*American Educational Research Journal*
*Summer 2003, Vol. 40, No. 2, pp. 495–538*

# Integrating Curriculum, Instruction, Assessment, and Evaluation in a Technology-Supported Genetics Learning Environment

Daniel T. Hickey
*University of Georgia*
Ann C. H. Kindfield
*Educational Designs Unlimited, Inc.*
Paul Horwitz
*Concord Consortium*
Mary Ann T. Christie
*Lesley University*

*This article describes an extended collaboration between a development team and an evaluation team working with GenScope, an open-ended exploratory software tool. In some respects, this was a routine evaluation, documenting substantial gains (of roughly 1 SD) in genetics reasoning ability in all but 1 of 17 classes, despite challenges presented by school computer-lab settings. Relative to matched comparison classes, larger gains were found in technical biology and general science courses but not in college prep or honors biology courses. In other respects, our effort illustrates the value of new views of assessment, technology, and research. The alignment of a sophisticated research assessment and simple classroom assessments shed light on initial failures, spurring revision. By refining the GenScope activities and extending the classroom assessments, we supported worthwhile whole-class discourse around the shared understanding of the software. A follow-up study in a laptop-equipped classroom yielded the absolute and relative gains (3.1 SD and 1.6 SD) that proponents of such innovations have long promised. In retrospect, the strengths and weakness of the study illustrate the value of newer "design-based" approaches to educational research.*

**KEYWORDS**: assessment, design-based research, evaluation, genetics learning, model-based reasoning, validity.

Genetics is a particularly challenging topic for science teachers and their students. It involves relationships between events that occur at various levels of biological organization and describes probabilistic phenomena that are not directly observable because they take place too quickly or too

slowly, or on a scale that is too small or too large. Thus mastery of the genetics content and reasoning goals defined in current science education standards (e.g., National Research Council [NRC], 1996) can be daunting. To help meet this challenge, science education researchers have invested heavily in computer-based tools for teaching genetics (e.g., Jungck & Calley, 1985; Stewart, Hafner, Johnson, & Finkel, 1992). Starting in 1991, a team at Bolt, Beranek and Newman Labs (the team is now at the Concord Consortium) began developing and refining the GenScope software, developing curricular activities, and piloting those activities (Horwitz & Christie, 2000; Horwitz, Neumann, & Schwartz, 1996).[1] The GenScope software has been acknowledged as a noteworthy example of the synergy between educational technology and contemporary pedagogical principles (e.g., NRC, 1999a, Chap. 9) and is consistent with the vision advanced in a report issued in 1997 by the President's Committee of Advisors on Science and Technology (PCAST).

Classroom research began as soon as the first working version of GenScope was developed. The present article focuses on a research collaboration that was initiated in 1995 between the GenScope development team (headed by Paul Horwitz and including Eric Neumann, Mary Ann Christie, Joyce Schwartz, and others) and an "outside" assessment team at the Educational Testing Service (headed by Ann Kindfield and including Dan Hickey, Drew Gitomer, Linda Steinberg, and others). The stated goal of this collaboration was developing curricular materials that would allow the GenScope software to be widely and readily deployed, developing a tool for

Daniel T. Hickey is a Research Scientist in the Learning and Performance Support Laboratory and an Assistant Professor in the Department of Educational Psychology, 611 Aderhold Hall, University of Georgia, Athens, GA 30602; e-mail *dhickey@coe.uga.edu*. His areas of specialization are classroom assessment, motivation, technology, and sociocultural theory.

Ann C. H. Kindfield is a Geneticist and Science Education Consultant with Educational Designs Unlimited, Inc., 325 Zion Road, Hillsborough, NJ 08844-2511; e-mail *akindfield@patmedia.net*. Her areas of specialization are curriculum, assessment, representations, and model-based reasoning.

Paul Horwitz is a Senior Scientist with the Concord Consortium, 10 Concord Crossing, Concord, MA 01742; e-mail *paul@concord.org*. His area of specialization is computer-supported modeling environments for science and mathematics education, across curricular domains.

Mary Ann T. Christie is an Assistant Professor of Education and Technology at Lesley University, 29 Everett Street, Cambridge, MA 01238; e-mail *mchristi@mail.lesley.edu*. She specializes in research on educational practice, learning technologies, and graduate teacher education.





assessing learning outcomes, and implementing both in a wide range of life sciences classrooms.

Our collaboration was initiated at the suggestion of project officers at the National Science Foundation (NSF). From the outset, we targeted the kind of research on educational technology called for in the PCAST (1997) report. Reflecting the immaturity of constructivist practices and opposition from some quarters, the report called for increased research on constructivist applications and a broadening of such research beyond a focus on formative questions and interpretive methods, to focus on "well-designed, carefully controlled experiments having sufficient statistical power to distinguish genuine effects of a relatively modest size from differences that can easily be explained as chance occurrences" (p. 94). The report further advocated that 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" (p. 95).

Since the publication of the PCAST report, increased focus on accountability has increased the demand for such research. Pressure from stakeholders outside the educational research community, along with changing assumptions within, have led to dramatically heightened concerns regarding evidence of program effectiveness:

> Technology design as educational research can no longer focus on just imagination and inquiry. Research on technology is like a three-legged stool and an explicit quest for impact is the third leg required to stabilize research programs. Without this third leg research totters between boutique studies which produce much excitement and knowledge about circumstances that defy replication, and large demographic studies which provide knowledge about the success and failure of today's educational technology but little direction for tomorrow. (Roschelle & Jackiw, 2000, p. 779)

When our project was initiated, it was one of the more ambitious studies carried out in this spirit. We believe that the collaboration that emerged is worthy of consideration because it embodies aspects of the type of educational research subsequently called for by expert panels in North America. For example, in 1999, the National Educational Research Policies and Priorities Board called for more "extended collaborative effort directed at pressing practical problems"; another committee at the NRC (1999b) called for "focused, multidisciplinary, cumulative, sustained, solutions-oriented research." Similar calls have been made by European experts (e.g., DeCorte, 2000). A more detailed summary of these reports, in light of our effort, is provided in Hickey, Kindfield, Horwitz, and Christie (1999); this article focuses on the nature of our research and findings as the collaboration unfolded, and considers them in light of new "design-based" research methods (e.g., Kelly, 2003).

We first outline the challenges of teaching and learning introductory genetics, describe the GenScope software and how it promised to help meet these challenges, and describe the NewWorm assessment system. We then describe pilot studies carried out during Year 1 and subsequent revisions and expansions of the GenScope curriculum. We outline findings and conclusions from large-scale implementation and evaluation conducted during Year 2 and Year 3. Finally, we describe a focused "follow-up" study carried out during Year 4 to address unresolved questions about using GenScope in computer labs (rather than biology classrooms) and about the validity of the NewWorm assessment in light of curricular extensions developed by the assessment team.

## Learning, Teaching, and Assessing Introductory Genetics

### The Challenge of Introductory Genetics

Secondary life science instruction leaves most students with little more than disconnected notions of biological processes such as meiosis and a set of rudimentary algorithms for solving basic inheritance problems using the familiar Punnett Square (Slack & Stewart, 1990). Thus mastery of secondary introductory genetics has commonly been associated with the ability to solve problems such as this: *Curly hair in guinea pigs is a recessive trait. What is the probability that mating a homozygous recessive* (cc) *and a heterozygote* (Cc) *will yield offspring with curly hair?* To determine the correct response (50%), one need only insert appropriate letters into the boxes of the 2 × 2 Punnett square (i.e., *cc* × *Cc*) and see how many of the four offspring boxes in the matrix contain a combination that will yield the trait (two of the boxes contain *Cc* and so do not yield curly hair; the other two contain the *cc* needed to yield curly hair). Stewart and Hafner (1994) and others have argued that such problems can be and often are solved with very little actual knowledge of the domain of genetics; rather, students merely plug the letters into a well-learned algorithm and generate the correct answer. For example, although the previous example appears to require understanding the distinction between dominant and recessive traits, students often learn that distinction as a simple algorithm regarding uppercase and lowercase letters, constraining its meaning to this specific problem. This leaves most students unable to solve more complex problems such as those involving dihybrid inheritance (solving for two characteristics simultaneously, using a 4 × 4 Punnett square) or sex-linked traits. Fewer still attain the understanding needed to solve problems that geneticists might be concerned with—the *effect-to-cause* problems that require reasoning from an outcome, such as the expression of specific traits of a genetic characteristic or the results of a mating, back to general rules for the inheritance of the characteristic. Consider, for example, the following problem from the National Assessment of Educational Progress (National Center for Education Statistics, 1996) secondary science assessment:





> A mother with attached earlobes and a father with free earlobes have 5 children—4 boys and a girl. All of the children have the father's type of earlobes. What can be predicted about the genotype of the father? Construct a genetic diagram to support your prediction. What additional information, if any, is needed to determine the genotype of the father? (p. 148)

A correct answer would indicate that (a) free earlobes were a dominant trait; (b) the father was probably homozygous dominant (i.e., *LL*) and the mother was probably homozygous recessive *(ll);* and (c) additional information about the father's parents' genotypes would be helpful in determining his genotype. With answers receiving .33 for each of the elements included, the mean score on this item in the 1996 administration was only .28, making it one of the most difficult items on the entire assessment.

A further challenge in introductory genetics is the need to apply one's understanding of the various biological events that relate to genetics. For example, *crossover* is an event that occurs when pairs of chromosomes entwine during meiosis, swapping entire segments of DNA in the process. Most students learn about crossover when learning about meiosis, but few understand how crossover affects inheritance. For example, the Punnett square will accurately predict both the range and the probability of outcomes of dihybrid inheritance problems when the two genes are in different chromosomes (or far from each other in the same chromosome); however, when they are near each other in the same chromosome, crossover alters the probability of outcomes predicted by the Punnett square. More complex cause-to-effect problems and most effect-to-cause problems are challenging for many students because (a) they require a cognitive model of the domain that can be "run" forward and backward to generate a correct answer; and (b) most textbooks, and therefore most science instruction, teach cause-to-effect reasoning only (Stewart & Hafner, 1994). GenScope, like most other software for teaching introductory genetics, was developed specifically to help learners develop such a cognitive model and gain practice in such reasoning.

### The GenScope Software

The GenScope software was designed to run on the 33 Mhz Macintosh computers that were widely available in schools starting in the early 1990s.[2] As shown in Figure 1, the various levels of biological organization relevant to introductory genetics are represented in GenScope by different software windows. Each window graphically represents the appropriate information alongside easy-to-use tools for manipulating that information. Just as genetic information flows between the levels of biological organization, the information flows between the levels of the software, so that manipulations made at any one level are immediately reflected in the others.

Although a variety of actual organisms (humans, dogs, etc.) are represented in GenScope, most of the activities involve *dragons*. These fanciful

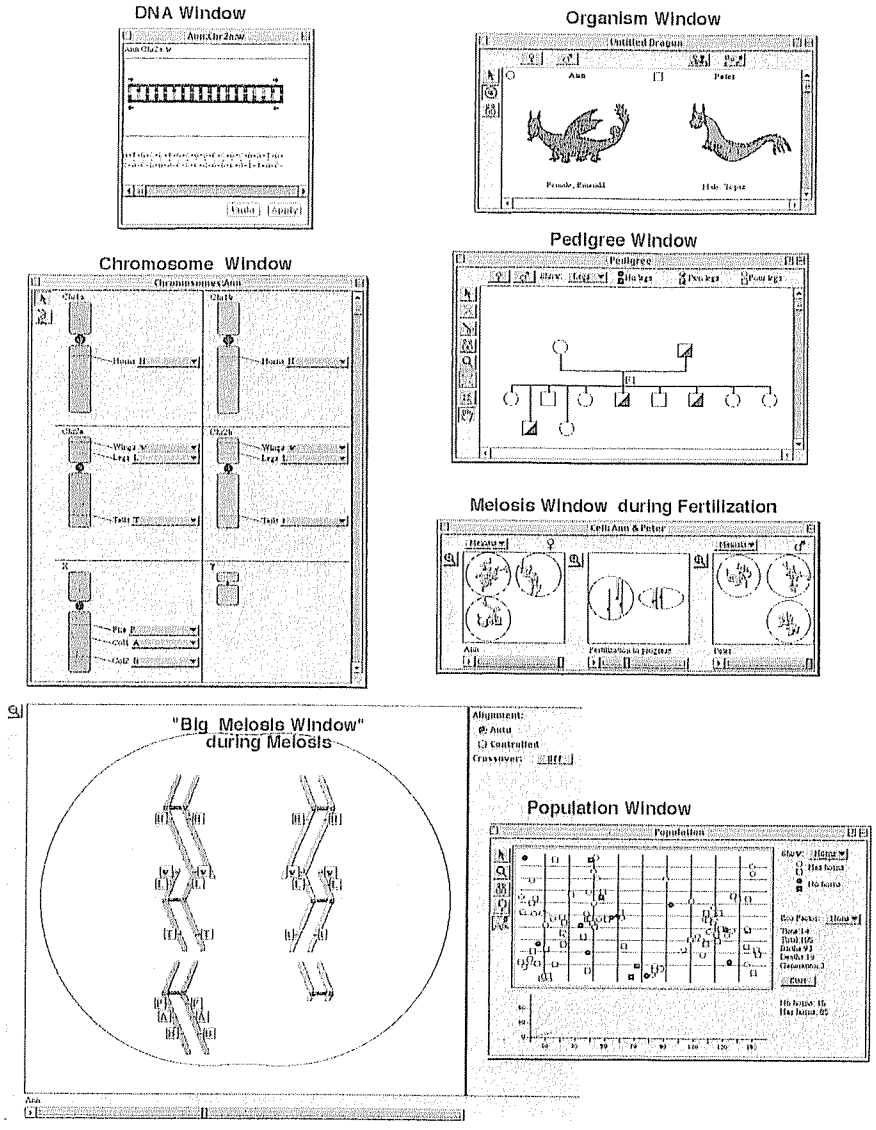



*Figure 1.* **Examples of Screens in the GenScope software.**





creatures feature various genetically simplified traits (e.g., color, number of legs, presence or absence of wings and horns), but using three pairs of chromosomes. Simple pulldown menus generate a male or a female organism in the *organism window.* This window displays the organism's *phenotype* (the expression of its physical traits) but not its *genotype* (genetic makeup). Two more clicks reveal an organism's *chromosome window.* GenScope represents chromosomes schematically, with the genes marked at their respective locations, just as in textbooks. Clicking on the gene labels changes the gene from one *allele* to another (e.g., *A* to *a*). Such changes are accompanied by possible changes in the organism's appearance or *phenotype,* following Mendel's Laws.

For example, our dragon has horns and displays the alleles *HH* ("homozygous dominant") for horns in the chromosome window. This is because the presence of horns happens to be a dominant trait. Illustrating the core concept of dominance, the horns remain if the alleles are changed to *Hh* ("heterozygous") but go away if the alleles are changed to *hh* ("homozygous recessive"). In contrast, having wings, a recessive trait, occurs only in the homozygous recessive (*ww*) individuals. Dragon legs illustrate incomplete dominance, where the heterozygote (*Ll*) expresses an intermediate characteristic (two legs), while *LL* and *ll* produce four and zero legs. Two more clicks open the *DNA window,* revealing a short strand of the chromosome's DNA. This window includes a physical representation where colored rectangles represent the base pairs and an informational representation with the letters ATGC (standing for adenine, thymine, guanine, and cytosine). A text-editing cursor can be used to alter the sequence of letters, producing *mutant* alleles that can be named and used like the predefined ones. Their default effect is to mimic the recessive allele, but they randomly trigger preprogrammed mutant traits (e.g., albinism, double wings, and a unicorn). The *cell window* lets users witness and control the creation and fertilization of *gametes* (sperm and egg cells). A larger window shows accurately detailed animation of the "dance of the chromosomes" that occurs during meiosis and mitosis. During meiosis, learners can control key events (*crossover* and *alignment*) as the single parental cells each divide into four gametes. The learners can then select one gamete from each parent and run fertilization. Mendel's laws of inheritance are graphically followed throughout the process and determine the resulting offspring's genotype and corresponding phenotype. The *pedigree window* allows learners to create multiple offspring at once, using the standard "family tree" pedigree representation for each trait, where female and male offspring are represented by circles and squares, respectively. Finally, learners can place a specified number of organisms in the *population window,* where each is represented by smaller circles and squares that function like those in the pedigree window. This window lets learners "run" evolution for a population of organisms with randomly chosen genotypes for many generations, consider several environmental variations, and see traits disappear from the population. Interested readers should note that the functionalities of GenScope were subsequently incorporated into the more powerful and flexible BioLogica, developed within in the Concord Consortium's *Modeling Across the Curriculum* project.[3]

## The Initial GenScope Curriculum

The development team designed a set of increasingly complex curricular activities when our research collaboration was initiated and continued refining them and developing new ones. Most were 1-to-3-page discovery-learning activities that could be completed within a single class period. For example, an activity called Fire Breathing was designed to teach about sex-linked inheritance, which is exemplified in dragons by the fire-breathing trait. As in many higher organisms, one gender has an XY chromosome pair, so any genes located in either the X or the Y chromosome have only a single "copy" of the gene in that gender. The other gender has an XX chromosome pair and thus contains two "copies" of any X-linked genes and no "copies" of any Y-linked genes.[4] The fire-breathing activity directs learners to launch a file that opens up the Pedigree window with the three dragons represented. Learners are instructed to use the various tools to cross the dragons and make predictions about how the trait is inherited. Additional items prompt students to explain their predictions, and a teacher version provides sample answers and key points to explain to the students. The development team created 17 such activities, and workshop teachers were strongly encouraged to develop their own activities and share them with others.[5] The curriculum for Year 1 included roughly 20 GenScope activities and assorted supplemental activities such as homework, videos, and quizzes.

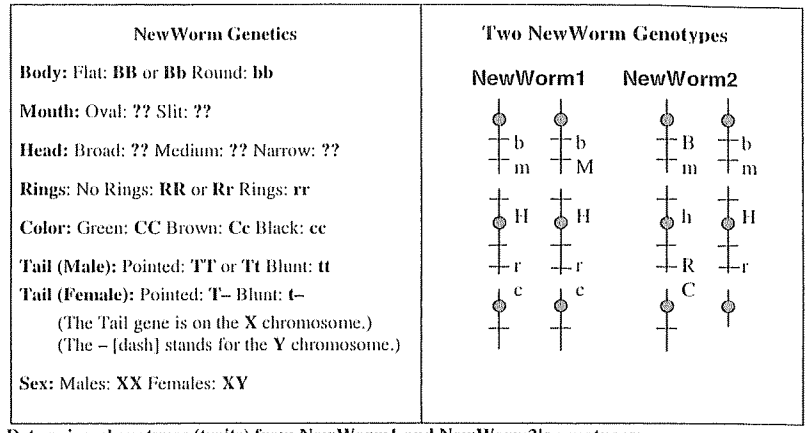
## Assessment System

The assessment team's first goal was to develop a tool for assessing understanding of the domain of introductory genetics. Constraints on the design of the assessment were the need to (a) use a paper-and-pencil format; (b) satisfy both (immediate) research and (ultimate) dissemination goals; (c) assess multilevel reasoning; (d) compare GenScope and non-GenScope users; and (e) assess a broad range of student populations. This was a substantial challenge, requiring two versions and four revisions across roughly 2 years. The resulting NewWorm assessment addressed 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. The items were carefully sequenced to scaffold student performance across increasingly complex problems.

The initial set of items on the NewWorm is shown in Figure 2. The first problem (1a and 1b, *What body shape?*) was designed to be solvable by most secondary students prior to instruction in genetics. These initial problems introduce the students to the organism, its genome, and the assessment environment. Success on the initial problems was expected to yield motivation and understanding that would scaffold performance on the more difficult subsequent problems. Notably, the inclusion of such items was a subject of extended debate within the assessment team. Some argued that such problems were so trivially easy that they should not be included. Accordingly, the first version of the assessment (the *NewFly*) did not provide the 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 (initial NewWorm item set).**

for any of the traits (akin to 6a and 6b in Figure 2). This required students to understand and apply the meaning of the terms *dominant* and *recessive* to solve the problem correctly. Others argued that assessments should give all students as much scaffolding as they needed and then see how far each student could progress as the scaffolding was removed (as argued by Wolfe, Bixby, Glenn, & Gardner, 1991, and by Frederiksen & Collins, 1989). The debate was settled when pilot studies revealed that some of the students in both GenScope and comparison classes struggled with items such as 1a and 1b even after instruction. Like the items in Figure 2, many of the items on the NewWorm called for categorical, single-word responses (or selection from multiple verbal or diagrammatic choices). However, as shown in Figure 3, the items assessing more complex reasoning also asked students to explain their reasoning for a particular categorical response.

Our assessment system reflects persistent recommendations that assessment should reflect what is known about the development of expertise in

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 items assessing effect-to-cause, between-generations reasoning.**

the domain (e.g., Glaser, Lesgold, & Lajoie, 1987; Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999; Wiggins, 1993). The content and organization of the NewWorm performance assessment embodies state-of-the-art understanding of the development of expertise in introductory genetics rather than traditional curricular scope and sequence. For example, traditional life science curricula often teach meiosis separately from genetics; this is problematic because events that occur during meiosis (e.g., crossover) are critical to understanding basic inheritance. The NewWorm includes items that directly assess whether students understand the consequence of meiotic events for inheritance.

Reflecting the prior research of Stewart (1988) and Kindfield (1994), Table 1 shows that developmental expertise in genetics can be represented by crossing two primary dimensions: (a) Domain-General Reasoning Type (*cause–to-effect, effect-to-cause,* and *process reasoning*), and (2) Domain-Specific Reasoning Type (*within-generation* and *between-generations reasoning*). For the most part, reasoning within generations is easier than reasoning between generations; reasoning from causes to effects (from genotypes to phenotypes)[6] is easier than reasoning from effects to causes (from phenotypes to genotypes) (Stewart, Stewart, & Hafner, 1994); and reasoning from effects to causes, in turn, is easier than reasoning about processes (Kindfield, 1993/1994, 1994). Figure 3 provides an example of the items used to assess effect-to-cause between-generations reasoning. GenScope was designed to support the development of reasoning in all of these categories, and thus all categories were represented in the NewWorm assessment.[7] Table 1 shows how the two primary dimensions of reasoning were crossed in the assessment design and describes the type of items that fell into each cell of the design. In addition to these primary dimensions, most items can



*Table 1*

**Primary Dimensions of Reasoning Represented by Items in the NewWorm Assessment**

| Domain-specific dimension of reasoning: Simple ↔ Complex | | Domain-general dimension of reasoning: Novice ↔ Expert | | |
|---|---|---|---|---|
| | | Cause-to-effect | Effect-to-cause | Process reasoning |
| 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 the genetic makeup of an organism and the products of a single meiosis, describe the meiotic events that resulted in this set of products. |
| Within generation | | *Genotype-to-phenotype mapping:* Given genotypes and information about NewWorm genetics, predict phenotypes. | *Phenotype-to-genotype mapping:* Given phenotypes and information about NewWorm genetics, predict genotypes. | None |

also be distinguished according to the particular genetics involved, the explicitness of provided information, and the type of information used or sought. For example, 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 involve explicit information, whereas Item 6 involves implicit information; Item 5 concerns a more complex sex-linked characteristic. In contrast, the problem shown in Figure 3 is an example of between-generations effect-to-cause reasoning and asks for both categorical and short-answer explanatory responses. Further details about the NewWorm instrument, including its psychometric properties, were presented in Kindfield, Hickey, and Wolfe (1999).

## Year 1: Piloting, Revisions, and Formative Assessments

During Year 1, the development team conducted the two consecutive implementations in general (also known as technical) life sciences classes in a suburban secondary school. Several team members worked alongside the teacher for approximately 25 class periods. Students then completed the precursor version of the NewWorm performance assessment. We found disappointingly modest proficiency. Students clearly had learned to navigate the GenScope environment and appeared to understand the relationships between the various windows and had been able to complete the activities without excessive guidance. However, they seemed to be left with little knowledge of introductory genetics. For example, in the two classes in the second pilot implementation, only 20 of 44 students were able to solve cause-to-effect, between-generations problems involving autosomal inheritance (akin to, given the information in Figure 2, selecting *definitely yes, maybe,* or *definitely no* as the answer to the question *Would an offspring of NewWorm 1 and NewWorm 2 have an oval mouth?*). More critically, none of the students could confidently solve such problems involving the sex-linked characteristic (i.e., *Would a male offspring of NewWorm 1 and NewWorm 2 have a pointed tail?*); and all seemed utterly baffled by effect-to-cause problems like the one shown in Figure 3.

The modest evidence of learning in light of students' seemingly thoughtful engagement immediately raised the issue of *transfer*. The students were able to complete the assigned worksheets with reasonable amounts of guidance and were able to complete the fairly informal quizzes that were included in the original GenScope curriculum. The question was, Did doing so leave students with meaningful knowledge of genetics that they somehow could not transfer to the assessment environment? Or were they simply learning to navigate the software? A key disagreement emerged in what turned out to be a critical point in our collaboration. One view was that students were indeed learning genetics but that this knowledge did not transfer to the NewWorm assessment environment. One theory in support of that view argued that knowledge performances differ across modalities (e.g., the written versus situated performances) in fundamental ways. Another related theory argued that the written tests might not contain enough information and therefore that the students may have had the conceptual understanding but applied incorrect rules. In each of these cases, we would conclude that the students were indeed learning introductory genetics but that the assessment instrument was requiring too much additional inference. This position was certainly bolstered by the prior concerns that the initial items on the NewWorm precursor were unnecessarily difficult.

An alternative explanation for the poor assessment performance was that the GenScope activities focused insufficient attention on the genetics concepts that could be expected to transfer to the assessment task. The specific theory was that much of the learning that was occurring was quite specific to the GenScope environment but was not directly helpful in navigating the NewWorm environment. Our understanding of this possibility was informed



by emerging *situative* views of transfer (e.g., Greeno, Smith, & Moore, 1993; Greeno et al., 1998). Situative analyses of transfer compare the resources that support meaningful participation in the learning environment with the resources that support participation in the transfer environment. Specifically, this means that one must consider the constraints and affordances that simultaneously bound and scaffold successful participation in the learning environment and in the transfer environment; one must then consider "transformations" between the two. For transfer to occur, some constraints and affordances must be the same (be "invariant") across both situations, and the learner must learn (become "attuned" to) these *invariants* in the initial learning environment. Our analysis revealed numerous transformations between GenScope and the NewWorm, including transformations of media, organism, genome, characteristics, and social setting. For example, one transformation concerned the way the organism's genotype was represented. GenScope's chromosome window (shown in Figure 1) provides a colorful detailed depiction of alleles, whereas the NewWorm assessment items use the traditional "stick figure" representation. If students' understanding of genotype and chromosome representation is to transfer to the assessment environment, they need to become attuned to both the aspects of the environment that are particular to GenScope (the chromosome "window") and the aspects that are invariant (i.e., the domain-relevant information that is conveyed by both representations). Given the number and nature of the transformations in this study, it was indeed possible that students were becoming attuned to the invariant aspects of the GenScope environment; but there were too many transformations between that environment and the NewWorm.[8]

### Transfer Sub-Study

Following the initially disappointing results in the first set of pilot implementations and the disagreement over transfer, we developed an additional set of outcome measures. Screen captures were used to create analogues of the NewWorm items, but using the GenScope organisms, traits, and representations that students would be quite familiar with. This part of our effort is best understood according to the multilevel analysis of instructional sensitivity provided by Ruiz-Primo, Shavelson, Hamilton, and Klein (2002). From this view, the GenScope students' completion of activities and quizzes represented *immediate* assessments of student learning (i.e., artifacts from the enactment of the curriculum). At the next level, our new assessments provided *close* evidence of learning. Although designed to be formally administered, the close-level assessments were also designed to be extremely sensitive to the content and activities of the GenScope curriculum. Our NewWorm assessment was more akin to what Ruiz-Primo et al. called a *proximal* assessment, "designed to consider knowledge and skills relevant to the curriculum" (p. 371).[9] Thus our new "close-level" classroom assessments were counterbalanced and administered alongside the "proximal" NewWorm items in the second pilot implementation. The pattern of results in the sub-study supported

the view that students were not learning underlying genetics concepts in the GenScope environment. For example, 22 of the 44 students could solve the "close-transfer" effect-to-cause, between-generations, autosomal inheritance problems; they included 19 of the 20 students who could solve the corresponding "proximal" assessment item. And again, none of the students could solve close-level transfer inheritance problems involving the sex-linked characteristic. If students were learning the underlying domain concepts in GenScope but could not transfer them to the NewWorm, we would have seen greatly improved performance on the close transfer items.

Given the results on the close-level assessments, we all agreed that although students were clearly learning in the activities during the pilot implementation, they were not learning aspects of the GenScope environment that were thought to represent meaningful domain understanding. In the words of Greeno, Smith, and Moore (1993), it appeared that students were not "becoming attuned to the invariants" in the GenScope learning environment. Rather, the learning that was occurring, as indicated by their increasingly skillful participation in the GenScope activities, concerned the "variant" aspects of the environment. As such, the learning was very specific to the GenScope environment and its specific actions and features.

### Revised Curriculum and the "Dragon Investigations"

As part of ongoing enhancements, with the added incentive of the disappointing initial learning outcomes, the development team revised and enhanced the software and continued developing and refining curricular activities. New dragon characteristics were added in a way that did not allow students to view the underlying characteristics genome (necessitating effect-to-cause reasoning); a new enlarged window made it possible for students to witness and control meiotic events such as crossover that are essential for understanding inheritance. In addition, the various curricular activities were refined and additional activities were developed that focused on specific aspects of domain reasoning.

In keeping with contemporary perspectives on assessment and instruction (e.g., Frederiksen & Collins, 1989; Wiggins, 1993; Wolfe, Bixby, Glenn, & Gardner, 1991) the assessment team began developing ways to help students learn the specific reasoning skills that were being targeted in the assessment system then under development. The ultimate outcome of this effort was a set of formative assessments known as Dragon Investigations. In practice, these materials evolved from the close-transfer classroom assessment items. They were designed to be completed away from the computer, either as homework or in class. An example is shown in Figure 4. For each there was also a teacher's answer key that included detailed explanations of the relevant domain content in the context of solving the particular problem. The activities were designed to foster a focused whole-class discussion by building on the classes' shared, simplified understanding of the domain as represented by the dragon genome and the various GenScope windows. The ultimate set of



**DRAGON INVESTIGATION #11**

**"From Offspring to Mode of Inheritance"**

We often don't know the genotypes of individuals or the genetics of the species for a particular characteristic. One way to figure out the genetics of a particular characteristic is to carefully study the patterns of inheritance of phenotypes.

**Fangs**

**Another inherited characteristic in dragons is Fangs. Both Sandy and Pat have no fangs. But when you look at 100 of their offspring, you find the following:**

- 29 (13 males and 16 females) have fangs
- 71 (37 males and 34 females) have no fangs

Use the information about the offspring to explain the mode of inheritance. Remember that in dragons, males are XX and females are XY.

1. The Fangs gene has two alleles–*fangs* and *no fangs*. The relationship between the two alleles is **simple dominance** (rather than incomplete dominance).

   What is it about the **offspring phenotypes** that indicates that the relationship is simple dominance?

2. The *no fangs* allele is **dominant** to the *fangs* allele (rather than the *no fangs* allele being recessive or incompletely dominant to the *fangs* allele).

   What is it about the **offspring data** that indicates that the *no fangs* allele is dominant to the *fangs* allele?

3. The gene for Fangs is **autosomal** (rather than X-linked).

   What is it about the **offspring data** that indicates that the Fangs gene is autosomal?

*Figure 4.* **Condensed example of a Dragon Investigation formative assessment targeting effect-to-cause, between-generations reasoning.**

11 activities was carefully sequenced across increasingly complex aspects of inheritance and increasingly expert kinds of domain reasoning.

## Large-Scale Implementation and Evaluation

During Years 2 and 3, we formally implemented the GenScope curriculum and the NewWorm, assessing learning outcomes in secondary schools in metropolitan areas in New England and in the southeastern United States.

### Study Overview

#### *Participants*

The participating teachers came from eight different schools and were recruited in various ways: through GenScope workshops, school systems, and e-mails to individuals who had independently downloaded the software from the website. Table 2 describes the 31 classes taught by 13 teachers, where students were assessed before and after genetics instruction. As described above, the NewWorm precursor was piloted during Year 1 in 11 classes and is not included here.[10]

#### *Research Design*

Eight of the classes listed on Table 2 are comparison classes. The way we used comparison classes reflects a shift beyond merely demonstrating that new technology can be used to teach as effectively as, or more effectively than, conventional methods:

> The probability that elementary and secondary education will prove to be the one information-based industry in which computer technology does not have a natural role would at this point be appear to be so low as to render unconscionably wasteful any research that might be designed to answer this question alone. (PCAST, 1997, pp. 93–94)

This suggests that the use of tools such as GenScope to teach introductory genetics is inevitable. Of course, it was important that we show that the GenScope curriculum was at least as effective as the curriculum that it was designed to supplant. But we also wanted to be sure that the comparisons *within* various GenScope classes would enhance our understanding of how this specific tool, and these types of tools in general, could enhance student learning (as suggested in Donovan, Bransford, & Pellegrino, 1999; and Collins, 1999).

Essentially, we used contemporary performance assessment methods and powerful psychometric techniques to measure gains in genetics reasoning ability in a broad range of GenScope and comparison classes. The development team carefully examined the teaching and learning experiences that

(*text continues on page 514*)

*Table 2*

**Study Classroom Characteristics, Implementation Details, and Learning Outcomes (Transformed to *M* = 50, *SD* = 10)**

| School and classroom description | | | | | Genetics curriculum[a] | | | NewWorm score | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Course context | School no. and type | Teacher and experience level[a] | Implementation status | No. of classes, students[b] | Periods on genetics[a] | Periods on GenScope | Number of Dragon Investigations | Number of other GenScope activities | Pretest | Posttest | Gain[c] |
| Ninth-grade Unified Science | School 1, urban | Mr. H, low | GenScope, Year 2 | 1, 11 | 25 | 25 | 0 | 15 | 31.7 | 46.0 | 14.3 |
| | | | Comparison, Year 2 | 2, 37[d] | 25 | 0 | 0 | 0 | | 47.6 | |
| | | | GenScope, Year 2, w/o computers | 1, 14 | 32 | 32 | 11 | 10 | 32.1 | 48.9 | 16.8 |
| | | | GenScope, Year 3, w/ computers | 1, 12 | 32 | 32 | 11 | 10 | 27.5 | 41.8 | 14.3 |
| | | Ms. L,[e] low | GenScope, Year 3, w/ computers | 1, 8 | 36 | 36 | 7 | 8 | 12.7 | 30.9 | 18.2 |
| | | | GenScope, Year 3, w/o computers | 2, 12 | 36 | 36 | 7 | — | 15.5 | 39.0 | 23.5 |
| | School 5, urban ESL | Ms. Q, medium | Comparison, Year 3[f] | 2, 31 | 20 | 2 | 2 | 2 | 36.0 | 40.1 | 4.1 |

(*continued*)

*Table 2*

**Study Classroom Characteristics, Implementation Details, and Learning Outcomes (Transformed to $M = 50$, $SD = 10$) (*Continued*)**

| | School and classroom description | | | | Genetics curriculum[a] | | | | NewWorm score | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Course context | School no. and type | Teacher and experience level[a] | Implementation status | No. of classes, students[b] | Periods on genetics[a] | Periods on GenScope | Number of Dragon Investigations | Number of other GenScope activities | Pretest | Posttest | Gain[c] |
| College Prep Life Science | School 2, suburban | Ms. L, low | Comparison, Year 2 | 1, 15 | < 20 | 0 | 0 | 0 | 40.3 | 50.4 | 10.1 |
| | School 5, urban ESL | Ms. P, high | GenScope, Year 3 | 2, 55 | 34 | 15 | 4 | 7 | 45.4 | 51.2 | 5.8 |
| | School 3, suburban/ industrial | Ms. M, high | GenScope, Year 3 | 1, 20 | 38 | 25 | 4 | 20 | 38.3 | 51.6 | 13.3 |
| Honors Life Science | School 3, suburban/ industrial | Ms. M, high | GenScope, Year 3 | 2, 32 | 29 | 20 | 4 | 20 | 45.8 | 54.7 | 8.9 |
| | School 4, suburban | Mr. B, high | GenScope, Year 3 | 1, 23 | 12[g] | 12 | 8 | 13 | 46.6 | 56.9 | 10.3 |
| | | Mr. D, high | Comparison, Year 3 | 1, 23 | 14[a] | 0 | 0 | 0 | 46.3 | 58.8 | 12.5 |
| | | | GenScope, Year 3 | 2, 39 | 14[a] | 14 | 5 | 13 | 44.6 | 59.1 | 14.5 |
| | School 6, suburban magnet | Ms. S, medium | GenScope, Year 3 | 3, 56 | 10[h] | 10 | 8 | 17 | 52.2 | 58.6 | 6.4 |
| General Life Science | School 2, suburban | Ms. B, medium | Comparison, Year 2 | 1, 18 | < 20 | 0 | 0 | 0 | 45.3 | 54.0 | 8.7 |
| | | Mr. R, high | GenScope, Year 3 | 1, 16 | 15 | 13–15 | 11 | 14 | 42.4 | 58.0 | 15.6 |
| | | | GenScope, Year 3 | 1, 16 | 15 | 13–15 | 11 | 14 | 46.4 | 57.1 | 10.7 |
| | School 3, suburban/ industrial | Ms. M, high | GenScope, Year 3 | 1, 10 | 35 | 25 | 2 | 12 | 33.5 | 43.3 | 9.8 |
| | | | GenScope, Year 3 | 1, 10 | 35 | 35 | 2 | 12 | 31.8 | 32.7 | 0.9 |
| | School 8, suburban/ rural[i] | Ms. F, high | Comparison, Year 4 | 1, 13 | 25 | 0 | 0 | 0 | 26.6 | 39.99 | 13.3 |
| | | Ms. T, high | GenScope, Year 4 | 1, 14 | 25 | 20 | 0 | 15 | 27.5 | 50.1 | 22.6 |
| | | | GenScope, Year 4 | 1, 18 | 25 | 14 | 6 | 10 | 22.5 | 53.2 | 30.7 |

[a] Self-reported on teacher survey.
[b] Includes only students who completed both pretest and posttest.
[c] Because of rounding, scores do not always represent posttest minus pretest.
[d] Posttest only.
[e] Included students identified as having learning disabilities and behavioral disabilities.
[f] This class started out as a GenScope classroom but abandoned the GenScope curriculum entirely after the 2nd day because of problems with computer access.
[g] Represents 90-minute class periods.
[h] Represents a class period > 50 minutes but < 90 minutes.
[i] Roughly half of the students in each of these classes was on an Individual Educational Program (IEP), having been identified as having a learning or behavioral disability.

were emerging in the context of the GenScope implementations and found that such experiences were simultaneously confronting teachers' and students' traditional classroom rituals and shaping new ones (e.g., Christie, 1999; Horwitz & Christie, 2000). The dynamic relationship between the assessment team and the development team continued throughout the implementation. Reflecting newer "design-based" approaches to educational research (e.g., Kelly, 2003), the induced and natural variations in GenScope implementations were not viewed through a conventional lens of "implementation fidelity." Rather, they were viewed as opportunities to build evidence-based understanding of various curricular enactments in a range of implementation contexts. We relied on comparison classes to give us a baseline for determining how much students in non-GenScope classes typically learned in introductory genetics. At the end of the implementation cycles, we carried out a more tightly controlled study of the factors that had emerged as key issues.

### *Curriculum*

At the beginning of Year 2, the Dragon Investigations and a subset of GenScope computer activities were organized into six curricular units around major-domain reasoning concepts (Introduction, Basic Inheritance, DNA & Meiotic Events and Inheritance, Two-gene Inheritance, Alignment & 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 to 5 GenScope computer activities and 1 to 3 Dragon Investigations. A package including a teacher guide and student worksheets was reproduced and distributed to the GenScope teachers. Some of the activities from the original set 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 targeted by the assessment practice. This is not a trivial point. These changes de-emphasized the more purely discovery-oriented activities that some contend are essential to learning scientific inquiry. Arguably then, the constraints on our assessment practice (especially the need for efficient large-scale administration) diminished participation in and learning of inquiry in the learning environment. (Alternatively, the constraints also provided more practice in the kind of systematic domain-specific reasoning that underlies the ultimate ability to successfully pursue more open-ended discovery-oriented activities). Across Years 2 and 3, teachers reported an average of 26 class periods on genetics, and the GenScope teachers reported an average of 21 periods working with GenScope. The GenScope teachers used an average of 12 of the GenScope curriculum activities developed by the development team and average of 6 Dragon Investigations; some of the GenScope teachers used all 11 of the Dragon Investigations, whereas others used none (see Table 2).

### *Data Collection and Analysis*

As shown in Table 2, we grouped the GenScope and comparison classes into four sets according to the broader course context in which introductory biology was taught: (a) general "unified" science, (b) general/technical biology, (c) college prep biology, and (d) honors biology. The primary statistical and interpretive analyses were comparisons within each of these sets of classes.

We had about 500 sets of pre and post assessments, which were scored by two graduate research assistants. Of the 87 individual items on the posttest, 32 required some sort of interpretation to score. For 12 of the 32 items, scores were dichotomous (right or wrong), and the other 20 items were given either no, partial, or full credit. Inter-rater reliability was examined by using 89 posttests scored by two research assistants. These scores came from four of the highest-scoring classes, ensuring a maximum number of completed responses on the more difficult items. However, even in this sample, there were too few responses on the most difficult items (the "Pedigree II" items and the "Process" items) to yield meaningful percentage agreement. Averaged across the remaining items, the percentage agreement was 0.76. The scaling results provided more comprehensive reliability evidence.

*Scaling.* Student scores for all of the classes shown on Table 2 were analyzed using multifaceted Rasch scaling with the Facets software (Linacre, 1989). This latent-trait modeling procedure locates each assessment item and each individual's score on a single linear scale. Given that we designed the NewWorm to capture a wide range of proficiency, scaling the results was a necessity. When more weight is given to items that were answered correctly by only a few of the respondents (who answered most of the other items correctly), scaling yields a much wider and much more accurate scale of proficiency than simply summing up correct answers.

Scaling also provides an estimate of the relative difficulty of each item and relative proficiency represented by each student's assessment performance on the same linear scale. For ease of interpretation, we transformed the scale scores to a *T* scale, with a mean of 50 and a standard deviation of 10. Scaling equates differences in item difficulty and individual proficiency regardless of where they appear, estimates the precision and reliability of the entire scale, and estimates the degree to which each individual's and each item's pattern of scores fit the expectation of the model. Our scaling results revealed a broad range of item difficulty and student proficiency, and high reliabilities for the proficiency scores.[11]

*Construct validation.* We examined mean difficulty of the five clusters of NewWorm items to confirm our assumptions about dimensions of domain reasoning and to provide a basis for interpreting scale scores. Figure 5 shows the relative difficulty of the NewWorm item in terms of the dimensions of reasoning shown in Table 1, confirming our assumptions about the development of proficiency in genetics. The within-generation items were easier than the between-generations items; cause-to-effect reasoning was easier than effect-to-cause reasoning, which in turn was easier than process reasoning.

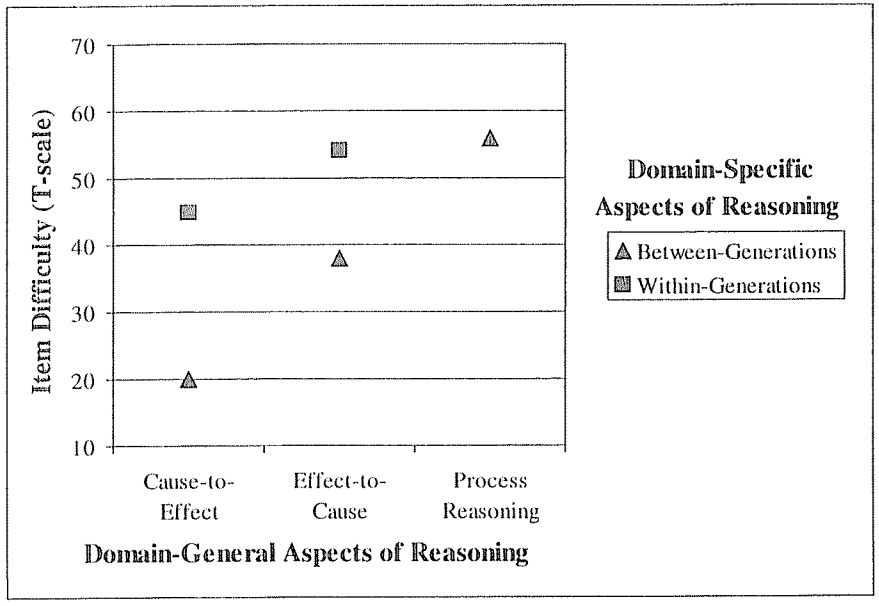



*Figure 5.* **Relative difficulty of clusters of NewWorm items, by reasoning type.**

We used the relationships depicted in Figure 5 to characterize gains in reasoning throughout the study. For example, the difference between the algorithmic cause-to-effect reasoning and the more expert effect-to-cause reasoning is roughly 20 points on the *T* scale, or 2 standard deviations in our sample. A change from 20 to 55 represents a fundamental shift from within-generation, cause-to-effect reasoning ability to between-generations, effect-to-cause reasoning ability. This is the type of dramatic learning outcome that proponents of innovations such as GenScope seek.

To further validate the assessment content and the range of reasoning skills captured, the NewWorm was administered to six pairs of students and faculty members in a college biology department (nonscience majors of any year; freshman, junior, senior, and graduate biology majors; and biology faculty). Figure 6 shows the mean score for each, along with the average pretest and posttest scores of the four different groups of high school students in our sample (described next). As expected, the increasingly more advanced college students and faculty demonstrated increasing proficiency.[12] Somewhat surprisingly, posttest performance for some groups of high school students reached the same level as for the college undergraduates. Results not presented here further validated assumptions about more specific dimensions of reasoning and other aspects of the domain (Kindfield, Hickey, & Wolfe, 1999); other data supporting the substantive and structural validity of





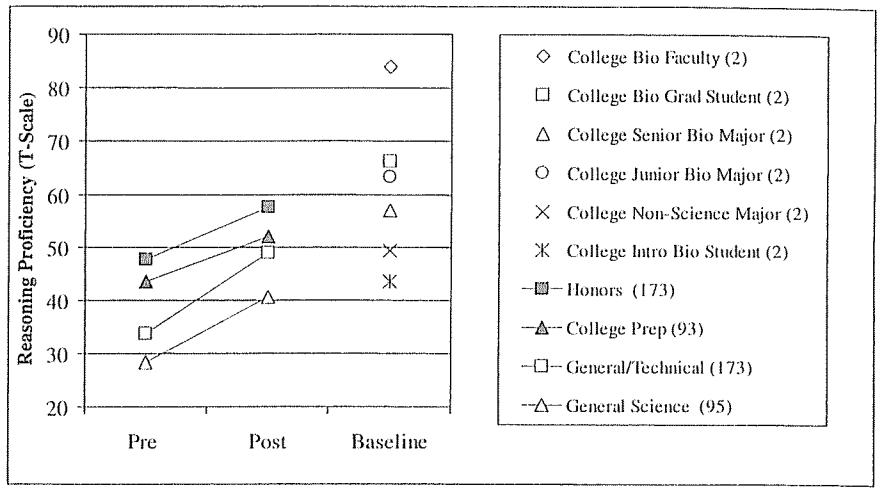



*Figure 6.* **Genetics reasoning gains in four instructional contexts and in six pairs of college biology students and faculty. Number of subjects in each group is in parentheses.**

the assessment practice (following Messick, 1994, and Shepard, 1993) were derived from think-aloud protocols and retrospective interviews (reported in Hickey, Wolfe, & Kindfield, 2000).

## Results

Given the diversity of implementation contexts, it is necessary that the outcomes in GenScope and in the comparison classes be compared within each of the four course types (Unified Science, Honors Life Science, College Prep Life Science, and General Life Science).[13]

### *Unified Science Classes*

Two teachers in a large, struggling inner-city school implemented GenScope and the NewWorm assessment in six classes following a district directive to incorporate genetics into the ninth-grade "Unified Science" curriculum. One of these teachers also administered the NewWorm after giving genetics instruction in one non-GenScope comparison class. A similar teacher at another school in the district abandoned GenScope but continued to administer the NewWorm.

At School 1, "Mr. H" had "modest" knowledge of biology and introductory genetics and was employed as a curriculum consultant to the GenScope development team. During Year 2, development team members were in the class nearly every day and provided most of the genetics instruction during the reported 25 days devoted to GenScope. For the 11 students who completed both the pretest and the posttest administration of the NewWorm, mean proficiency scores increased from 31.6 to 46.0, a gain of 14.4 points,

or 1.4 standard deviations. Figure 7 shows that the mean posttest score in the GenScope class was higher than mean posttest scores for students in Mr. H's other two general sciences classes (39.6, the open square). These two classes had a similar group of students, and Mr. H used a textbook, worksheets, and lectures to teach introductory genetics. The difference did not quite reach statistical significance, $F(1, 50) = 3.49$, $p = .067$; the lack of pretest data leaves pre-instructional differences unexplained, and instruction in the GenScope class was managed and delivered by individuals who were better prepared to teach introductory genetics.

In the summer following the Year 2 implementation, Mr. H completed a 40-hour GenScope teacher-training workshop. For Year 3, he took over the instruction in introductory genetics in both of his Unified Science classes, using the revised GenScope curriculum described previously to cover introductory genetics in both of his classes. However, only one of his classes had access to the computer lab. Partly because of this, Mr. H relied very heavily on the Dragon Investigation activities in both classes. During the 36 days that 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 class independently completed the GenScope computer activities, Mr. H would either go over the activities 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.

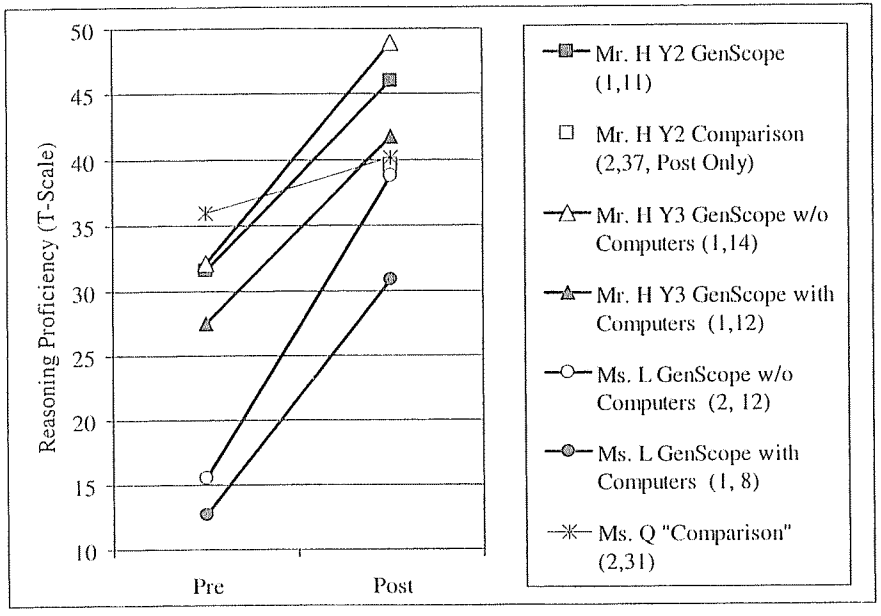



*Figure 7.* **Reasoning gains in Unified Science classes. Y = Year. Numbers of classes and students are in parentheses.**





Another teacher at School 1, "Ms. L," conducted a similar implementation in her three general science classes. Like Mr. H, Ms. L reported "modest" knowledge of biology and genetics. Her classes served students who had been identified as overcoming learning or behavioral disabilities and other students who were struggling to keep up with their peers. In all three classes, Ms. L closely followed the new curriculum package and reported relying very heavily on both the student worksheets and the teacher versions of the Dragon Investigations. Like Mr. H, Ms. L had limited access to the computer lab for her students: in just one class, her students were able to use it for one or two periods per week. When those students completed the GenScope computer activities, the other students completed the GenScope worksheets as whole class activities.

Figure 7 shows similar and substantial NewWorm gains in both of Mr. H's classes (the triangles) and in both sets of Ms. L's classes (the circles). Notably, the gains in the classes that had access to the computers (the closed triangles and circles) were somewhat smaller than the classes that did not (the open triangles and circles). However, neither of the differences in gains within the group taught by each teacher reached statistical significance (for both, $F < 1$). It was certainly noteworthy that Ms. L's two no-computer classes showed average score gains of 23.5. Although partly due to the very low pretest scores in those classes, that average gain was one of the largest gains observed in the study. Further investigation confirmed that the GenScope teachers encountered numerous difficulties in gaining access to the computer labs and carrying out the GenScope computer activities. The difficulties included scheduling changes, hardware and software problems, and confusion and problems with some of the GenScope computer activities.

We compared the students at School 1 with a pair of classes at another school in the same district; overall achievement was slightly higher, but the school served many non-native English speakers. "Ms. Q" was an experienced biology teacher who described herself as "very comfortable" teaching genetics, and she elected to participate in the study after participating in the summer workshop. However, difficulties with the software and the computer lab led her to entirely abandon GenScope on the 2nd day. She reported spending an additional 18 class periods covering genetics using a mix of lectures, demonstrations, textbooks, and worksheets. As shown in Figure 7 (the asterisks), these students gained only 4.1 on the NewWorm—significantly less than the gain across the five GenScope classes at School 1, $F(1, 82) = 8.75$, $p = .004$. However, Ms. Q answered "not very seriously" to the survey question, *How seriously did your students take the GenScope assessment?* Like most other teachers in the study, Mr. H and Ms. L reported that their students took the assessment "seriously." Although Mr. H reported assigning a grade to student performance on the assessment, her students may indeed have tried less seriously on the posttest than did the GenScope students.

These results support the overall effectiveness of the GenScope curriculum in urban Unified Science classes, compared with the curriculum that GenScope was designed to supplant or replace, in terms of the learning out-

comes assessed on the NewWorm. This is noteworthy, given that these students represent a population that is particularly at risk for academic failure. The small number of scores for each class in School 1 was largely due to the high absenteeism and turnover in those classes. Of the students who left the classes during the term, the teacher reported that several had become pregnant or incarcerated. Perhaps the most interesting is the somewhat surprising finding that students who used the GenScope curriculum without actually completing the GenScope computer activities showed the same reasoning gains as their schoolmates who completed the computer activities. It appears that while one group of students was independently trying to understand complex concepts under challenging conditions, their schoolmates were participating in a teacher-managed classroom activity that targeted the same concepts, taking advantage of the shared representation of the GenScope dragons. This finding provided an early indication both of the challenges of implementing GenScope in computer labs and of the value of whole-class discourse around a well-understood organism.

### *Honors Life Science Classes*

Four teachers in three schools implemented GenScope in eight honors classes, and one of these teachers administered the NewWorm in a non-GenScope comparison class.

At School 4, both "Mr. D" and "Mr. B" were pursuing science education doctorates. They reported "above average" knowledge of genetics and comfort in teaching it, but "low comfort" with integrating computer-based activities into their curriculum. Mr. B implemented GenScope in one course, and Mr. D implemented it in two; Mr. D's other honors biology course provided a seemingly ideal comparison class.

Mr. D described his teaching style as being "relatively socio-constructivist" and encouraged his students to approach the computer activities as inquiry learning activities. Five of the Dragon Investigations were assigned as homework and later reviewed in class. As at School 1, Mr. D reported various difficulties with the computer activities, including software glitches, computer crashes, and access problems, leaving his students somewhat frustrated. Figure 8 shows that despite these difficulties, Mr. D's GenScope students (solid squares) gained 14.4 points (14.3 in one class and 14.7 in the other). In the comparison class, Mr. D reported successfully using his normal mix of lectures, discussion, chalkboard diagrams, and textbook-reading and problem-solving assignments to teach the concepts in the GenScope curriculum (which corresponded closely to the genetics concepts that he normally covered). Figure 8 shows that these students (open squares) reached the same level as his GenScope class. Although they showed a smaller increase, the difference did not reach statistical significance, $F(1, 60) = 1.38$, $p = .245$.

In contrast to Mr. D, Mr. B indicated that his approach was more consistent with direct instruction practices. During fourteen 90-minute class periods devoted to genetics, he presented the GenScope activities as a way of illus-



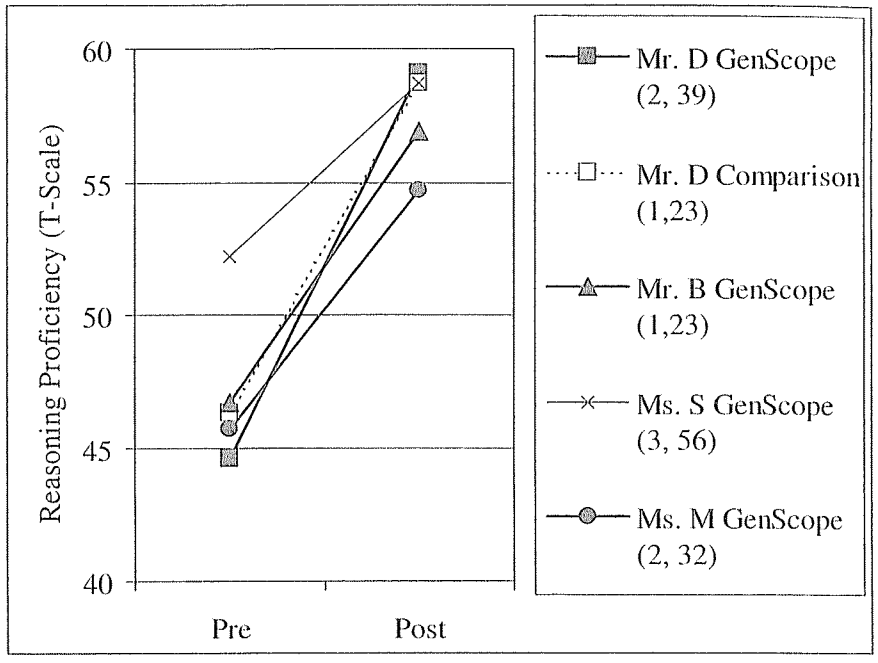



*Figure 8.* **Reasoning gains in Honors Life Science classes. Numbers of classes and students are in parentheses.**

trating and reinforcing concepts that he first introduced through lectures and textbook assignments. Like Mr. D, he reported that his students experienced frustrating computer problems. Eight of the Dragon Investigations were assigned as homework and later reviewed in class. As indicated by the triangles on Figure 8, Mr. D's students gained 10.3 points. This gain was significantly smaller than in Mr. D's GenScope classes, $F(1, 60) = 5.51$, $p = .022$, but not in Mr. D's comparison class, $F(1, 44) = 1.35$, $p = .252$.

Two teachers at two other schools implemented GenScope in honors biology classes during Year 3. One of these teachers, "Ms. S" (School 6) was a very experienced biology teacher who taught three sections of International Baccalaureate (akin to advanced placement) biology to 11th and 12th graders at a suburban science magnet school. They spent only 10 periods on GenScope. Showing the highest mean pretest scores in the study, 52.2, the scores increased only 6.4 from pretest to posttest (Figure 8), with similar gains across all three classes. We concluded that the students had already learned much of what was covered in the curriculum. At School 3, "Ms. M" had incorporated the GenScope software into her biology curriculum after independently obtaining it on the Internet. After she participated in (and helped to facilitate) the summer workshop, all five of her biology classes participated in the study. In her two honors classes, she reported covering genetics during roughly 30 class periods spread across an entire semester, including

4 Dragon Investigations and 20 GenScope activities.[14] These students (Figure 8, the circles) gained 8.9 points, with very similar gains across the two classes.

We concluded that although the gains in these classes were smaller than in Unified Science, these teachers had devoted substantially less course time to genetics than did the general sciences teachers. Furthermore, *pretest* proficiency for the honors classes was generally higher than *posttest* proficiency in the general sciences classes. This is not surprising given that some of the honors students were 11th and 12th graders and some had already completed one life science course. Unfortunately, computer labs again presented challenges to the GenScope students. Thus, again, while the students in the GenScope classes were struggling to learn independently under somewhat difficult circumstances, the comparison students were engaged in focused classroom learning directed at the same topics. Nonetheless, because GenScope was relatively new and the computer problems would eventually be solved, we conclude that GenScope is a viable option for use in typical honors classes.

### *College Prep Life Science Classes*

Most students in the United States learn introductory genetics in college prep life science courses. These classes typically have a handful of students overcoming learning disabilities, and sometimes do not include the highest achievers. One of the college prep GenScope classes was taught by Ms. M at School 3 (described previously). As shown in Figure 9 (circles), these students gained

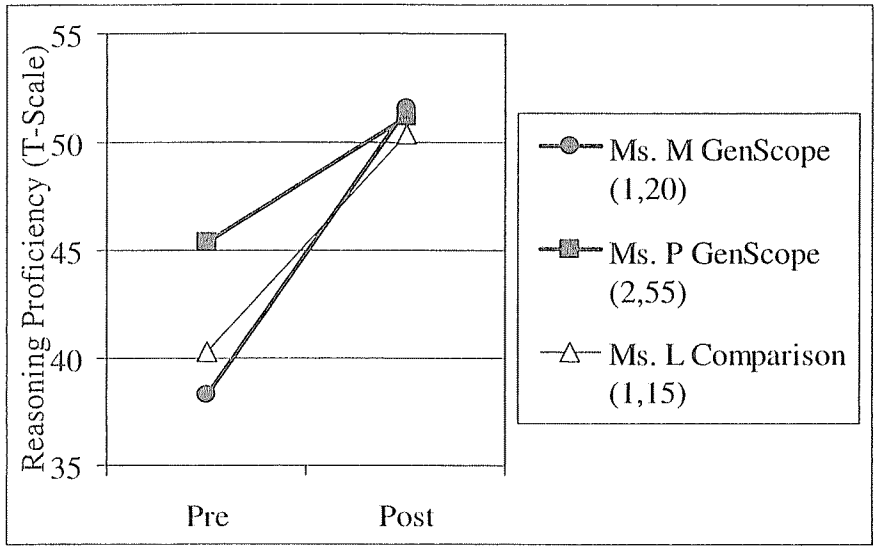



*Figure 9.* **Reasoning gains in College Prep Life Science classes. Numbers of classes and students are in parentheses.**



13.0 points, which is somewhat higher than average. Another college prep implementation occurred at School 9, a language magnet school in the same urban school district as Schools 1 and 5. The teacher, "Ms. P," described herself as being "extremely comfortable" teaching genetics. She participated in the summer workshop but reported being "not very confident" about integrating computer technology. Like Ms. M, she used genetics partly to organize other biology content, inflating the number of class periods devoted to genetics (34). Ms. P used seven of the GenScope computer activities and four of the Dragon Investigations to replace the lab sessions that were normally used to exemplify and extend the topics that had already been introduced in class discussion. As shown in Figure 9 (squares), Ms. P's students showed a modest gain of just 5.8 points on the NewWorm, with similar gains across the two classes. This gain was substantially smaller than the gain in Ms. M's prep class, but the difference did not quite reach statistical significance, $F(1, 68) = 3.52$, $p = .065$. The modest gains by Ms. P's students may be qualified by the report that her students were upset about having to complete the NewWorm assessment, particularly about its impact on their grade. She had informed them that their performance "would not lower their course grade but would count as extra credit."

During Year 2, comparison data were collected from one college prep class at School 2, which served a relatively advantaged suburban population. Ms. L 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, that teacher is no longer at the school, and efforts to obtain additional information about the curriculum or how many class periods were devoted to it were unsuccessful. As shown on Figure 9 (triangles), Ms. L's students showed a fairly typical gain of 10.1. This gain was smaller than that of Ms. M's GenScope class, but not significantly ($F < 1$); it was larger than Ms. P's GenScope classes, but not significantly, $F(1, 73) = 2.94$, $p = .091$.

We conclude that, although the results in the college prep classes again show that GenScope is an effective environment for developing domain reasoning skills, it was no more so than an existing comparison curriculum. It is noteworthy that the comparison curriculum in this case was a self-paced, programmed instruction module developed by a former biology teacher at the school. However, there is insufficient information about that curriculum or how much time students devoted to it—or even the circumstances under which the assessments were administered.

### *General Life Science Classes*

General Life Science (also called "ABC" or "technical") classes typically include many students identified as overcoming behavioral or learning disabilities, or both. One comparison teacher, Ms. B at School 2, had her students use the locally developed programmed instruction unit to cover genetics in her general biology class. As shown in Figure 10, the mean

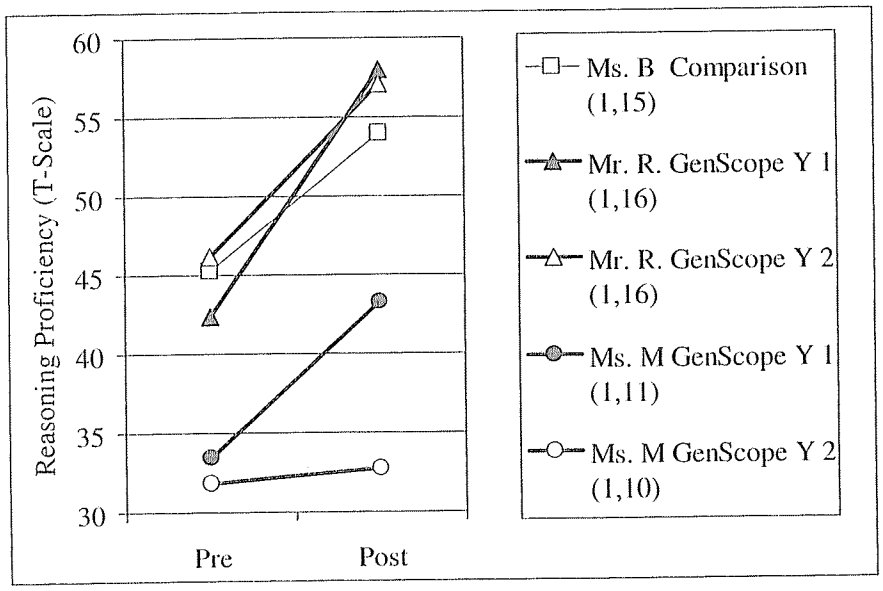



*Figure 10.* **Reasoning gains in General Life Sciences classes. Numbers of classes and students are in parentheses.**

NewWorm scores for Ms. B's student (the open squares) increased 8.7 points. Another teacher at School 2, "Mr. R," implemented the GenScope curriculum in his two general biology classes. Mr. R was the science program director and reported "above average" knowledge of genetics and comfort with teaching it. Having spent a prior sabbatical working with the GenScope development team, he was "very comfortable" integrating technology into his biology curriculum. Roughly half of the students in the two classes were on individual educational programs, and both classes where characterized as "challenging." All 11 Dragon Investigations were assigned as ungraded homework. Although half of the students reportedly did not complete them, Mr. R used them in an intensive review prior to the posttest. As shown by the triangles in Figure 10, one class gained 10.7 and the other gained 15.6, a difference just reaching significance, $F(1, 30) = 4.51$, $p = .042$. The gain in Mr. R's two GenScope classes together was not significantly larger than that in Ms. B's comparison class, $F(1, 48) = 2.40$, $p = .128$, but the difference between the gains in Mr. R's higher-gaining class and Ms. B's comparison class did reach statistical significance, $F(1, 32) = 6.16$, $p = .019$.

A second GenScope implementation was in two of Ms. M's classes at School 3 (described above). These 20 students showed a disappointing 5.4 gain. However, as shown in Figure 10 (circles), the differences in gains in these two classes differed substantially, 9.8 as opposed to 0.9, although the small sample and within-group variance precluded statistical significance, $F(1, 18) = 2.8$, $p = .149$. Further examination of the second class revealed that scores for 4 of the 10 students actually declined, with some students



missing items on the posttest that they answered correctly on the pretest. Along with highly suspect fit statistics, this raised obvious concerns about the posttest conditions.

### Conclusions from the Large Scale Implementation and Evaluation

As a whole, the results described above show that many teachers were able to use GenScope effectively to enhance students' ability to reason about introductory genetics. Although we found statistically larger gains only in the Unified Science and General Science course contexts, the findings in all course contexts were quite promising. Clearly, the numerous challenges presented by using the software in a remote computer lab served to constrain the effectiveness of the GenScope software. This setting effectively removed the teacher from the instruction and presented numerous problems with access, hardware, and software. Remote labs, and the model of practice that they constrain, essentially require the computer to teach the children. In contrast, the PCAST report and other influential reports on teaching and learning (e.g., NRC, 1999a) emphasize the ways that technology can be used to amplify existing classroom practices.

We also concluded that the Dragon Investigation formative assessments were a promising addition to the GenScope curriculum. They were clearly effective for building on the shared understanding afforded by the GenScope environment—even to the point that in some classes, scores on the NewWorm increased more when teachers relied exclusively on the Dragon Investigations without using the GenScope software at all. At least in terms of the kinds of reasoning assessed by the NewWorm, it seemed that the real value of the GenScope environment was that it offered teachers and students an understandable but sufficiently complex context in which to discuss and learn introductory genetics.

Our formal collaboration was now over. Despite substantial progress and promising findings, several issues remained to temper our conclusions. In addition to the obvious problems with using the software in a computer-lab setting, other issues concerned the difficulty of identifying "fair" comparison classes. Although Mr. D's honors classes promised an ideal quasi-experimental, within-teacher comparison, there was a clear "carryover" from the GenScope curriculum into comparison classes. Another issue was the relationship between the Dragon Investigations formative assessment and the NewWorm summative assessment. The Dragon Investigations were shown to increase NewWorm performance substantially, but we could not rule out a "training effect." Following the validity model outlined by Messick (1994), it was possible that the Dragon Investigations had introduced "construct irrelevant variance" by excessively familiarizing GenScope students with the NewWorm format and content.

### Follow-up Study

An additional implementation was undertaken during the year after the original 3-year project was completed. This study was conducted in three classes at School 8, a suburban/rural school that served a broad range of students.

This implementation was carefully designed to address three unresolved issues from the previous implementations. The first issue concerned the problems that most students encountered in completing the GenScope activities in computer labs. The second issue concerned the degree to which the Dragon Investigations had undermined the evidential validity of the NewWorm. The third issued concerned the validity of Mr. D's honors biology comparison class, where the students may have benefited from the organizational guidance of the GenScope curriculum but avoided the substantial computer lab challenges.

*Design*

Three General Life Science classes at School 8 served a single pool of technical track (i.e., non-university-bound) students, with roughly half of the students in each class on Individual Educational Programs for learning or behavioral problems. "Ms. T" implemented GenScope in two of the classes. Ms. T was a 1st-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. Thus she was very familiar with the reasoning targeted by the NewWorm. Addressing the issue of computers, the GenScope activities were further refined and debugged, and the students completed them on 10 laptop computers installed in Ms. T's wetlab classroom. Addressing the issue of the Dragon Investigations, one of Ms. T's classes completed 15 GenScope computer activities (and no Dragon Investigations) over approximately 25 class periods devoted to genetics. In contrast, the other class completed just 10 of the GenScope computer activities and 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 paper-and-pencil activities designed to teach very specific aspects of domain reasoning. Regular observations and daily videotapings showed that Ms. T nevertheless initiated many whole-class discussions that were fairly similar to the discussions that were scaffolded by the Dragon Investigations. Thus there was some carryover of the broader goals of the Dragon Investigations, but those students were never exposed to NewWorm-type items and formats.

Addressing the third issue concerning the carryover effects of the GenScope curriculum and the associated lack of valid implementation or 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 (which generally followed the district's standardized curriculum). She was encouraged to do her very best, using the methods that she normally used (lectures, worksheets, textbook, and discussion) to help during the same number of class periods as in Ms. T's GenScope classes.

*Results and Conclusions*

Observations revealed few technical difficulties with the GenScope computer activities or software in either class. Furthermore, during the computer activities, Ms. T and a paraprofessional wandered among the students to answer questions and keep students on task. Ms. T provided a brief introduction to most computer activity and would sometimes call the class to attention to review or clarify a particular point. Reflecting the number of behaviorally disabled students and overall low achievement, videotapes of the classes revealed a good deal of "horsing around" during the computer activities, fairly extended stretches of off-task activity, and substantial effort devoted to maintaining order.

Figure 11 shows an above-average gain of 13.3 in Ms. F's comparison class (triangles). Much to our satisfaction, the mean scores in Ms. T's GenScope class that did not use the Dragon Investigations (squares) increased by 22.6 points. Even more impressively, Ms. T's class that used the Dragon Investigations (circles) gained 30.7 points, the largest gains of any class in the study. The difference between the gains in these two GenScope classes just reached statistical significance, $F(1, 30) = 5.0$, $p = .033$. The combined gains in the two GenScope classes were significantly greater than the gains in the comparison class, $F(1, 43) = 15.7$, $p < .001$.

Given Ms. T's knowledge of the GenScope curriculum and continued refinements of the GenScope curriculum, this was one of the most successful implementations of GenScope ever undertaken. We believe that it demonstrates the potentially dramatic knowledge gains possible when teacher knowledge,

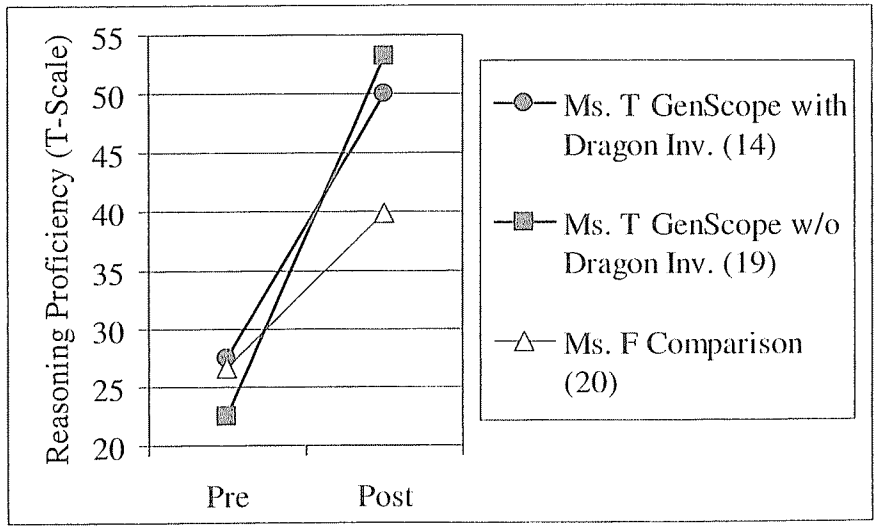



*Figure 11.* **Reasoning gains in General Life Sciences classrooms during follow-up implementation. Inv. = Investigation. Number of subjects in each group is in parentheses.**

curriculum, technology, classroom assessment, and external assessment are aligned towards well-defined, ambitious goals.

We also concluded that the Dragon Investigation presented a small and very acceptable degree of compromise to the NewWorm's evidential validity. The smaller gain in the GenScope class that did not complete the Dragon Investigations suggests that these activities do indeed train students to do better on the NewWorm. Nonetheless, there should have been a much larger difference in the two gains if the Dragon Investigations had more fundamentally compromised performance on the NewWorm (by reducing the complexity of the problems to the degree that they could be solved algorithmically). Because the organism, genotypes, and phenotypes of the two instruments (and the format of some of the problems) were entirely different, it appears that the Dragon Investigations had precisely the desired effect: developing transferable domain reasoning skills.

## Discussion and Conclusions

### The Impact of the GenScope Learning Environment

We first reflect on the GenScope learning environment's potential for developing students' ability to reason about introductory genetics. The findings suggest that the GenScope curriculum is indeed useful for addressing this challenge. We documented worthwhile reasoning gains in nearly every GenScope class; relative to their comparison classes, the gains in general science classes and the general/technical life science classes were statistically unlikely to have occurred by chance. The follow-up study at School 8 provided dramatic evidence of the potential of this learning environment. In light of the dimensions of domain reasoning described earlier, mean performance went from reasoning at the within-generation, cause-to-effect level to the between-generations, effect-to-cause level. These gains were far larger than those in any of the comparison classes. We argue that this represents fundamental, qualitative improvement in domain-specific reasoning.

Nonetheless, several factors preclude a direct conclusion that GenScope "works better" than the conventional approaches that it was designed to supplement or supplant. Some of these factors are common to quasi-controlled, school-based evaluations. These include the difficulty of establishing "fair" comparison groups, temporary local obstacles to the innovation, the conditions under which research instruments are administered, and the localized effects of teachers and classes where the innovation is enacted. Other confounding factors were specific to our effort. This included the challenge of using the computer labs and the surprisingly large gains in several comparison classes that used a locally developed programmed instruction module. The most important specific factor was the curricular revisions carried out in a focused effort to improve scores on the NewWorm performance assessment. As we point out, these changes represented a modest narrowing of the curriculum along the lines of our assessment and a reduced emphasis on



the more discovery-oriented aspects of the environment (which was not as readily assessed within our constraints). We return to this point shortly, as it goes to the heart of another conclusion.

Another point worth consideration is the seemingly positive consequences of using laptop computers in the classroom to complete the GenScope activities (as opposed to computer labs). Other factors certainly contributed to the success of the GenScope classes in the follow-up study. But it seems clear that the classroom implementation allowed Ms. T to scaffold and structure student discourse and inquiry and to stop and start the computer activities in a way that supported engagement and learning. Although we lack detailed data in this regard, we suspect that many of the other teachers sent their students to the computer lab on their own or stayed in the lab long enough to get students started. Using laptops in the classroom appeared to better situate each group's completion of the GenScope activities within the larger classroom context. It is certainly possible to create such participation structures in computer labs, but their typical configuration (i.e., computers arranged against the walls) and history of use (for isolated activities by individuals or small groups outside class) seem to discourage the kind of whole-class discourse among collaborative groups that appeared so powerful in the follow-up GenScope classes.

### The Potential of Classroom Assessment

Our efforts were strongly influenced by contemporary assessment research. Much of this research was subsequently included in two recent NRC reports: *Classroom Assessment and the National Science Education Standards* (NRC, 2001a), and *Knowing What Students Know: The Science and Design of Educational Assessment* (NRC, 2001b). Several of our primary conclusions buttress recommendations in these reports. For example, the first report concludes that classroom assessments can powerfully enhance learning and teaching—provided that they are accompanied by feedback that learners use to advance their understanding and that teachers use to evaluate and refine their instructional practices. In the case of our Dragon Investigation classroom assessments, many of the teachers (especially those with only modest knowledge of genetics) reported that the detailed answer explanations that we provided were quite useful in both regards. Although our study convinced us of the value of such practices, it also pointed to the need for additional research. For example, our study highlighted the issue of *formality*. Compared with some approaches, our classroom assessments were relatively formal events. They were paper-and-pencil activities that were administered at the end of instructional units, often for a grade. This formality appeared to motivate students to work, which seems crucial if students are to benefit fully from feedback. We wonder whether more informal approaches, where assessments are more seamlessly embedded in the curricular activities, might be less effective for this reason. Meanwhile, within more formal approaches like ours, we wonder about several issues, such as different ways of grading classroom assessments.

Both of the NRC assessment reports argued that learning and achievement are increased when classroom assessment and external testing are better aligned; a new committee recently established by the NRC's Board on Testing and Assessment (NRC, in press) is focusing directly on this issue. By using classroom assessments to highlight and refine alignment between the GenScope activities and the NewWorm assessments, we made significant progress in what has traditionally been a very challenging topic for secondary life science students. In retrospect, we found that the framework advanced by Ruiz-Primo, Shavelson, Hamilton, and Klein (2002) was invaluable for characterizing the distance between our various measures and the enacted GenScope curriculum (i.e., *immediate, close, proximal, distal,* and *remote*). We also found this framework useful for considering how we might extend our effort to include more distal measures, such as conventional high-stakes tests. For example, this framework helps to illuminate the common practice of "cherry-picking" items from among existing high-stakes items. Once selected from a larger "distal" instrument, such items become more proximal, limiting claims about the generalizability of resulting scores. This issue is currently being explored in another GenScope study that is still under way (Hickey, 2001). In addition to the Dragon Investigations and the NewWorm, this study also includes a battery of quasi-randomly selected items from a larger pool of released high-stakes items covering genetics.

Both NRC reports also highlight the need to use current knowledge about the development of domain expertise when creating (or selecting) and aligning assessments. Our NewWorm assessment was organized around a research-based model of the development of domain expertise. Crossing a domain-general dimension of expertise with a domain-specific dimension yielded the simple matrix in Table 1; we then used the resulting framework to (a) validate, interpret, and communicate NewWorm scores; (b) create our classroom assessments; and (c) realign the existing curriculum. This framework took our entire effort in a different direction than a conventional textbook scope and sequence, or established science education standards, would have done. We believe that our approach provides a useful model that should be readily adaptable for a range of innovations in other content domains.

Our assessment practice raises other important issues that are central to current assessment research. Our NewWorm assessments led to a more narrow curriculum and less focus on the purely discovery-oriented activities that many see as essential for learning scientific inquiry. Although this was trivial in comparison with the narrowing caused by conventional large-scale assessment, some would still consider it a compromise. Thus we wonder how the curricular revisions might have proceeded had we presented the assessment item in Figure 3 without the scaffolding of the three discrete questions. The resulting item would surely have been more difficult to score reliably, but the corresponding changes to the classroom assessments and GenScope activities might well have led to more engagement in inquiry. Leading scholars have made substantial progress in the assessment of inquiry (e.g., Duschl, in press; White & Fredericksen, 1998; Wilson & Sloane, 2000).



Yet most of these efforts are more akin to "close-level" classroom assessments, in that they are embedded into the fabric of the curriculum. Our findings suggest value in finding ways to include such assessments at the more "proximal" and "distal" levels as well (akin to the level of our NewWorm and beyond). This would allow them to be validly administered to comparison classes and might eventually have more influence on large-scale, high-stakes tests.

We also acknowledge that our classroom assessments partly compromised the evidential validity of our NewWorm assessments. Specifically, our initially "distal" NewWorm became more "proximal" once we introduced the Dragon Investigations and realigned the curriculum. We believe that our results show that it is possible to sacrifice a small, measurable degree of evidential validity in exchange for increases in the positive consequences of the assessments practice. From our perspective, sacrificing the pedagogical and motivational power of classroom assessment to maximize evidential validity is potentially inappropriate and unethical. Given that learning is the ultimate goal of educational systems, we concur with the arguments of Frederiksen and Collins (1989), Shepard (2000), and the NRC (2001a) supporting "systemically valid" assessment practices that emphasize the entire range of positive assessment consequences while attempting to minimize the negative consequences. We believe that our study suggests new ways to understand and accomplish this process.

Both of the NRC assessment reports argue that technology promises to dramatically improve assessment practice. Some of the ideas developed in the present effort are being put in place within the much more sophisticated BioLogica software that Horwitz and colleagues are developing. BioLogica takes advantage of more powerful computers and software tools (e.g., Java) that make it possible to build assessments and feedback directly into the software environment. Such sophisticated environments present the intriguing possibility of using technology-based tools to manage the logistical challenges of providing formative feedback, while simultaneously collecting useful summative data.

## New Models of Theory Development and Research Collaboration

Our study's ostensible goal was evaluating the GenScope software and curriculum. Given that many observers and policymakers are still skeptical about the value of such tools, we believe that the evidence outlined above is useful new knowledge. With the guidance of contemporary perspectives on evaluation, assessment, and educational research, we were able to contribute considerably more knowledge that should ultimately be even more useful. It seems to us that narrowly defined views of program evaluation (e.g., Chemlinsky, 1998; Sechrest, 1992) are of very limited utility for exploiting the educational potential of technology.

As elaborated in more detail in Hickey, Kindfield, Horwitz, and Christie (1999), our collaboration embodied new pragmatic approaches to the conduct of educational research and the development of educational theory. In

addition to the PCAST report described above, two reports that were released after our collaboration began called for systematic, sustained collaboration among educators, developers, and researchers—communities that heretofore have pursued relatively distinct agendas. The National Educational Research Policy and Priorities Board (1999) called for "extended collaborative efforts directed at pressing practical problems" and "developing and testing general principles of education that can travel to locations beyond where the research was done" (p. 26). A separate report, by the NRC (1999b), called for renewed efforts to incorporate research on cognition, development, and learning into educational practice, helping educational institutions to continuously improve their practice and increasing the use of research knowledge in educational institutions. We believe that our effort provides a useful illustration of this type of research. Consider, for example, the relationship between the development team and the "outside" assessment and evaluation team. This was clearly a departure from the typical approach of dealing with assessment and evaluation once implementations are planned or under way. The resources devoted to assessment and evaluation were constant across the project; the collaborative environment allowed the assessment team to revise and extend the curriculum in ways that the development team might have otherwise resisted. We do not mean to imply that our collaboration was free of argumentation. In addition to the disagreements described above, there were numerous other issues that emerged between and within the two teams. But the iterative nature and extended duration of our collaboration allowed us to use empirical methods and theoretical arguments to settle these disagreements, contributing potentially useful new knowledge to others who face similar issues. Fortunately, as described in the next section, recent advances in research design suggest ways that this process might be dramatically streamlined.

A related point is that the NSF elected to fund three successive projects for a single development team targeting one long-standing educational problem. After funding the development of the initial version of GenScope, NSF provided additional funding for the implementation effort described here—contingent on the developer's collaboration with an assessment and evaluation team. The insights described in this article and others that emerged from the overall project were central to the subsequent development of the BioLogica software. Furthermore, the level of support for GenScope was sufficient to initiate a broad community of inquiry and practice around these tools. The tools have helped a community of educators, researchers, and developers to come together with shared goals for enhancing the learning of introductory genetics. This has led to worthwhile continuing activity within this community beyond the scope of the funded project—including work funded by other agencies and practical collaborations that are not externally supported. Our experience leads us to share the apparent enthusiasm of policymakers for this sort of effort.

### Design-Based Educational Research

In part, the policy reports described earlier reflect a more fundamental shift in the relationship between theoretical and practical work in educational



research (Lagemann, 1999). Leading researchers are increasingly attempting to develop scientific understanding while designing learning environments, formulating curriculum, and assessing learning. For many, coherence, parsimony, and predictive validity are no longer the sole questions or even the initial questions being asked of theories. Rather, the primary question is *whether the concepts and principles inform practice in productive ways.* As described by Greeno, Collins, and Resnick (1996),

> It becomes a task of research to develop and analyze new possibilities for practice, not just to provide inspiring examples, but also to provide analytical concepts and principles that support understanding of the examples and guidance for people who wish to use the examples as models in transforming their own practice. (p. 41)

This means that embedding research in the activities of practical reform should yield theoretical principles with greater scientific validity than those developed in laboratories or in disinterested observations of practice.

These new views of theory development are embodied in "design-based" approaches to educational research, through what have come to be called "design-experiments." Aspects of these approaches can be traced back to early "teaching experiments" by math educators (e.g., Steffe, 1983). Design-based methods were first fully articulated by Collins (1992, 1999) and Brown (1992) and are exemplified in the widely cited efforts of the Cognition and Technology Group at Vanderbilt University (e.g., 1997) and Greeno et al. (e.g., 1998). Recent collaborative efforts (i.e., Design-Based Research Collective, 2003; Kelly, 2003; Kelly & Lesh, 2000) have further clarified design-based methods and provide useful context for our study. The central notion is that the design of learning environments and the development of theories are "intertwined" and occur within "continuous cycles of design, enactment, analysis, and redesign" (Design-Based Research Collective, p. 10). Rather than simply evaluating the GenScope curriculum as it existed, we repeatedly applied scientific methods and our assumptions about learning to meet clearly defined expectations. In doing so, we developed some nascent theories that should generalize to a broader class of curricular innovations. It is in this sense that design-based methods view theoretical advance in terms of "prototheory" (Design-Based Research Collective, p. 10), targeting an "intermediate" theoretical scope (diSessa, 1991).

Our study also illustrates the value of specifying (a) the significant disciplinary ideas and forms of reasoning that one is working toward, (b) the conjectured starting points, and (c) the elements of a trajectory between the two (as outlined by Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003). Our endpoint was structured by a robust prior body of knowledge about the development of expertise in the domain; this served to focus our initial inquiry when key disagreements emerged during the 1st year. However, our delayed clarification of a conjectured starting point (highlighted when the original form of the assessment was made too difficult) cost substantial time

and effort. Likewise, we could have done a much better job of clarifying "elements of the trajectory" between the starting and ending points. Specifically, our efforts would have benefited from a clearer picture of the role of whole-class and collaborative-group discourse at the outset.

We conclude by expressing our enthusiasm for design-based studies of assessment practices for promising instructional innovations. Design-based methods seem ideal for refining the alignment of innovative curriculum, classroom assessments, and external assessments and for maximizing the impact of formative feedback at the various levels. As a caveat, our study supports Sloane and Gorard's (2003, p. 30) insistence that any such efforts seriously consider the possibility of "artifact failure" at every iteration and always consider that one's final model may be suboptimal. It seems to us that such studies could yield the consistently large gains on high-stakes assessments that have so far eluded many otherwise promising innovations. Such evidence seems essential for continued progress in instructional innovation, in light of current policy tensions (e.g., Feuer, Towne, & Shavelson, 2002; NRC, 2002; Pellegrino & Goldman, 2002).

## Notes

This article and the research that it describes were supported by an NSF award (RED-955348) to the third author and by a postdoctoral fellowship from the Educational Testing Service (ETS) Center for Performance Assessment and an NSF award (REC-0196225) to the first author. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF or ETS. Allan Collins, Janet Kolodner, Bill Sandoval, Chandra Orrill, and four anonymous reviewers provided essential feedback on this article or a previous version of it. We gratefully acknowledge the contributions of our numerous collaborators, particularly Drew Gitomer, Linda Steinberg, Joyce Schwartz, Iris Tabak, Edward Wolfe, Joan Heller, and graduate students Jessica DeCuir, Bryon Hand, Alex Heidenberg, Brandon Kyser, Marina Michael, Kirsten Mixter, Krista Herron, Annette Parrott, Art Russell, Nancy Schafer, and Steve Zuiker. We also thank the many administrators, teachers, and students who made this research possible. For more information, e-mail the authors (see author information on the second page of this article).

[1] Visit *http://genscope.concord.org/* for more information on the GenScope program, including software downloads, reports, the assessments, and curricula.

[2] For a more detailed description, see Horwitz & Christie (1999, 2000), or visit *http://genscope.concord.org/*, where you may also download a copy of the software.

[3] Visit *http://biologica.concord.org/index.html* for a more detailed description and downloads.

[4] In contrast to mammals, *female* dragons have an XY chromosome pair. Making female dragons XY was a design feature intended to focus learners' attention on the concept of sex-linked inheritance patterns rather than on the algorithms typically used to solve sex-linked inheritance problems.

[5] A number of 40-hour professional development workshops for teachers 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.

[6] *Genotype* refers to the genetic makeup for a particular characteristic (e.g., *TT* as opposed to *Tt*, as opposed to *tt*), whereas *phenotype* refers to the observable aspects of that characteristic (plants that are tall as opposed to short).

[7] 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-generations, domain-specific reasoning. Within-generation processes such as transcription and translation were not dealt with in the GenScope curriculum or the NewWorm assessment.





[8] Subsequently, Lobato (2003, p. 17) argued that this analysis exemplified an "observer-oriented" view of transfer and was therefore "not informed by data regarding the specific generalizations that the students may have formed and how the instructional environment may have afforded those connections." As an alternative, Lobato advances an "actor-oriented" view of transfer that "seeks to understand the process by which individuals generate their own similarities between problems." This suggests a fruitful direction for further refining our understanding of transfer across the various levels of instruction and assessment.

[9] The model also includes *distal* evidence of learning, based on established standards in particular content domains (i.e., standardized content tests), and *remote* evidence, based on general measures of achievement or student success. A roughly similar continuum is represented by Kennedy's (1999) four-level model.

[10] During Year 1 we also assessed posttest proficiency, in an additional four GenScope classrooms and seven comparison classroom using the more difficult NewFly assessment (reported in Hickey, Wolfe, & Kindfield, 2000). Because this instrument was ultimately abandoned and because there were no identical items with which to scale performance on the NewWorm assessment, those results are not reported here.

[11] The separation index for the items (a measure of the spread of the estimates relative to their precision) was 14.0. According to Fisher (1996), this means that we ended up with 19 statistically distinct strata of item difficulties, strata = [(4 * separation index) + 1]/3. A separation index of 4.1 for individuals indicated 5.6 statistically distinct strata of proficiency, $\chi^2(566) = 8{,}305$, $p < .005$. The latent-trait reliability coefficient $\alpha$ equivalent was .94

[12] 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, whereas the seniors had taken that course the previous year.

[13] For clarity, *class* refers to the students in a single classroom, *classroom* refers to all of the classes taught by a single teacher, and *course* refers to the curricular context (e.g., honors or college prep).

[14] Reflecting her own organization of life sciences around genetics, Ms. M reported using genetics to organize most of the biology curriculum. Therefore, the number of days spent teaching "genetics" is somewhat inflated.

## References

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences, 2,* 141–178.

Chemlinsky, E. (1998). The role of experience in formulating theories of evaluation practice. *American Journal of Evaluation, 19,* 35–55.

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

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13.

Cognition & Technology Group at Vanderbilt. (1997). *The Jasper project: Lessons in curriculum, instruction, assessment, and professional development.* Mahwah, NJ: Erlbaum.

Collins, A. (1992). Towards a design science of education. In E. Scanlon & T. O'Shea (Eds.), *New directions in educational technology.* New York: Springer.

Collins, A. (1999). The changing infrastructure of educational research. In E. C. Lagemann & L. B. Schulman (Eds.), *Issues in educational research: Problems and possibilities* (pp. 289–298). San Francisco: Jossey-Bass.

De Corte, E. (2000). Marrying theory building and improvement of school practice: A permanent challenge for instructional psychology. *Learning and Instruction 10,* 249–266.

Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5–8.

DiSessa, A. (1991). Local science: Viewing the design of human-computer systems as cognitive science. In J. M. Carroll (Ed.), *Designing interaction: Psychology at the human–computer interface* (pp. 162–202). New York: Cambridge University Press.

Duschl, R. (in press). *Assessment of inquiry*. Arlington, VA: National Science Teachers Association.

Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher, 31*(8), 4–14.

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.

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. Ronning, J. Glover, J. C. Conoley, & J. Witt (Eds.), *The influence of cognitive psychology on testing and measurement: The Buros-Nebraska Symposium on measurement and testing* (Vol. 3, pp. 41–85). Hillsdale, NJ: Erlbaum.

Greeno, J. G., Collins, A. M., & Resnick, L. (1996). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 15–46). New York: Macmillan.

Greeno, J. G., McDermott, R., Cole, K. A., Engle, R. A., Goldman, S., Knudsen, J., et al. (1998). Research, reform, and aims in education: Modes of action in search of each other. In E. C. Lagemann & L. S. Shulman (Eds.), *Issues in education research: Problems and possibilities* (pp. 299–335). San Francisco: Jossey-Bass.

Greeno, J. G., & Middle School Mathematics through Applied Projects Group. (1998). The situativity of knowing, learning, and research. *American Psychologist, 53,* 5–26.

Greeno, J. G., Smith, D. R., & Moore, J. L. (1993). Transfer of situated learning. In D. K. Detterman & R. J. Sternberg (Eds.), *Transfer on trial: Intelligence, cognition, and instruction* (pp. 99–167). Stamford, CT: Ablex.

Hickey, D. T. (2001). *Assessment, motivation, and epistemological reconciliation in a technology-supported learning environment* (Grant REC-0196225, National Science Foundation).

Hickey, D. T. (2003). Engaged participation vs. marginal non-participation: A stridently sociocultural model of achievement motivation. *Elementary School Journal, 103*(4), 401–429.

Hickey, D. T., Kindfield, A. C. H., Horwitz, P., & Christie, M. A. (1999). Advancing educational theory by enhancing practice in a technology-supported genetics learning environment. *Journal of Education, 181*(2), 1–33.

Hickey, D. T., Wolfe, E. W., & Kindfield, A. C. H. (2000). Assessing learning in a technology-supported genetics environment: Evidential and consequential validity issues. *Educational Assessment, 6*(3), 155–196.

Horwitz, P., & Christie, M. (1999). Hypermodels: Embedding curriculum and assessment in computer-based manipulatives. *Journal of Education, 181,* 1–24.

Horwitz, P., & Christie, M. (2000). 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.

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





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.

Kelly, A. E. (Ed.). (2003). Theme issue: The role of design in educational research. *Educational Researcher, 32*(1).

Kelly, A. E., & Lesh, R. A. (Eds.). (2000). *Handbook of research design in mathematics and science education*. Mahwah, NJ: Erlbaum.

Kennedy, M. M. (1999). Approximations to indicators of student outcomes. *Educational Evaluation and Policy Analysis, 21,* 345–363.

Kindfield, A. C. H. (1993/1994). Biology diagrams: Tools to think with. *Journal of the Learning Sciences, 3,* 1–36.

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., & Wolfe, E. W. (1999, April). *Tools for scaffolding inquiry in the domain of introductory genetics*. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Lagemann, E. (1999). An auspicious moment for education research? In E. Lagemann & L. S. Shulman (Eds.), *Issues in education research: Problems and possibilities* (pp. 3–16). San Francisco: Jossey-Bass.

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

Lobato, J. (2003). How design experiments can inform a rethinking of transfer and vice versa. *Educational Researcher, 32*(1), 17–20.

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

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and Human Behavior, 15,* 335–374.

National Center for Education Statistics. (1996). *NAEP 1996 Assessment: Science—Public Release, Grade 12*. Washington, DC: Author.

National Educational Research Policy and Priorities Board. (1999). *Investing in learning: A policy statement on research in education*. Washington, DC: Department of Education.

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

National Research Council. (1999a). In J. D. Bransford, A. L. Brown, & R. R. Cocking (Eds.), *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

National Research Council. (1999b). *Improving student learning: A strategic plan for education research and its utilization*. Washington, DC: National Academy Press.

National Research Council. (2001a). In J. M. Atkin, P. Black, & J. Coffey (Eds.), *Classroom assessment and the National Science Education Standards*. Washington, DC: National Academy Press.

National Research Council. (2001b). In J. W. Pellegrino, N. Chudowski, & R. W. Glaser. (Eds.), *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

National Research Council. (2002). In R. J. Shavelson & L. Towne (Eds.), *Scientific inquiry in education*. Committee on Scientific Principles for Educational Research. Washington, DC: National Academy Press.

National Research Council. (in press). *Assessment in support of instruction and learning: bridging the gap between large-scale and classroom assessment* (workshop report, Committee on Assessment in Support of Instruction and Learning, Board on Testing and Assessment). Washington, DC: National Academies Press.

Pellegrino, J. W., & Goldman, S. R. (2002). Be careful what you wish for—You may get it: Educational research in the spotlight. *Educational Researcher, 31*(8), 15–17.

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

Roschelle, J., & Jackiw, N. (2000). Technology design as educational research: Interweaving imagination, inquiry, and impact. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 777–798). Mahwah, NJ: Erlbaum.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39,* 369–393.

Sechrest, L. (1992). Roots: Back to our first generation. *Evaluation Practice, 13,* 1–7.

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

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.

Slack, S. J., & Stewart, J. (1990). High school students' problem-solving performance on realistic genetics problems. *Journal of Research in Science Teaching, 27,* 55–67.

Sloane, F. C., & Gorard, S. (2003). Exploring modeling aspects of design experiments. *Educational Researcher, 32*(1), 29–31.

Steffe, L. P. (1983). *The teaching experiment methodology in a constructivist research program*. Paper presented at the Fourth International Congress on Mathematics Education, Boston.

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.

Stewart, J., Hafner, R., Johnson, S., & Finkel, L. (1992). Using computers to facilitate learning science and learning about science. *Educational Psychologist, 27,* 317–336.

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

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

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181–208.

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.

Manuscript received April 30, 2002
Revision received February 14, 2003
Accepted March 5, 2003