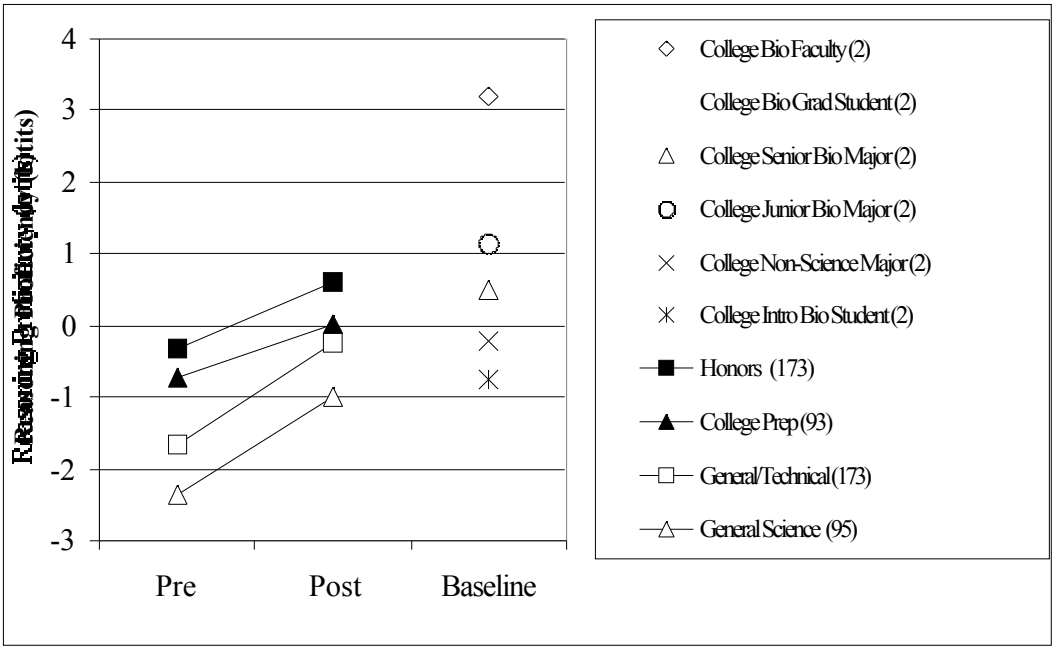


Figure 5. Genetics reasoning proficiency before and after instruction in four sets of high school classroom and in six pairs of college biology students and faculty.

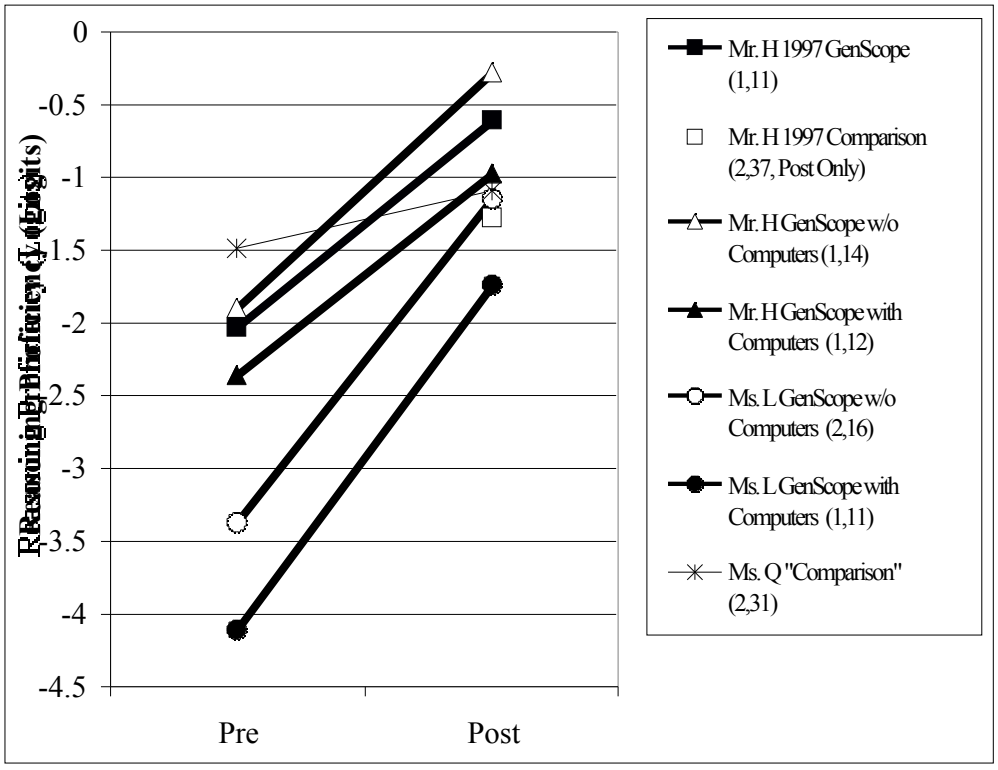


Figure 6. Gains in genetics reasoning proficiency in general science classrooms.

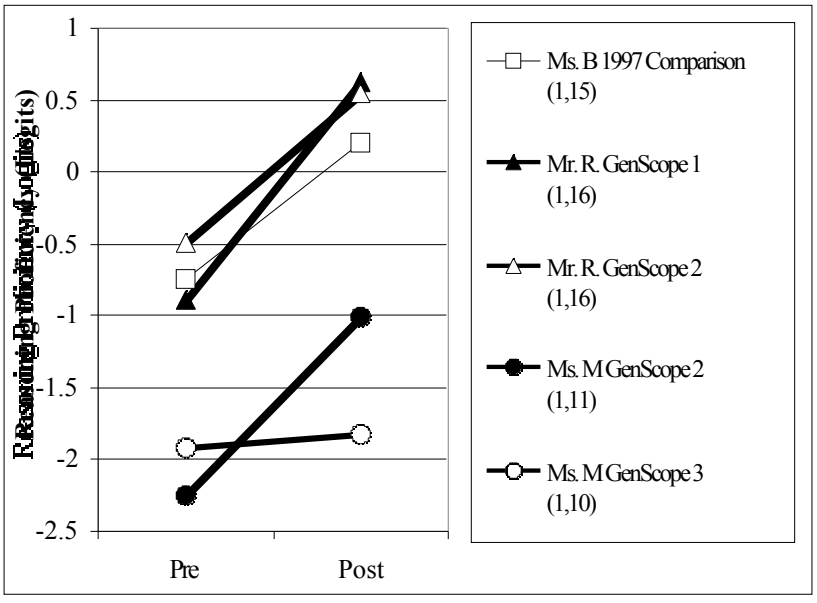


Figure 7. Domain reasoning gains in general/technical biology classrooms .

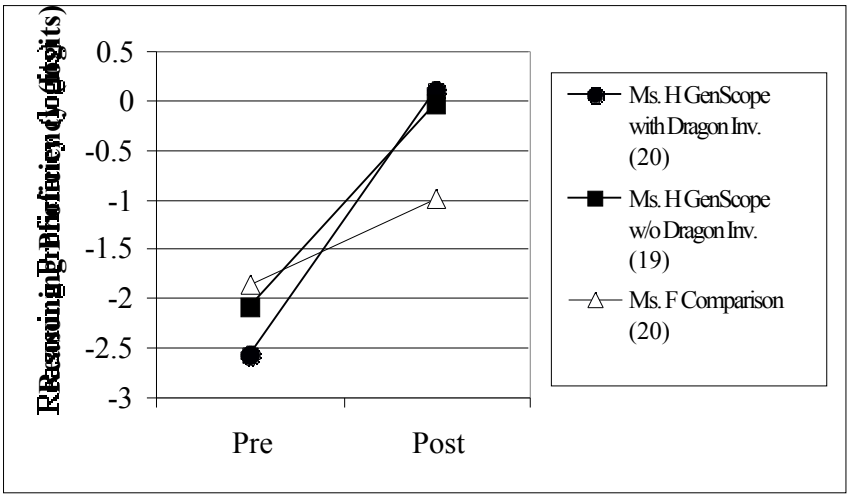


Figure 8. Reasoning gains in general/technical biology classrooms during follow-up implementation.