Hickey, D.T., Kindfield, A.C.H., Horwitz, P., & Christie, M. (2000). Integrating Instruction, Assessment, & Evaluation in a Technology-Based Genetics Environment: The GenScope (TM) Follow-up Study. In B. Fishman & S. O'Connor-Divelbiss (Eds.), *Fourth International Conference of the Learning Sciences* (pp. 6-13). Mahwah, NJ: Erlbaum.

Integrating Instruction, Assessment, & Evaluation in a Technology-Based Genetics Environment: The *GenScope*™ Follow-up Study

Daniel T. Hickey
Department of Educational Psychology & Special Education, Georgia State University, Atlanta, GA 30303
Tel: (404) 651-0127, Fax: (404) 641-4901
Email: dhickey@gsu.edu

Ann C. H. Kindfield
Educational Designs Unlimited, 325 Zion Road, Neshanic Station, NJ 08853
Tel: (908) 369-8960, Fax: (801) 749-0952
Email: annk@net-lynx.com

Paul Horwitz & Mary Ann Christie, The Concord Consortium, 37 Thoreau Street, Concord, MA 01742
Tel: (978) 371-5856 (PH), (978) 371-5851 (MAC), Fax: (978) 371-0696
Email: paul@concord.org, maryann@concord.org

Abstract: A previously reported study in 24 secondary science classrooms showed the *GenScope* computer-supported learning environment to be at least as effective as conventional curricula at enhancing genetics reasoning ability. A follow-up study in three more classrooms yielded the dramatic reasoning gains we had been seeking, partly by addressing four unresolved issues from the prior research. First, addressing previous difficulties implementing GenScope in shared computer labs, the follow-up study used laptop computers installed in the biology classroom. Second, addressing problems establishing valid comparison classrooms, we established a more valid comparison classroom that did not encounter "carry over" from the GenScope classrooms. The third issue concerned continuing refinements made to the GenScope curriculum. The final issue concerned one aspect of that curriculum, formative assessments that used the familiar GenScope dragons to scaffold reasoning targeted in our summative assessments. By withholding these activities from one of two GenScope classrooms, the present study confirmed that this enhancement to "systemic" validity (Frederiksen & Collins, 1989) presented a small, acceptable degree of compromise to "evidential" validity. The specific results and the broader collaboration are considered in light of recent federal policy reports regarding educational technology, educational research, and assessment practices.

Keywords: assessment, learning environments, scaffolding, science education

In 1998, we reported the findings from a three-year collaboration to implement and evaluate the *GenScope*™ software and associated curriculum in 20 secondary science classrooms and 4 comparison classrooms (Hickey, Kindfield, Wolfe, & Heidenberg, 1998). This paper provides a more comprehensive summary and interpretation of that research and presents a follow-up study that addressed four issues left unresolved in the prior research. While GenScope was shown to be as effective or more effective than conventional curricula for enhancing genetics reasoning skills, these issues may have diminished the relative gains for the GenScope classrooms and clouded interpretation of the results. The first issue concerned the challenges students faced when independently completing the GenScope activities in the school computer lab (a problem that was exacerbated by the demise of Macintosh computers in secondary schools). A second associated issue concerned the difficulty of identifying "fair" comparison classrooms in within-teacher contrasts because of "carryover" from the GenScope curriculum into comparison classrooms. The third issue concerned the need for further refinement and organization of the GenScope curriculum. The fourth issue concerned the impact of one key aspect of that curriculum, a set of formative assessments known as *Dragon Investigations* designed to use the familiar GenScope dragons to scaffold the kind of reasoning assessed in our *NewWorm* assessment. While the Dragon Investigations were shown to increase performance substantially, we had yet to show they were supporting genuine domain reasoning gains—rather than compromising the evidential validity of our summative assessment.

Background

Genetics is a particularly challenging topic for science teachers and their students. It involves relationships between events that occur at different levels of biological organization and involves probabilistic phenomena that are not directly observable because they take place too quickly or slowly, or on a scale that is too small or too large. As such, mastery of the genetics content and reasoning goals as defined in current science education standards (e.g., National Research Council, 1996) can be daunting. To help meet this challenge, science education researchers have invested heavily in computer-based tools for teaching genetics (Jungck & Calley, 1985; Stewart, Hafner, Johnson, & Finkel, 1992). Starting in 1991, a team at BBN Labs (now at the Concord Consortium) began developing and refining the *GenScope*™ software, developing curricular activities, and piloting those activities (Horwitz & Christie, in press; Horwitz, Neumann, & Schwartz, 1996).[1] GenScope is acknowledged as a noteworthy example of the synergy between educational technology and contemporary constructivist pedagogical principles (see Bransford, Brown, & Cocking, 1999, p. 204, and is consistent with policy recommendations for K-12 educational technology issued by the President's Committee of Advisors on Science and Technology (PCAST, 1997).

The GenScope Software.

Within the GenScope software, the various levels of biological organization relevant to introductory genetics are represented by different windows. Each window dynamically represents the relevant information alongside easy-to-use tools for manipulating that information. While a number of species are included, most of the activities involve the fanciful dragons that features just three chromosomes and nine traits, but exemplifies most of the relationships covered in introductory genetics (sex linkage, incomplete dominance, polygenic traits, lethal alleles, etc.). The *organism* window (the logical starting place for student inquiry) displays the organisms' phenotype (the collection of their physical traits), but gives no direct information concerning their genetic makeup. Clicking a button in the organism window opens the *chromosome* window, with pull-down menus for changing the gene from one variant, or "allele," to another. A button in the chromosome window takes the learner to the *DNA* level for each allele, revealing its molecular sequence either as paired strings of colored rectangles representing the DNA base pairs or as a linear sequence of paired letters with the abbreviations for adenine, thymine, guanine, and cytosine. Mutations created at the DNA level (by adding or deleting letters in sequences) are treated as new alleles.

By dragging individuals from the organism window to the *cell* window, students can observe and explore meiosis, mitosis, and fertilization. They can directly control the relative alignment of chromosomes during meiosis as well as the crossover of DNA between them, or allow them to behave randomly, as in nature. After creating gametes (e.g., eggs and sperm), students can again observe fertilization and then see the outcome of these processes in new offspring. At the *pedigree* level, students create "family tree" structures of related organisms in order to observe and investigate inheritance patterns. The *population* level introduces time and space. Organisms move about on the screen, randomly mating with each other, and different portions of the screen can be assigned different "environments" that selectively favor one or another phenotype, which allows for the observation and quantification of the impact of various forces of evolution.

The GenScope Curriculum

The GenScope development team had designed numerous activities when the present research was initiated and continued refining them and developing new ones. Most of these were 1-3 page "puzzle" exercises that typical students could complete within a single class period. As part of on-going enhancements, and partly in response to disappointing initial learning outcomes, the GenScope development team made substantial revisions and enhancements to the software and continued developing and refining curricular activities. In keeping with newer perspectives on assessment and instruction (e.g., Frederiksen & Collins, 1989; Wolfe, Bixby, Glenn, & Gardner, 1993) the assessment team developed a set of paper-and-pencil formative assessments known as *Dragon Investigations*. These materials were designed to foster a focused whole-class discussion by building on the teacher's and students' shared, simplified understanding of the domain as represented by the GenScope dragons. Individual Dragon Investigations were designed to be useful away from the computer, either as homework or in class, and each was accompanied by a teacher's answer key that included detailed explanations of the relevant domain content in the context of solving particular problem. The eleven Dragon Investigations and a subset of GenScope computer activities were then organized into six curricular units around major domain reasoning concepts, including *Introduction*, *Basic Inheritance*, *DNA & Meiotic Events and Inheritance*, *Two-gene Inheritance*, *Alignment and Crossover*, and *Reasoning about Inheritance*. A package including a teacher guide and a packet of

student worksheets was reproduced and distributed to implementation teachers partway through the final implementation year when most of the data were collected.

Prior Research Methods and Findings

The *NewWorm* assessment was developed and used to assess students' ability to reason about genetics. It uses a species whose genetics mimics that of GenScope dragons, but is novel and understandable to both GenScope and non-GenScope students (Kindfield, Hickey, & Wolfe, 1999). The items were carefully sequenced to scaffold student performance across increasingly complex problems. The initial NewWorm problems were designed to be solvable by most secondary students prior to any instruction, and introduced students to the new organism, genome, and assessment environment. Success on the initial problems was expected to yield motivation and understanding that would scaffold performance on the more difficult subsequent problems. Some of the items called for categorical, single-word responses (or selection from multiple verbal or diagrammatic choices), while the items assessing more complex reasoning also asked students to explain why the categorical response was correct. Following from a developmental model of genetics reasoning (Kindfield, 1994; Stewart & Hafner, 1994) and as shown in Table 1, the various NewWorm items can be classified along two primary dimensions: (1) Domain-general Reasoning Type (cause-to-effect, effect-to-cause, and process reasoning) and (2) Domain-specific Reasoning Type (within-generations and between-generations). The items can further be distinguished according to the particular genetics involved (e.g., autosomal vs. X-linked inheritance), the explicitness of provided information, and/or type of information used/sought (i.e., categorical, probabilistic, diagrammatic, short answer, definitive vs. indeterminate).

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

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

Method

During our three-year collaboration, the NewWorm was used to assess reasoning gains in 20 GenScope classrooms and 4 comparison classrooms. Student scores were analyzed using multi-faceted Rasch scaling (Linacre, 1989) to locate each assessment item and each individual's score on a single linear scale. This provided an estimate of the relative difficulty of each item and relative proficiency of each student on a common metric (*logits*). Figure 1 shows that relative difficulty of the NewWorm items closely matches our assumptions about those dimensions (Table 1). The information in Figure 1 is useful for characterizing differences and gains in domain reasoning ability. For example, the difference between the algorithmic cause-to-effect reasoning and the more expert effect-to-cause reasoning is roughly 2 units of the 6-unit range of the scale. Figure 2 shows what we found when we administered the NewWorm to pairs of college students and faculty. The more expert pairs performed increasingly well, and the faculty members reveal the upper bounds of expertise on the assessment—process reasoning.

Results

Figure 2 shows the mean proficiency of students in four groups of GenScope and comparison classroom before and after genetics instruction. Given the diversity of the GenScope implementations in different types of classrooms, the relative gains in GenScope and comparison classrooms were considered within each of the four

classroom types. As described in more detail in Hickey, Kindfield, and Wolfe (1999), differences from year to year and from school to school further called for many different analyses and statistical tests. For example, the 9th-grade

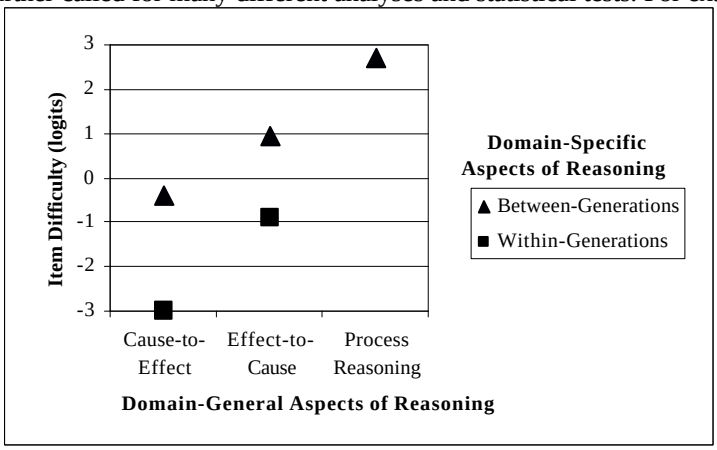

Figure 1. Relative difficulty of various types of NewWorm items according to the underlying model of domain reasoning. (Scale is in logits, a logistic odds consisting of the probability of getting a dichotomous item correct divided by the probability of getting an item incorrect, or the probability of getting n points on a partial credit item divided by the probability of getting n-1 points.)

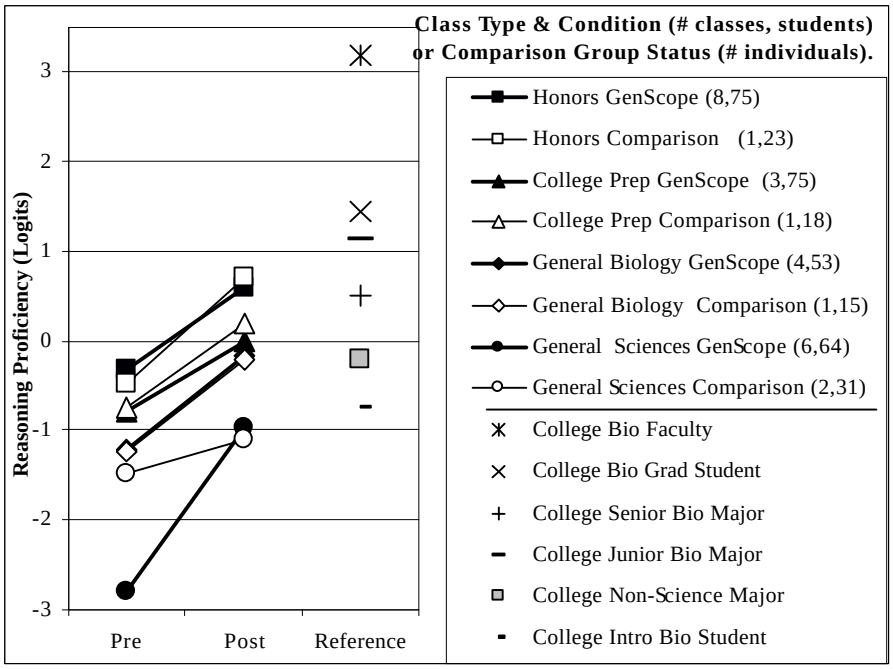

Figure 2. Genetics reasoning proficiency before and after instruction in GenScope and comparison classrooms and in six pairs of college biology students and faculty.

general science classrooms were at a struggling inner-city school that served extremely disadvantaged students. Three of the six classrooms for the two GenScope teachers did not have access to the computer labs; while their counterparts completed GenScope activities in the computer labs, their teachers went over the various activities with the class on the chalkboard. The gains in all six GenScope classrooms were similar and large, and were significantly

larger than the gains in the two comparison classrooms at a similarly disadvantaged school where difficulties accessing the computer led the teacher to abandon the GenScope program and curriculum entirely, (but continue to participate in the evaluation). While the gains in the GenScope classrooms were quite large-as much as two units of the six-unit scale, the extremely low pretest scores meant that at posttest proficiency was still below the level of between-generations problems—well below college prep and honors students *before instruction*.

In four general biology classrooms at one suburban school, gains in the three GenScope classrooms were similar to the gains in the one comparison classroom, where the teacher relied on a mix of lectures and a programmed instruction module to cover genetics. In five general biology classrooms at suburban schools, two teachers implemented GenScope in four classrooms. The gains were (barely) statistically larger than in the comparison classroom, where the teacher also relied on lectures and programmed instruction. In the nine honors biology classrooms at suburban schools, one of three teachers agreed to withhold GenScope from one of his classrooms; however, this teacher, per our instructions, did his best to help those students do as well on the test as his two GenScope classrooms, who independently struggled to complete the activities in the computer lab. To the teacher's surprise, the gains in the GenScope classrooms were larger, but only slightly and non-significantly so.

Conclusions from Primary Study

We concluded that GenScope was certainly as effective as the existing practices it supplanted or replaced, and in some cases, more effective. However, with the exceptions of the 9th grade general sciences classrooms, the gains in the GenScope classrooms were still quite modest, and few of the students developed the domain reasoning skills needed to solve between-generations effect-to-cause problems. Further, the relative gains in the GenScope classes were apparently diminished by (1) the challenges of accessing computer labs (2) carryover from GenScope classrooms to comparison classrooms, and (3) various difficulties with the curriculum. In addition, while comments and data showed that the Dragon Investigation formative assessments did enhance reasoning gains, we could not prove that our efforts to enhance the consequential validity of our assessments had not compromised evidential validity of scores on the NewWorm assessment. While we expected that the Dragon Investigations would give GenScope students a small advantage over the comparison students by familiarizing them with the format of the items on the NewWorm, we could not prove that completing the formative assessments had not fundamentally compromised the test by teaching students simple algorithms that would allow them to solve the assessment problems without using domain reasoning skills.

Follow-up Study

The follow-up study was conducted in three classrooms at a suburban/rural school that served relatively advantaged students. Three classrooms served a single pool of technical track (i.e., non university-bound) students and roughly half of the students in each classroom were identified as having learning or behavioral disabilities. The GenScope teacher was a first-year teacher, and had participated in the GenScope research (primarily scoring assessments and evaluating curricular activities) during the previous year while she was a science education graduate student. From the outset, this implementation was structured to address the four issues that were still unresolved from the prior research.

Method

In order to address the first issue concerning problems with computer access, these students completed the GenScope activities on 10 laptop computers installed in the biology classroom/lab for the duration of the implementation. Regarding the second issue, the carryover effects of the GenScope curriculum and the associated lack of valid implementation/comparison pairs, a very experienced biology teacher was recruited to provide an "ideal" comparison classroom. This teacher taught general biology to the same population of students, and was provided with a detailed summary of the reasoning concepts assessed in the GenScope curriculum and the NewWorm assessment. She was encouraged to do her best using the methods that she normally used (lecture/worksheets/textbook/discussion) to help her students develop the targeted domain reasoning skills during roughly the same number of class periods as the GenScope classroom.

In order to address the third issue, the curriculum was further enhanced. The activities were further refined and organized into 6 units of instruction to cover the 25 periods to be allocated to genetics. Regarding the fourth issue, the impact of the Dragon Investigations, the first GenScope class completed 15 GenScope computer activities—but no Dragon Investigations. The second class completed only 10 GenScope computer activities, but completed 6 Dragon Investigations as in-class activities in lieu of the computer activities. Thus, one group of

students had roughly one third of their computer-based activities replaced by paper-and-pencil activities designed to teach very specific aspects of domain reasoning.

Daily videotapes made by a high-school student assistant in the GenScope classrooms revealed no technical difficulties with the computer activities or software in either classroom. As we had hoped, while the students were completing the activities, the teacher wandered among the students to answer question, occasionally calling for everyone's attention during or following the activity to review or clarify a particular point. Reflecting the number of behaviorally disabled students and the overall modest proficiency of these students, the videotapes did reveal plenty of "horsing around" during the computer activities, stretches of off-task activity, and repeated, effective efforts by the teacher and an aide to maintain order. The videotapes during the first (non-Dragon Investigation) class showed that the teacher initiated many whole-class discussions that were fairly similar to the discussions that were instigated by the Dragon Investigations in the other classroom. In other words, there appeared to be some carryover of the broader goals of the Dragon Investigations into the other classroom, but those students were never asked to complete problems that were like those on the NewWorm assessment.

Results

Figure 3 shows the reasoning gains in the three classrooms in the followup study. The gain in the comparison classroom (triangles) was a modest 0.83 logits; in contrast the gain in the GenScope classroom that did not use the Dragon Investigations (squares) was an impressive 2.14 logits. Most impressively, the gain in the GenScope classroom that used the Dragon Investigations (circles) was 2.67 logits, the largest of any classroom in the study. The gains in both of the GenScope classrooms were significantly larger than the gain in the comparison [$F(1,37) = 9.10$, $p = .005$, and $F(1,38) = 14.02$, $p < .001$, respectively]. The differences in the gains in the two GenScope classrooms was not statistically significant $F(1,37) = 2.24$, $p = .143$. Fortunately, this implementation provided about as valid a comparison group as can be established in classroom-based research. Given the validity of the comparison pairing and close observations of the implementation, these results provide conclusive evidence that the GenScope learning environment is substantially more effective than the typical conventional learning environment that it would replace—at least in terms of the domain reasoning skills assessed with the NewWorm.

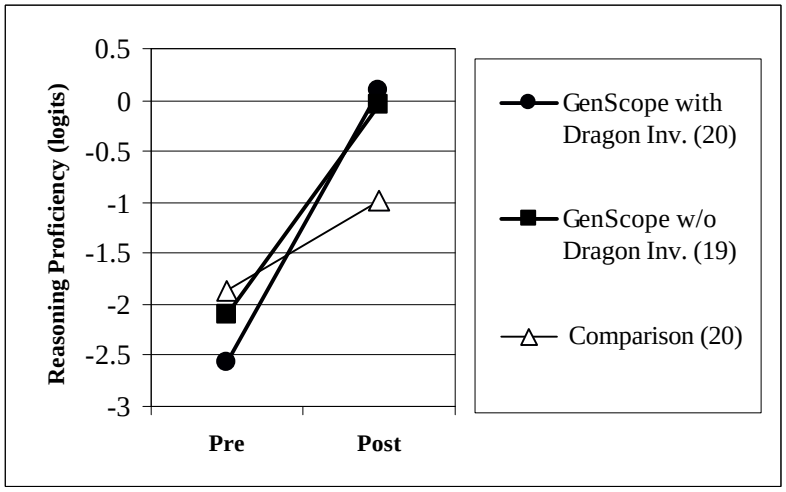

Figure 3. Reasoning gains in follow-up implementation.

Conclusions

The first overall conclusion concerns GenScope's power for enhancing students' ability to reason about introductory genetics. In light of the dimensions of domain reasoning described earlier, the gains in the follow-up GenScope classrooms represent roughly the difference between within-generation and between-generation reasoning, or the difference between cause-to-effect and effect-to-cause reasoning—a fundamental, qualitative change in students' ability to reason in the domain of introductory genetics. These are the sort of gains whose absence in typical biology classrooms has long been lamented by genetics education researchers (see Stewart &

Hafner, 1994). These gains can be attributed to a learning environment that could not be accomplished without technology, providing evidence that educational technology proponents seek but seldom find.

The second conclusion is that providing computer access *in the classroom* enhances the teacher's ability to use the technology-supported curriculum to support meaningful learning. This provides additional support to the many arguments for placing of computers in content area classrooms rather than computer labs, and lends additional credence to the conclusion of the PCAST (1997) report that priority should be given to using computers to teach subject area content rather than teaching about computers themselves.

The third conclusion is that our Dragon Investigations formative assessments presented a small and acceptable degree of compromise to the NewWorm's evidential validity. There would have been a much larger difference in the two GenScope followup classrooms if the Dragon Investigations had more fundamentally compromised performance on the NewWorm. More generally, these results argue that sacrificing the pedagogical power of assessment in order to preserve evidential validity is inappropriate and unethical. Indeed, these results show that it is possible to sacrifice a small, knowable degree of evidential validity in exchange for dramatic increases in the positive consequences of the assessment practice. These results support theorists like Frederiksen and Collins (1989) who argue that assessment events are too valuable for supporting learning to preserving every bit of evidential validity.

The results of the follow-up study provide one example of how educators can take advantage of the powerful affordances of assessment practice for structuring curricula and providing instruction while still providing the degree of evidential validity called for in current policy documents such as the PCAST (1997) report. These results provide both a justification and a framework to further develop linked instructional and assessment activities within the new *BioLogica* software. When paired with the validity inquiry described in Hickey, Wolfe, & Kindfield (2000) and used in the type of implementation studies described here, it should be possible to further develop "systemically valid" assessment practices that simultaneously maximize learning and preserve evidential validity.

A final conclusion concerns the research collaboration behind the findings described here. As described in more detail in Hickey, Kindfield, Horwitz, and Christie (in press), we conclude that our collaboration is one of the few recent examples of the systematic, sustained collaboration between educators, developers, and researchers called for by the National Educational Research Policy and Priorities Board (1999) and the National Research Council (1999). We believe that aspects of our research provide a modest initial illustration of the kind of "focused, multidisciplinary, cumulative, sustained, solutions-oriented" research programs outlined by the NRC. The collaborative relationship between the development team and the assessment team was relatively unique, and allowed the assessment team to reorganize the curriculum and include curricular activities that ostensibly presented a threat to evidential validity. This was possible because the National Science Foundation elected to fund successive initiatives targeting a long-standing educational problem. This work is continuing within the *BioLogica* development effort that uses newer development tools to provide a scriptable, platform-independent package.

The level of support for this work by the National Science Foundation was sufficient to have instigated a community of inquiry and practice—educators, research, and developers—around this tool with a shared goal of enhancing 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 collaboration in the context of practice that is not externally supported. Our experience leads us to share the apparent enthusiasm of educational research policy makers for this sort of effort.

Endnotes

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

References

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

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

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

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

Hickey, D. T., Kindfield, A. C. H., and Wolfe, E. W. (1999, April). *Assessment-oriented scaffolding of student and teacher performance in a technology-supported genetics environment.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada (in review, *The Journal of the Learning Sciences*).

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

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

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

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, Quebec, Canada (in review, *Science Education*).

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

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

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

National Research Council: Committee on a Feasibility Study for a Strategic Education Research Program. (1999). *Improving student learning: A strategic plan for education research and its utilization*. Washington, DC: National Academy Press.

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

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.

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.

Acknowledgments

This research was supported by NSF Applications of Advanced Technology Program Grant RED-95-5348 to the third author, and by a postdoctoral fellowship from the Center for Performance Assessment at Educational Testing Service to the first author. The authors acknowledge the substantial contributions of our collaborators Edward Wolfe and Joyce Schwartz, and graduate students Alex Heidenberg, Brian Davis, Kirsten Mixter, and Krista Herron. We also thank the many administrators, teachers, and students who made this research possible.