DOCUMENT RESUME

ED 478 483 TM 035 082

AUTHOR Hickey, Daniel T.; Kruger, Ann Cale; Frederick, Laura D.; Schafer, Nancy Jo; Zuiker, Steven

TITLE Design Experimentation with Multiple Perspectives: The GenScope Assessment Project.

SPONS AGENCY National Science Foundation, Arlington, VA.

PUB DATE 2003-04-00

NOTE 37p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).

CONTRACT REC-0196225

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01/PC02 Plus Postage.

DESCRIPTORS *Computer Assisted Instruction; *Formative Evaluation; Genetics; *High School Students; High Schools; *Performance Based Assessment; *Research Methodology

ABSTRACT

The GenScope Assessment Project is studying assessment in the context of a month-long computer-supported learning environment for introductory genetics. Across three annual iterations with multiple teachers, project researchers manipulated the materials, incentives, and contexts in which students were invited to use formative feedback on challenging classroom performance assessments. The consequences of these manipulations on engagement and learning were systematically examined from behavioral/empiricist, cognitive/rationalist, and situative/sociohistoric perspectives. This "comparative approach" was intended to provide new insights into unresolved issues over extrinsic rewards and accountability-oriented reforms.It turned out that the comparative approach also provided a powerful framework for refining and improving theories about classroom assessment. Essentially researchers "tuned" the classroom assessment environment to maximize gains on carefully aligned external performance assessments. Successive improvements led to correspondingly larger gains on an external achievement test that was more aligned with conventional genetics instruction. This study shows that design-based research around classroom assessment can help meet the wider educational goals of researchers within the increasingly narrow policies of reformers. Five appendixes contain examples of feedback and rubrics. (Contains 10 figures and 80 references.) (SLD)

Design Experimentation with Multiple Perspectives: *The GenScope Assessment Project*

Daniel T. Hickey*
Learning & Performance Support Laboratory
Department of Educational Psychology
University of Georgia

Ann Cale Kruger
Laura D. Fredrick
Nancy Jo Schafer
Department of Educational Psychology and Special Education
Georgia State University

Steven Zuiker
Learning & Performance Support Laboratory
Department of Educational Psychology
University of Georgia

Abstract

We are studying assessment in the context of a month-long computer-supported learning environment for introductory genetics. We are trying to increase the value of classroom assessment practices for directly advancing student knowledge, and indirectly enhancing learning environments. Across three annual iterations with multiple teachers, we manipulated the materials, incentives, and context in which students were invited to use formative feedback on challenging classroom performance assessments. We systematically examined the consequences of these manipulations on engagement and learning from behavioral/empiricist, cognitive/rationalist, and situative/sociohistoric perspectives. This "comparative" approach was intended to provide new insights into unresolved issues over extrinsic rewards and accountability-oriented reforms. It turned out that the comparative approach also provided a powerful framework for refining and improving our theories about classroom assessment. Essentially, we "tuned" the classroom assessment environment to maximize gains on carefully aligned external performance assessments. Our successive improvements led to correspondingly larger gains on an external achievement test that was more aligned with conventional genetics instruction. This study shows that design-based research around classroom assessment can help meet the wider educational goals of researchers within the increasingly narrow policies of reformers

* Presented at the annual meeting of the American Educational Research Association, Chicago, April, 2003. This research is supported by the National Science Foundation Grant REC-0196225 to the University of Georgia. The opinions presented here belong to the authors and do not necessarily represent the positions of the University of Georgia or the National Science Foundation. We gratefully acknowledge the contributions of many graduate assistants and educators on this project, including Novella Abbott, Bobby Bable, Bryon Hand, Marina Michael, Marcus Norman, Annette Parrott, John Price, and Art Russell. We also gratefully acknowledge the district administrators who helped coordinate our efforts and the student who participated in it. Daniel Hickey, Room 611 Aderhold Hall, Athens, GA, 30602, dhickey@coe.uga.edu.

Our project represents the convergence of several lines of inquiry. This paper illustrates how newer design-based research methods can be used to enhance classroom assessment practices around innovative curriculum. This paper further illustrates how such an effort is enhanced when it systematically differentiates between different, competing views of the knowing, learning, and instruction. Following is a summary of each of these lines of inquiry and how they are manifested in our project.

Curricular Inquiry

At the most basic level, our project concerns using multimedia computers to teach introductory genetics. Key genetics phenomena are not directly observable, and secondary genetics is often students' first formal exposure to probabilistic reasoning. The factors that make genetics difficult to teach and learn also made it a promising candidate to profit from classroom multimedia technology. Starting in 1991, with the support of the National Science Foundation, a team at BBN Labs (headed by Paul Horwitz, now at the Concord Consortium) began developing and refining software for teaching introductory genetics in middle and secondary science classrooms, (Horwitz, Neumann, & Schwartz, 1996; Horwitz & Christie, 2000). The resulting *GenScope* software has been widely acknowledged as a noteworthy example of the synergy between advances in educational technology and contemporary constructivist pedagogical principles (e.g., Bransford, Brown, & Cocking, 1999, Chapter 9).

The learning environment afforded by *GenScope* is generally consistent with the software recommendations for K-12 educational technology issued by the President's Committee of Advisors on Science and Technology (PCAST, 1997). During a four-year collaboration funded by NSF's AAT Program (Grant RED-95-5348) GenScope's developers and a team including Dan Hickey and Ann Kindfield (initially at Educational Testing Service) implemented and evaluated GenScope in over 40 classrooms. This research found GenScope to be an effective tool for enhancing or supplanting conventional introductory genetics instruction. (Hickey, Kindfield, Horwitz, & Christie, 1999; in press). The effort also yielded formative and summative assessment tools that are central to the present project. The former were shown to be very effective at improving learning in the GenScope environment; the latter was shown to be an effective tool for measuring learning in GenScope and conventional introductory genetics environments.

Assessment Inquiry

Our efforts draw directly from the emerging consensus around the value of formative assessment (e.g., Black & Wiliam, 1998, Gipps, 1999; Graue, 1993, National Research Council, 2002a Turnstall & Gipps, 1996). In key respects, we have followed Duschl and Gitomer's (1997) inspiring portfolio-oriented research on *assessment conversations* into the current accountability-oriented reform climate. We are attempting to define modest, scaleable practices that engage students in worthwhile discourse around formative assessment feedback. Such activity promises to dramatically enhance student learning. Our idealized characterization of such activity is involves vibrant authentic scientific argumentation (e.g., Driver, Newton, & Osborne, 2000) where students are making and warranting knowledge claims based on evidence and on theory of the specific scientific domain (e.g., Jimenez-Aleixandre, Rodriguez, & Duschl, 2000).

Our efforts are guided by newer views of educational assessment and of motivation that follow from emerging situative/sociocultural perspectives of knowing and learning (e.g., Vygotsky, 1978; Lave, 1988). In a seminal article, Frederiksen & Collins (1989) advanced the notion of *systemic validity,* as a fundamental reconceptualization of testing:

> A systemically valid test is one that induces in the educational system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure (Fredriksen & Collins, 1989, p. 27).

The notion of systemic validity challenged conventional assumptions about assessment because it blurred the distinctions between evidential and consequential validity, as well as between formative and summative assessment (Hickey, Wolfe, & Kindfield, 2000). Fredriksen and Collins proceeded to outline a set of principles for the design of systemically valid assessment systems, including the *components* of the system (a representative set of tasks, a definition of the primary traits for each subprocess, a library of exemplars, and a training system for scoring tests), *standards* for judging the assessments (directness, scope, reliability, and transparency) and *methods for fostering self-improvement* (practice in self-assessment, repeated testing, performance feedback, and multiple levels of success). One goal therefore is developing a better understanding of systemically valid assessment practices.

The research that guided our efforts has recently been catalogued in two National Research Council reports: *Classroom Assessment and the National Science Education Standards* (NRC, 2001a), and *Knowing What Students Know: The Science and Design of Educational Assessment* (2001b). For example, the first report concludes that classroom assessments can powerfully enhance learning and teaching—provided they are accompanied by feedback that learners use to advance their understanding, and that teachers use to evaluate and refine their instructional practices. Both 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, 2002a) is focusing directly on this issue.

Theoretical Inquiry

A third line of inquiry concerns the complex tensions between the contemporary situative/sociocultural views of knowing and learning and prior behavioral/empiricist views and cognitive/rationalist views. These goals of our project are perhaps best understood as an effort to use arguments set forth by Greeno (e.g., Greeno et al, 1998; Greeno, Collins, & Resnick, 1996) to address the tensions between empiricist models of testing and reform, and modern rationalist models of motivation. To this end, our project is attempting to study learning and engagement from these three very different perspectives. This required us to first define ways of conceptualizing and assessing engagement and learning from each perspective. This has itself been a worthwhile outcome of our project. This supports a subsequent goal of revealing the contradictions that emerge when examining engagement and learning in this manner. This in turn supports our ultimate goals of comparing different ways of reconciling the tensions created by contradictory conclusions. To the extent that these tensions are undermining efforts to reform teaching and testing, we believe that such an inquiry has far-reaching ramifications.

As outlined in more detail in Hickey and McCaslin (2001) and Hickey (2003) researcher have traditionally reconciled the activity of individual and the activity of groups by characterizing social contexts using aggregated individual-level constructs. While following quite naturally from behavioral and cognitive perspectives, this "levels of aggregation" approach seems problematic. The present study attempts to explore the value of a "dialectical" reconciliation. Following quite directly from Greeno et al. (1998), this approach characterizes both the patterns of behavior of individual organisms and the patterns of human information processing as special cases of a broader form of situated human activity. This approach is controversial because it advances situative/sociocultural approaches as higher-order synthesis that balances the strengths and weakness of the prior approaches in the process of subsuming them. Our initial goal was exploring whether a dialectical approach could advance the seemingly intractable debate over competition and extrinsic rewards. The extended analyses and interpretation needed to explore this issue are still underway. This paper focuses more on the apparent value of such

an approach for refining classroom assessment practices to maximize the impact of innovative learning environments on high-stakes external achievement tests. Specifically, we use the more readily analyzed learning outcome measures to show how this perspective can help enhance and demonstrate the power of otherwise-promising curricular innovations.

As a caveat, we acknowledge the potential objections our deliberately comparative approach. More judicious and detailed applications of competing perspectives (e.g., Tudge and Winteroff, 1993) have been critiqued for overemphasizing initial conceptualizations of competing theories and ignoring their evolution (i.e., Zimmerman, 1993). From our perspective, such concerns are most relevant when one's primary concern is developing coherent and parsimonious theories of cognition and information processing. As described next, our primary concern is directly improving teaching and learning. Hence, we insist that the value of a comparative approach and a dialectical reconciliation should be judged in terms of their direct impact on educational practice.

Methodological Inquiry

Our project raises issues about research methods in education. Our approach reflects a 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, & 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. Following the inspiration of Stoakes' influential book *Pasteur's Quadrant,* we are developing "use-inspired basic research" about classroom assessment practices.

These fundamental shifts in educational research are embodied in "design-based" approaches, 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; also 1999) and Brown (1992), and are exemplified in the widely cited efforts of the Cognition and Technology Group at Vanderbilt (e.g., 1997) and Greeno et al. (e.g., 1999). Recent collaborative efforts (i.e. Kelly & Lesh, 2000; Kelly 2003; Design-Based Research Collective, 2003) 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" (DBRC, 2003, p. 10).

METHODS

This paper reports the results of three annual implementations, conducted in 2001, 2002, and 2003. With each implementation, we systematically manipulated the incentive context in which student completed classroom performance assessments (i.e., graded vs. ungraded) and used formative feedback about their performance (with and without public recognition of self-assessed proficiency after reviewing the answer explanations). These manipulations were studied within a relatively powerful in

within-teacher/between-class design using multiple teachers. Across each of three implementation cycles, we refined these manipulations in a search for significant consistent effects on learning outcomes.

These manipulations were ultimately unsuccessful, in terms of the learning outcomes being reported here. Specifically, we found that our manipulations explained little of the variance in pre-post scores on two learning outcome measures. When statistically significant differences were observed, they were not consistent across teachers. As such, we are omitting these details from the present consideration. Rather, we focus on the overall refinements to our classroom assessment practice that were made from one year to the next, in all classrooms, and their impact on all of the students taught by each teacher.

Participants

Across the three years, six teachers implemented the GenScope curriculum in a total of 30 9th grade life science classes. These classes included both honors life sciences course as well as AC (adjusted-curriculum)[1] life science. Five of these teachers were recruited via a request sent out by the district science coordinator, and were paid a non-trivial honorarium for their participation; these teachers in turn helped us recruit two additional teachers at their schools who agree to allow us to use four of their classroom for comparison purposes. The following details are relevant to some analyses, but not others; the most important point is that this paper focuses on the two GenScope teachers who participated during all three project years.

GenScope Teachers.

School 1 was lower SES suburban school where over 30% of the students had qualified for the federal lunch subsidy. Nearly every student in the school (99.5%) was African American. The school typically posted school wide achievement scores that were below average overall but higher than most of the other schools in this district that also served predominantly African American students. Published figures reported that 61% of these students passed the science component of the high school graduation test on their first try. One teacher at this school, Mr. N, played a central role in our efforts. He implemented GenScope in four of his classes during each of the four project years. He was African American, and had an undergraduate degree in biology and five years experience when the project started. Mr. P, a Euro American, implemented GenScope in four of his classes during Year 1. He left teaching for a position in the private sector the subsequent year.

School 2 was a middle-SES suburban school where 18% of the students qualified for a lunch subsidy. The school typically posted standardized achievement scores somewhat above average, with 89% passing the science graduation test on their first try. Roughly 40% of the students at this school were African American, and some of those students were continuing as participants in a court-ordered desegregation plan that had been abandoned several years earlier. Ms. P, an African American, was a doctoral student in science education who also participated in the curriculum development effort as a graduate research assistant. She implemented GenScope in two of her AC life science classes during each of the three project years.

School 3 was a high-SES suburban school that reported school-wide achievement scores well above the overall average, with a 95% pass rate on the science graduation test. Only 1.5% of the students at this school were qualified for the lunch subsidy and most (88%) were non-minority. Ms. L,

[1] AC refers to "adjusted-curriculum" meaning that the district approved college preparatory curriculum could be modified to meet the needs of special education students. While the curriculum in the AC courses was ostensibly the same as in the non-AC regular biology courses, non-AC course could, and typically did, include students identified as having a learning or behavioral disability.

an African American, implemented GenScope in three honors life science courses and three non-AC regular life science courses during Year 1; she left teaching for the private sector the following year.

Comparison Teachers

We succeeded in collecting comparison data from two other teachers. During Year 1, Ms P. at School 3 administered our learning assessments before and after genetics instruction in two of her AC life science classes. Ms. P was an African American with an undergraduate degree in biology and masters degree in science education; coincidentally, she was quite familiar with the GenScope curriculum and our assessment tools, having completed her teaching practicum in the classroom that was participating in our previous research project. During Year 3, Mr. H at School 1 also administered our instruments in two of his AC life science classes. Mr. H was a Euro American with a graduate degree in science education and had just begun work on his doctorate. Both comparison teaches were in the same district, which mandated a relatively well-defined life science curriculum. Reflecting district directives, both allocated the same number of class periods to introductory genetic as used in the GenScope project. All indications suggested that the comparison curriculum were very similar to the curriculum that their corresponding GenScope teachers were using prior to their participation in the project.

Genetics Curriculum and Formative Assessments

The organizing framework for our curriculum and instruction was a robust model of the developmental course of expertise in genetics, based on Kindfield's (1994) prior research (also Stewart & Hafner, 1994). Table 1 shows how the various aspects of domain reasoning can be classified along two primary dimensions: (1) Domain-general Reasoning Type (cause–to-effect, effect-to-cause, and process reasoning) and (2) Domain-specific Reasoning Type (within-generations and between-generations). In general, reasoning within generations (i.e., not involving inheritance) is easier than reasoning between generations; reasoning from causes to effects (e.g., from genotypes to phenotypes[2]) is easier than reasoning from effects to causes (from phenotypes to genotypes), which in turn is easier than reasoning about processes.

Both our prior efforts and the present investigation illustrate the point made in a recent report on educational research methods in both of the NRC assessment reports (as well as in Donovan, Pellegrino, & Bransford 1999). They argue that assessment and instruction should more reflect what is known about the development of expertise in the domain, rather than the scope and sequence associated with typical classroom instruction in the domain. Traditional life sciences curricula present many of the key concepts needed to fully understand introductory genetics outside of the "genetics" curriculum. For example, while meiosis is generally isolated from Mendelian inheritance, events that occur during meiosis (e.g., alignment and crossover) are critical to understanding Mendel's laws. This kind of linkage is perhaps the most promising affordances of the GenScope software, and is the sort of unique learning outcome that we targeted in our assessment practice.

In our prior effort (described in Hickey, Kindfield, Horwitz, & Christie, 1999; 2003) and continuing in the present investigation, our robust understanding of the developmental course of domain expertise coordinated the many aspects and potentially competing goals. As described below, the framework was used to coordinate the computer-based learning activities, the *Dragon Investigation* formative assessment materials, and the *NewWorm* summative assessment. This provides a useful interpretive framework for understanding transfer of learning from formative to summative assessment environments; this framework in turn helps us understand the complex issues that emerge from the inevitable blurring of the conventional distinction between *consequential* and *evidential* validity (e.g., Messick, 1994; 1995) within efforts to create systemically valid learning environments.

---

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

GenScope Genetics Curriculum.

The genetics curriculum was built around small group activities carried out using the GenScope software. 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, linking them in such a way that the effects of manipulations made at any one level are immediately reflected in each of the others. Most of the activities were 1-3 page exercises that structured students' inquiry and learning of key phenomenon, and could be completed within a single class period.[3] Fifteen activities based on materials developed by the GenScope developers were organized into four units designed to supplant the curriculum previously used during the 20 class periods normally devoted to introductory genetics.

The GenScope software runs only on Macintosh computers. Because these computers have become scarce in secondary schools, laptop and desktop computers were obtained from university surplus. Ten computers were installed in each classroom for at least the duration of the implementation. This is a departure from the previous GenScope implementations where students typically went to the computer lab to complete the activities (and many reported encountering substantial logistical challenges and confusion or problems with the software activities and lab hardware). In the previous study, the one classroom where GenScope activities were completed on laptop computers installed in the teacher's biology lab/classroom, learning gains were nearly double those found in any other classroom (Hickey, Kindfield, Horwitz, & Christie, 2003)

Formative Assessment Unit Tests and The Original Feedback Materials

For each of the four curricular units, we developed ambitious unit exams based on the *Dragon Investigations* formative assessment. These were developed by in the prior study in response to disappointing learning outcomes with the initial GenScope activities. The activities were designed to scaffold students' understanding of complex problems on the NewWorm posttest measure, but using the more familiar GenScope dragons. An example showing two of the three items that made up the assessment of dihybrid inheritance is included in Appendix A. Each unit test consisted of two of three such assessments. They were designed to be comprehensive and quite challenging.

For each of the four unit tests, we created text-based formative feedback materials. For each part of each unit assessment, we crafted a set of *Key Points* providing detailed explanation of the concept targeted by each of the assessments. A part of one of them is presented in Appendix B. The formative feedback materials also included *Answer Explanations* for each assessment item. As shown in Appendix C, these provided a detailed explanation of how each problem was solved, in light of the Key Points. They were designed so that students would have to read and comprehend the explanation in order to determine whether the answered the item correctly. Finally, for each of the 2 or 3 assessments in each of the unit tests, a *Judge Your Understanding* rubric was developed. As shown in Appendix D, these outlined the different types of problems covered in the assessment, and guided students through the process of evaluating their understanding of the targeted concept after having completed reviewing the answer explanations.

As the materials were being finalized for the first implementation, debate emerged within the project over the dense, complex language of the answer explanations. Some argued (1) that these materials targeted complex concepts that were difficult to explain in text using simplistic terms and sparse prose, (2) that the absence of such authentic discourse and representations was a major shortcoming of typical introductory genetics instruction, (3) that teachers and students had reportedly

[3] The GenScope software, the curricular activities, the dragon investigation formative assessments, and the NewWorm summative assessment can all be downloaded from http://genscope.concord.org/

found earlier versions of these materials very useful, (4) that the formative assessment context would provide scaffolding to help students use and comprehend the materials, and (5) that the earlier versions of the materials had been most effective in the lowest achieving classrooms. Others argued that (1) many ninth-graders would still be unwilling and/or unable to use materials as written, (2) that was unethical to present materials to students that so obviously exceeded their reading grade level, (3) that many technical vocabulary terms that had not been systematically covered in the curriculum. Indeed, even cursory examination of the materials in the appendices reveals that they were quite challenging. In the end, the impending implementation and exhausted development resources forced us to move forward with the materials as they were written during the first implementation.

In an effort to compare the motivational consequences of the range of typical grading practices, the completed unit tests were retuned to students in one of three conditions. In the *grade-oriented* classes, unit tests were marked with the percentage of items answered correctly, and graded (90% = A, 80% = B, etc.) as is convention. In the *standards-oriented* classes, unit tests were marked according to a scoring rubric that the teachers used to judge whether the understanding of the 3-5 concepts targeted by the unit tests appeared *exemplary, accomplished, developing,* or *beginning.* This scoring rubric was given to these students along with the answer key and answer explanation. The *accountability-oriented* classes were like the standards-oriented classes. In an effort to induce the sort of extrinsic, competitive environment that is intended by new educational reform policies, we had teachers invite students whose performance was assessed as exemplary or accomplished to volunteer to have their names place on a special board that was prominently displayed in the classroom.

Dependent Variables

Following from the theoretical line of inquiry described above, this project is attempting to reconcile competing views of knowing, learning and instruction. To this end, we deployed three different measures of engagement and three different measures of learning. Each set of measures was intended to be consistent with the assumptions of behavioral/empiricist, cognitive/rationalist/ or situative/sociohistoric views of knowing and learning (as outlined in Hickey & McCaslin, 2001; also Greeno, Collins, & Resnick, 1996; and Case, 1996).

Engagement Measures.

Two of our measures of engagement are based on different analyses of videotape recording of classroom activity. During Year 1, we recorded five triads of students in six classes (two each for Mr. N, Ms. P, and Ms. L). For each triad we recorded one GenScope activity and the Unit 1 and Unit 3 feedback sessions. During Year 2 and Year 3, we followed just two triads in the four classes taught by Mr. H and the two classes taught by Ms. P. In these classes we recorded one or two GenScope activities, the feedback sessions for all three units, and the informal feedback session when students were given back their final exams.

Three teams who were tasked with devise an interpretive methodology for interpreting these recordings that (1) was consistent with particular assumptions about knowing and learning, (2) helped answer questions about students engagement on the project in a manner that is consistent with the general practices of each teams respective scholarly peers, and (3) could be accomplished within the constraints of the videotape we had captured. These analyses are continuing for the Year 2 and 3 recordings, and the coding schemes evolved somewhat over time. Following are the overall approaches.

Behavioral/empiricist video analyses. Laura Fredrick headed one video analysis team. She is an applied behavior analyst whose specialization is direct instruction methods in K-12 setting (e.g., Fredrick, Deitz, Bryceland, & Hummell 2000). Consistent with empiricist assumption that learning is the building and strengthening of many small behavioral or cognitive associations that directly represent associations in the environment, this team's scoring method defined engagement as any behavior that involved the intended course content. They analyzed the tapes by coding the behavior of each student

in the group independently. Each student's behavior at five-second intervals was coded according to the following mutually exclusive and exhaustive categories:

Off Task- talking about non-academic subjects with friends.
On-Task/Surface Engagement- talking about grades, scores, answers, etc, but not concerning the content knowledge that they represent.
On-Task/Substantive Engagement- arguing about or otherwise discussing issues directly related to genetics.

Situative/sociohistoric video analysis. Ann Kruger headed the other video analysis team. She is a developmental psychologist who specializes in the application of discourse-analytic methods to the study of classroom culture and informal learning environments (e.g., Kruger & Tomasello, 1995). Consistent with the situative assumption that learning is represented by enhanced participation of the rituals and tools associated with the particular knowledge domain, this team analyzed discourse patterns within the triads of students, searching for emerging ritualized use of desirable knowledge practices associated with introductory genetics.

During Year 1, mediocre audio quality precluded the comprehensive transcription that is a prerequisite for analyzing discourse. In lieu, the socio-constructivist team coded the collective functioning of the student groups. As such, when only one student was on-camera, the activity was not analyzed. The collective activity of the group was segmented according to naturally occurring shifts in the shared activity; these segments were scored as being either:

In Group- physically and conversationally focused on their assigned partners, or
Out Group- physically and conversationally focused on students other than their partners.

Group interactions were further scored as belonging to one of the following mutually exclusive and exhaustive categories:

Off-Task- concerning things unrelated to the intended classroom activity or content.
On-Task/Grade-Oriented- concerning evaluation received at feedback.
On-Task/Rule-Oriented- concerning surface level procedures of the assignments and activities.
On-Task/Content-Oriented- concerning the intellectual content of the activity.

We analyzed the feedback sessions on 40 of the 120 Year 1 videotapes. These sessions were analyzed by tracking the shifts in social interaction of the three individuals and coding the resulting segments.

Cognitive/Rationalist analyses. Reflecting the mainstream assumptions that meaningful learning involves intrinsic sense-making processes two other measures reflected the assumption that an orientation towards intrinsic "mastery" goals is desirable, while orientation towards extrinsic "performance" goals is generally undesirable, and that the two orientations are largely orthogonal. (Ryan & Deci, 2000). This perspective directed our attention towards the students' behavioral manifestation of intrinsic motivation, as well as the self-report of their goal orientation. During Year 2 and Year 3 we videotaped students' participation in the informal feedback session. Because no extrinsic recognition was offered and grades had already been assigned, students "free choice" engagement in the formative feedback around their final exam was viewed as a direct measure of their intrinsic motivation to learn genetics. Additionally, we also administered self-report assessments of students' *motivational experiences* during the GenScope activities, and their *motivational orientations* before and after the GenScope curriculum. These analyses are continuing and preliminary results have been reported elsewhere (Michael, Hickey, & Zuiker, 2003). This is one reason why these findings are not detailed in this paper.

Another reason why these engagement outcomes are not reported here is because we found that they provided relatively little useful insights or feedback for our iterative improvement of the learning environment. As will be described below, the process of conceptualizing both the behavioral and sociocultural video analysis provided more immediately useful frameworks for guiding our more informal observations and subsequent revisions of the learning environment.

Learning Outcome Measures.

Central to any analysis of learning is the consideration of *transfer*. Specifically, for knowledge that is presumably acquired or developed in some learning environment to be "meaningful", it must somehow be useful in some other subsequent "transfer environment". As highlighted by the *Transfer or Trial* volume by Detterman and Sternberg (1993), ones view of transfer is fundamentally bound to ones assumptions about knowledge and (therefore) learning.

"Far-transfer" multiple-choice test. Implicit in conventional multiple choice assessment practices is the assumption that knowledge consists of cognitive or behavioral representations of numerous specific associations that are presented in the environment, as well as associations between those associations. Thus it makes sense to assess understanding by testing whether students can recognize specific associations that represent a sample of the universe of associations that knowledgeable individuals are presumed to possess. To this end, we developed a short multiple-choice assessment consisting of nine-items taken from released forms of the SAT II-Biology and the AP Biology test. In order to address the primary research questions, we needed a "far-transfer" assessment that would give us an idea of how the GenScope curriculum might impact performance on the sort of high-stakes assessments that many students (including the ones in the implementation) would have to excel at to obtain a high-school diploma. In order to provide a "fair" test for the non-GenScope comparison students, it was critical that we not simply select the genetics items that most closely matched the GenScope curriculum. To this end we identified 45 released items that targeted genetics more broadly. These items were then ranked according to difficulty (based on percentage of students who had answered them correctly when they were operational items). We then selected every fifth item to yield a nine-item test that would cover the entire range of proficiency. Reflecting the randomness of the process, the test ended up including an item that tested the Lamarkian misconception that acquired traits are passed on (e.g., *A dog whose ears were clipped when it was a puppy has a litter of puppies. Which statement best describes those puppies?).* This is a key concept that is directly presented in most conventional genetics curricula, but is not directly addressed in the GenScope curriculum (because it is presumed that students will develop the requisite conceptual understanding). As such, this assessment is somewhat biased in favor of the comparison curricula. As such, we characterize it as a "far-transfer" measure. The test was administered by the researchers and no feedback was provided to students or teachers in order to preserve its evidential validity.

"Near-transfer" NewWorm performance assessment. From the cognitive/rationalist perspective, knowledge consists of higher-level cognitive schema and structures that are constructed as part of the uniquely human ability to adapt the mind to make sense of the world. As such, transfer is analyzed by examining whether students are able to use the higher-level concepts presumably constructed to make sense of the learning environment to solve (i.e., make sense of) new problems that require some of those same concepts in the transfer environment.

One of the key dependent measures of individual learning in the present investigation was the *NewWorm* assessment (Kindfield, Hickey, & Yessis, 1999). This paper and pencil based performance assessment consists of many short-answer items involving a fanciful species whose genetics mimics those of GenScope dragons, but is novel and understandable to both GenScope and non-GenScope students. The items were organized around the developmental model of expertise shown in Table 1, and were carefully sequenced to scaffold student performance across increasingly complex problems. The instrument was designed to accurately assess the broadest range of expertise possible; while the initial items were solvable by most secondary student prior to formal genetics instruction, some of the

subsequent items proved challenging even to university biology graduate students and faculty. The instrument was revised across several years. The abbreviated version administered before and after genetics instruction in the present study consists of about 25 items and can be completed by most students in less than 40 minutes.

An obvious issue with the NewWorm assessment is that the formative assessment activities included in the GenScope curriculum are designed to give students specific experiences solving the kinds of problems presented in the NewWorm assessment. This is not a problem when comparing GenScope classes with each other. In fact, the close connection between the formative assessments and the NewWorm promises a very accurate measure of the amount of knowledge students construct under the different formative assessment conditions. However, the fact that the comparison students have not had this unique exposure means that the NewWorm assessment is fundamentally biased in favor of the GenScope classrooms. As such we characterize the NewWorm as a "near-transfer" measure, to differentiate it from the more objective "far transfer" multiple-choice test[4].

Situative/socio-constructivist analysis of transferable knowledge practices. The inherently contextualist world-view of situative/socio-constructivist perspectives precludes the conventional clear distinction between engagement and learning (because engagement in knowledgeable activity *is* learning, and vice versa). From this perspective, learning is the process of becoming more *attuned* to (i.e., familiar with) the social and physical constraints and affordances that simultaneously bound and scaffold successful participation in knowledgeable activity. As such, students are presumed to have learned transferable knowledge when able to participate more successfully in some transfer environment that presents at least some of those same constraints and affordances that they learned to negotiate in the learning environment (see Greeno, 1998; Gruber, Law, Mandl, & Renkl, 1995). Needless to say, this is a complicated analysis that presents difficult issues about what constitutes an appropriate transfer environment and how to characterize the "transformations" that relate the two. We had intended to simplify it by characterizing the feedback activity as the learning environment and the subsequent GenScope computer activities as the corresponding transfer environment. Thus we could examine whether the knowledge rituals that emerged in the feedback setting (particularly the appropriate use of scientific terms and concepts in discussions) were also used by students during the GenScope computer activities. However, given the poor audio quality and the finding described below we did not pursue this analysis for the Year 1 implementation; analyses are still underway for Years 2 and 3.

Independent Variables

As indicated above, within each teacher we systematically manipulated the context in which students completed their performance assessments at the end of each unit and used formative feedback on their performance. This ultimately did not impact performance on the outcome measures presented here. What turned out to be most important were the changes we made from one implementation cycle to the next. Specifically, in a focused effort to maximize gains on the near-transfer performance assessment, we refined the context in which all classes completed their classroom assessments and used formative feedback. These manipulations are reported below, in the context of the gains obtained across classes for each of the teachers in that implementation year.

YEAR ONE IMPLEMENTATION AND FINDINGS

As described above, four GenScope teachers at three schools implemented our curriculum in 13 life science classes during Year 1; we also administered our learning outcome measures in two

[4] This is not to say that we could not have identified cognitive/rationalist measures that represented near transfer from the GenScope environment, or that behavioral/empiricist measures represent far transfer. These issues are central to program evaluation efforts and are discussed at length in Hickey & Holbrook (2000) and Hickey & Zuiker (in press)

comparison classes at the high-SES school in Year 1, and in two classes in the low-SES school in Year 3 (these time-insensitive data are used for comparison purposes across the project years).

During the first implementation year, our primary concern was observing the consequences of our motivational manipulations on students' engagement in formative feedback, and their subsequent learning. As such we elected to make students use of the formative feedback materials optional. Thus the teacher returned completed unit assessments to the students on the first day of the next unit, along with the answer explanations. Students were encouraged to use the answer explanations to collaboratively to review their unit assessments before beginning the first computer-based GenScope activity for the next unit. We launched our first implementation with substantial reservations about the density and technical language of the answer explanations, particularly given the free-choice context of their use.

Results

Our concerns about the difficulty of the formative feedback material were immediately borne out in our informal observations, and subsequently confirmed in the videotape analyses of engagement. On many occasions, the research assistant responded to students' uncertainty about how to proceed by walking them through the materials and explaining that they were to review the answer explanations to figure out how to solve the problems correctly. The few students who actually attempted to review the feedback on their own eventually "rolled their eyes" and put the materials down. While some students compared their answers, and some discussed differences, there was little of the hoped for argumentative discourse where students would use knowledge of genetics and the data on the assessments to argue for or against particular position. Most quickly launched into the next GenScope computer activity (in retrospect, a relatively attractive activity for most students).

Engagement

Given these observations, we chose to analyze only 40 of the 120 hours of videotape, focusing only on the first and third feedback sessions in one grade-oriented and one standards-oriented class in both the high SES and low SES schools, for a total of ten tapes from four classrooms.

Behavioral engagement. For the 84 students whose behavior during the first and the third feedback activity could be scored, the nature of student engagement was coded at five second intervals according the criteria listed above. By summing the number of coded intervals and multiplying by five, we could estimate the number of seconds that students were engaged. This analysis revealed that the *mean time engaged in any behavior involving the unit test was just 157 seconds.* While 66% of this behavior was deemed "on-task", nearly all was coded as "surface-level" engagement (talking about grades, scores, answers, etc, but not concerning the content knowledge that they represent). Just 1.9% of the behavior was coded as substantive engagement.

Sociocultural engagement. Study social activity was coded according the categories descried above. Students spent roughly equal amounts of time off-task, on-task/grade-oriented, and on-task/rule oriented. Since students spent an equal amount of time in-group and out-group, the team concluded that group membership was not a compelling factor and did not figure prominently in the feedback experience. The analysis revealed just a few brief instances of on-task/content oriented activity.

Learning

Scores on the NewWorm and the multiple-choice test were scaled separately using *Facets* (Linacre, 1989). This Rasch technique makes it possible to (1) directly compare scores for students across the entire range of proficiency, (2) characterize proficiency according to the specific items and general types of items that students at that level of proficiency are able to solve, (3) compare gains to previous years as long as some of the assessment items are the same, and (4) reference proficiency to other benchmarks, such as university biology students and faculty, who have previously completed the NewWorm. For interpretability, raw logits were transformed to a T-scale (mean = 50, SD = 10).

NewWorm ("near-transfer") proficiency gains. Figure 2 shows students' reasoning gains according to the NewWorm performance assessment. Not surprisingly, every teacher's students showed statistically significant and substantive gains in student understanding. However, these gains observed in the GenScope classrooms are in the same range as those observed in 32 GenScope classrooms 40 GenScope 1996 and 1999. As observed in our prior studies, we again see that the mean proficiency for many of the low-SES classrooms *after* instruction is near or below the mean proficiency of some of the high-SES classrooms *before* instruction.

We limit our statistical tests to the two within-school comparisons. In all cases, teacher x time repeated measures analysis of variance were used. A non-nested design was used, meaning that some potentially important between-class within-teacher differences may have been overlooked. At the low SES school, the students in Mr. N's three classes gained 6.5 points. This gain was substantially larger than the gain of 2.5 in the two non-GenScope comparison classes taught by a different teacher at the same school, but this difference was likely to have occurred by chance, $F(1,80) = 1.3, p = .25$. At the same school, the students in Mr. P's three GenScope classes gained 12.1. Relative to the gain of 2.5 for the comparison students, this difference was unlikely to have occurred by chance $F(1,66) = 5.34, p = .02$. At the high-SES school, the gain of 15.7 in Ms. L's six GenScope classes was double the 8.2 gain in Ms. A's two non-GenScope comparison classes, a difference was extremely unlikely to have occurred by chance, $F(1,149) = 13.6, p < .001$. The larger gains in the GenScope classrooms were certainly expected, as the curriculum was closely aligned to the NewWorm assessment.

Multiple choice (far-transfer) proficiency gains. Figure 3 shows proficiency scores on the multiple-choice items before and after instruction. As expected, the GenScope students' gains on the far transfer tests were somewhat smaller than the gains on the NewWorm assessment. At the low-SES school, The students in Mr. N's three GenScope classrooms gained just 2.1, while the students in Mr. P's three GenScope classes actually declined by –2.2 points. Subsequent analysis revealed a significant interaction of class with one of Mr. P's classes declining sharply while the other two increased moderately. The students in the two non-GenScope comparison classes at the low-SES school increased by 5.7. The difference in gains between Mr. N's GenScope students and the comparison students were likely to have occurred by chance, $F(1,80) < 1$; the difference between Mr. P's students and the comparison students was less likely to have occurred by chance, $F(1,66) = 3.02, p = .087$. At the High SES school, the students in Ms. L's six GenScope classes gained 10.2, compared to the gain of 8.5 in the two non-GenScope comparison classes, a difference that was likely to have occurred by chance, $F(1,149) = 1.4, p = .24$.

Engagement in Feedback and Learning Outcomes.

While estimated engagement in the feedback activity was limited, we did observe a range of engagement. One student's was involved in feedback during 82 intervals (6.8 minutes) while some were never engaged at all; the rest were fairly normally distributed. In order to consider the relationship between learning gains and engagement in formative feedback, we examined the partial correlation of estimated time engaged in feedback with posttest learning outcomes (after partialling out pretest scores). The correlation for the near-transfer NewWorm assessment was .46 ($p < .01$). This generally confirms our common-sense expectation that being engaged in the formative feedback would be strongly related to gains on the NewWorm test, given how closely aligned the two assessment were. Not surprisingly, engagement in feedback was not significantly correlated with scores on the far-transfer multiple-choice assessment. Because our motivational manipulations did not lead to any systematic differences in engagement, we cannot make claims regarding the regarding causality of increased time on feedback and increased scores on the NewWorm.

Conclusion

While the implementation of the GenScope curriculum went reasonably well, we were disappointed by the limited engagement in the feedback activity. It is worth noting that despite the

limited participation in the feedback activities, one of the GenScope teachers made tremendous accomplishments with the GenScope curriculum. Ms. L at the high-SES schools obtained gains over a full standard deviation on the far-transfer multiple-choice test. Nonetheless, we had begun planning our revisions to the feedback materials and the feedback routines while this first implementation was still underway.

YEAR TWO IMPLEMENTATION AND FINDINGS

At the end of the first school Year, we lost two GenScope teachers, Mr. P at the low-SES school and Ms L. at the high-SES school. While this was disappointing, it also allowed us to focus our efforts more closely on the two remaining GenScope teachers in a wholesale effort to improve the quality of the curriculum and maximize the effectiveness of our formative assessment practice. This effort is best understood as an effort to increase students' engagement in the formative feedback activities, which was expected to directly increase gains on the near-transfer NewWorm assessment and indirectly increase scores on the far-transfer multiple-choice assessment.

Curricular Revisions

Formative feedback practice. Perhaps the most significant change we made was devoting an entire class period to collaboratively reviewing the formative assessments (rather than the free-choice activity during Year 1). We also asked the teachers to model the effective use of the feedback materials by reviewing the first set of items as a whole class activity. We began to realize that what we really wanted was for students to go beyond merely determining whether they got the item correct. Instead, we wanted students to focus more attention on understanding and discussing *why the correct answer is correct.* To use the logic of Scardamalia and Berieter's *intentional learning* framework, the "problem" that students are trying to solve during formative feedback should be their incomplete understanding of the concepts underlying the correct answer—not just the actual answer to the problem. In other words, the focus of student and teacher discourse should be on reaching consensus as to *why* the particular response was the correct answer.

As a result of attempting to code the Year One videos, we also began to realize that this would be manifested by discourse around a specific problem continuing *after* consensus on the correct answer was reached, trying to reach consensus on what made the particular answer correct. In light of this, we invested substantial effort in coaching the teachers to coach the students to not move on from a problem until each student in a group understood "why the right answer was right and why any wrong answers were wrong". Essentially, either the principle investigator or the project director would begin the formative feedback session with the students during each teacher's first class period, and invite the teacher to take over once they felt comfortable. This occurred about half way through the first class period for the first unit's feedback for both teachers; the researchers continued providing informal coaching for the teacher and some of the students during the first unit, and less so during the second unit.

Formative feedback materials. Our experience in Year One confirmed that our *Answer Explanations* were indeed too hard for these students to comprehend. We completely reworked the answer explanation materials to make them more readable. Given that many of these 9th graders were functioning at least one grade level-behind expectation, this turned out to be an enormous challenge. Appendix E shows one of the revised Answer Explanations; comparing it to the original version in Appendix C reveals that it is indeed more user friendly. Nonetheless, these materials were still quite challenging to these students and still included sentence length and language that was ostensibly written at about the college freshman level.

Grading practices. During Year One it became apparent that the various grading conditions made little difference to the students. Students in the standards-based classes readily converted our performance categories into A-B-C-D grades, and appeared no more inclined to use the formative

feedback materials than the other students. Additionally the process of scoring the exams placed an undue burden on the teachers (research staff ultimately assisted teachers in the process). As such we abandoned the process of grading the unit tests. Teachers instead returned the unit tests without grading them and relied entirely on the GenScope final for assigning students grades for the genetics portion of the class.

GenScope curricular materials. The materials used in Year 1 were provided by the GenScope development team. Our teachers tended to treat the materials as entirely student-directed activities, and students appeared to have difficulty moving beyond the GenScope-specific knowledge practices (i.e., dragons, wings, chromosome windows, etc) and into the language and tools of Genetics (i.e., homozygous alleles, recessive traits, Punnett squares, etc). While the GenScope-specific practices were expected to support learning gains on our near-transfer *NewWorm* assessment, it is the genetics practices that are expected to transfer to the far-transfer multiple choice items in our assessment and in the high-school graduation test that our students would encounter in two years.

In order for the GenScope activities to support engaged participation in genetics knowledge practices, it seemed to us that the GenScope activities needed to be treated as a means of creating a shared context of understanding that would support worthwhile discourse within groups and in whole class discussions. To this end we invested substantial resources into revising the existing materials. Twelve activities were completely redesigned so that the first part of the activity would be completed as a whole class investigation, with the teacher using an LCD panel; during the second part of the activity the students would complete a similar investigation working in triads at the computers. We also rewrote all of the student worksheets and teacher versions. In particular we included extensive color-coded information in the teacher versions, including correct answers, logistical pointers, and "key points" concerning the genetics concepts that needed to be covered in the lesson. Finally, we reorganized the curriculum into three units (from four) and dropped some of the most difficult content that teachers were not able to cover in the allotted time.

The combined revisions of the assessments, feedback materials, and curricular activities turned into a very substantial undertaking, occupying two science education doctoral students (including one of the implementation teachers) for roughly four months, and roughly 10 days of our genetics learning specialists time.

Research Methods Revisions

We made several additional revisions or refinements to address shortcomings or issue we encountered during the first year.

Video recording configuration. In order to obtain the quality of audio needed to do discourse analysis, we switched from five video groups per classroom to two and replaced the tabletop "PZM" microphones. We isolated the two triads in the corners of the rooms, and placed an individual lapel microphone on each student. We then ran the microphones into a portable mixer that allowed us to maintain the left-center-right separation of the stereo audio track recorded on the Hi-8 camera.

Sociocultural discourse analysis methods. A central challenge for our project was further operationalizing sociocultural measures of engagement. Our efforts were informed by the phenomenon that Duschl (Jimenez-Alexandre, Rodriguez, & Duschl, 2000) labeled "doing the lesson." Relative to the knowledge practices associated with scientific domains, Duschl argues that the vast majority of activity in science classrooms is consistent with what Bloome, Puro and Theodourou (1989, p. 272) called *procedural display:* "procedures that themselves count as accomplishment of a lesson...not necessarily related to the acquisition of intended academic or nonacademic content or skills...." In other words, instead of learning to "do the science", most of the knowledge practices in school science involve coping with the demands of the class and still getting a good grade, regardless of whether the actual knowledge practices of science are involved (also see Schauble, et al., 1995). After substantial deliberation, the discourse team defined the following mutually exclusive and exhaustive categories of increasingly adaptive forms of engagement in classroom discourse:

Off-task - Discourse that serves to distract the learning of science (i.e. "I saw a great movie last night. What did you do?").

Neutral - Discourse that does not serve learning but does not distract from it either (i.e. "Oh, my pencil lead broke.").

Procedural - Discourse that serves to clarify directions or routines of the assignment (i.e. "What page do we start on?").

Factual ("Doing the Lesson") - Discourse that serves to simply obtain the correct answer to the assessment without explaining, supporting, criticizing, evaluating, extending, clarifying or refining ideas about science covered in GenScope (i.e. "I put complete dominance for number 1.1.").

Argumentative ("Doing Science") - Discourse that includes explaining, supporting, criticizing, evaluating, extending, clarifying or refining ideas about genetics tied to GenScope (i.e. "Why did you put incomplete dominance? I say it is complete dominance because there is only two possibilities.").

Argumentative ("Doing Science Beyond GenScope") - Discourse that includes explaining, supporting, criticizing, evaluating, extending clarifying or refining ideas about science that is entirely removed from the immediate curricular context (i.e., "You don't understand? Well, think of how some human have different color eyes.")

For each tape of the formative feedback class periods, we transcribed the first ten minutes of discourse during the student directed segment (i.e., starting once the teacher-directed introduction was finished). This time was selected because all triads had a minimum of ten minutes of discourse, allowing the research team to compare equal amounts of discourse across triads, assessments and conditions. Each conversational turn was coded as belonging to one the five categories above.

Revision of the content test. One of the central goals of our project is producing gains that will transfer to high-stakes assessment items. Our "far-transfer" test in Year 1 was made up of items taken from the SAT II subject test and the AP Biology test. Scale scores from the Year 1 results revealed that many of the items were still beyond the proficiency of many students. Thus we dropped some of the harder items and added a number of simpler items take from the district-assigned biology text. However, enough of the items were constant across the two tests to allow us to compare gains across Year 1 and subsequent years.

Results

The revised materials and practices were used in the two classes taught by Ms. P at the medium-SES school and the four classes taught by Mr. N at the low-SES school. Our initial observations revealed that the overall revisions were helpful. The general climate classes seemed quite improved, and the formative feedback routine clearly yielded much more of the kinds of assessment-oriented discourse that we had been seeking all along.

Engagement

Behavioral engagement. In all six classrooms, devoting an entire class period to reviewing unit assessments and self-assessing understanding led to a dramatic increase in the amount of time engaged in the formative feedback activities. We have so far only coded the videotapes the two triads in each of Mr. N's four classes. Each five second interval was coded as either taking place during a *teacher*

directed activity (generally the whole class discussion and modeling of the feedback activity at the beginning of the feedback period) or *student directed* activity (the subsequent small group collaborative feedback activity). The teacher directed intervals were then additionally coded as *off-task* or *on-task;* the student direct intervals were additionally coded as either *off task, on-task independent* (either reading or listening to the teacher) or *on-task collaborative* (students were actively engaged in verbal interaction with their classmates that was consistent with our curricular intentions).

Estimated times engaged in particular types of activity were obtained by counting each interval as five seconds of behavior. Averaging across the first feedback activity and the third feedback activity revealed an average of 26 minutes of intervals coded as teacher-directed activity, with an average of 4 ½ minutes of intervals coded as *off-task* and 21 ½ minutes were coded *on-task.* An average of 32 ½ minutes worth of intervals were coded as small-group activity, with an average of 5 minutes coded *off-task,* 9 ½ minutes *on-task independent* and 4 minutes *on-task collaborating.* Figure 4, shows the distribution of behavior, averaged across students during each feedback session.

Obviously this is an enormous improvement over the Year One feedback activity. However, we still only managed to support an average of four minutes worth of behavior per period that was consistent with the ultimate goal of our efforts.

Sociocultural engagement. The first ten minutes of student directed activity on each of the tapes was transcribed. Each conversational turn was coded in terms of the nature of the discourse, according to the categories above. This yielded 4337 conversational turns. As shown in Figure 5, over half of the conversational turns involved discourse that was *Off Task, Neutral,* or *Procedural.* From a sociocultural perspective, such discourse is unlikely to directly support meaningful learning. We see that 30% of the conversational turns were coded as *Factual—Doing the Lesson*, while only 20% were coded as *Argumentative-Doing Science.* Only a single conversational turn involved discourse that was *Argumentative--Doing science-beyond GenScope.*

This analysis also suggests substantial improvement over Year One. But it also points to the need for further improvement. The behavioral analysis confirms that students were actually engaged in the feedback for substantial amounts of times. Both the behavioral and sociocultural analysis suggest that a relatively small proportion of that time was devoted to activity that could be expected to support student learning.

Learning

Students in all six classrooms completed the *NewWorm* and the multiple-choice test before and after instruction. As shown in Figure 6, the GenScope students made substantial average gains on the on the New Worm near-transfer measure. Across his four classes, Mr. N's students showed an average gain of 15.2, more than double the average gain of 6.5 in Mr. N's student during Year One, a difference that was extremely unlikely to have occurred by chance, $F(1, 132) = 15.7, p < 0.001$. Perhaps most significantly, (as shown if Figure 6), the gains by Mr. N's GenScope students were *six times larger* than the gain of 2.5 across the two non-GenScope comparison classrooms at the same school. Somewhat surprisingly, the students in Ms P's two classes showed a Year Two gain of 10.1, which was actually *smaller* than the 12.8 gain in Year One, $[F(1, 81) = 1.90, p = 0.171]$

As shown in Figure 7, both of the GenScope teachers also showed substantial gains on the far-transfer multiple-choice test. Mr. N's students showed an average gain of 7.4, which was over three times larger than the average gain of 2.1 across his three classes during Year One $[F(1, 132) = 4.78, p = 0.031]$. The average gain across Mr. N's classes was also larger than the 5.7 gain across the students in the two comparison classroom at the same school, although this difference was likely have occurred by chance, $[F(1, 82) = 0.24, p = 0.625]$. The average gain across Ms. P's two classes was 5.2, which was a modest increase over the average gain of 3.2 in Year 1, but this difference was also likely to have occurred by chance $[F(1, 81) = 0.50, p = 0.479]$.

Conclusion

We concluded that our revisions of the curriculum, the feedback materials, and feedback activities appeared to have led to an overall improvement in the learning environment. In particular, Mr. N's GenScope students showed dramatically improved gains compared to his students in Year 1. Notably, we achieved one of our primary goals for the first time. By refining our formative feedback activity to maximize gains on the closely aligned *NewWorm* assessment, we appear to have supported learning that transferred substantially to improved performance on the far-transfer multiple-choice test, leading to gains that were larger than those for the comparison students. Because that test was actually biased toward the comparison students, we believe this provides our best evidence yet about the power of our formative feedback activity.

Nonetheless, we still saw substantial room for improvement. A relatively small proportion of the behavioral engagement was coded as involving actual collaboration between students; likewise a relatively small proportion of the conversational turns during the student directed feedback activity was actually focused on understanding introductory genetics. Fortunately both the behavioral and sociocultural analyses gave direct guidance for our efforts to do so. Specifically, we redoubled our efforts to get students to spend more of their time during the formative feedback activity engaged in the kind of discourse that is expect to support meaningful learning.

YEAR THREE IMPLEMENTATION AND FINDINGS

During Year Three, we again implemented in four classes taught by Mr. H. at the low SES school and in two classes taught by Ms. P at the medium-SES school. The Year Three implementation was carried out in January and February of 2003.

Revisions

A few minor revisions were made to the curriculum and feedback materials. Most of our effort was directed at further refining the formative feedback activity. After substantial consideration, we decided to further structure the feedback activity in an effort to get students to spend as much time as possible engaged in meaningful argumentative discourse around genetics. To this end, we made a large chart of *Test Review Steps* place it in the classroom, stating the following:

TEST REVIEW STEPS

The more you review your unit tests, the better you will do on the final and the graduation test. Spend the entire period reviewing with your group. Use data and knowledge of genetics to support scientific claims.

FOR EACH ITEM:

1. Each student must state and defend a solution. If you don't know—guess!
2. Work together to figure out the best solution. Why is the correct answer correct and why are the wrong answers wrong? Each student should agree on the solution or "agree to disagree" before the next step.
3. Read the answer explanation (the yellow sheet) together. Compare that solution with your own solution(s). Each student should state whether their solution was the same as or different from that solution.
4. Make sure that every student understands why the correct answer is correct and why wrong answers are wrong before going to the next item.

These steps were used to structure both the teacher modeling during the whole class activity at the beginning of the period, and during the student directed activity during the rest of the period. Our logic was guided by our prior observation that we could observe clear transitions when student moved from comparing answers to reviewing the answer explanation, and when student moved from one item to the next. Specifically, it was apparent if students reached a desirable consensus before making each transition. As such, both transitions offer the ideal opportunity for researchers to observe the structure of the student discourse. We reasoned that they would also provide an ideal structure for scaffolding students idealized participation in discourse around the formative feedback. We used this structure to coach students to make sure that every member of the triad had explained how they solved the problem before the group turned to the answer explanation, and that every member understood "why the right answer was right and why the wrong answers were wrong" before moving on from the item. We coached the teacher to first model this activity structure as a whole class feedback session on the first set of assessment items, and then carefully monitor the progression of discourse in the groups. In practice, a major part of this involved coaching students to slow down and focus on the solution processes rather than just comparing answers.

Results

While we have yet to complete coding the video from Year Three, our informal observations suggested that we did indeed make further progress in refining the feedback routines. Particularly in Mr. N's classes, we were quite pleased with the way that most groups seemed to correctly appropriate our intentions for the formative feedback activity. As a caveat, we point out that Ms. P was absent on the day when the posttest was administered. This led to administrative difficulties, with an unknown number of students reportedly unable or unwilling to complete both assessments in the allotted time. While the order of the two tests was counterbalance, the NewWorm performance assessment took most students 2-3 times longer to complete than the multiple choice test.

As shown in Figure 8, Mr. N's students gained an average of 19.8 points on the near-transfer performance assessment. This gain was 4.6 points larger than Mr. N's average gains in Year 2, [F (1, 119) = 3.81, p =.053]. In contrast, the gains in Ms. P's class on the NewWorm were somewhat disappointing, averaging just 8.8 across her two classes, lower than the 10.1 in Year 2 and the 12.8 in Year 1. Examination of the NewWorm scores showed that a number of students of the students actually showed score declines, indicating lack of motivation or time to complete this fairly lengthy assessment. Because the most difficult items at the end of the test are the ones that provide the highest scale scores, students who run out of time can be severely penalized on the test. Further analyses of these scores are underway.

As shown in Figure 9, Mr. N's students' gained an average 10.6 points on the far-transfer multiple choice test. This increased was over a full standard deviation, and almost double the average gain for the comparison students at the same school. However, due to the wide variation in gains and the small number of comparison students, this difference in gains had a roughly 1:5 possibility of occurring by chance [F (1, 67) = 1.59, p = 0.212]. Mr. N's students' Year 3 gain of 10.6 points was three points larger than his students in Year 2, but this difference also may have occurred by chance [F (1, 119) = 1.63, p = 0.205]. Ms. P's students gained 8.0 on the far transfer measure. While substantially larger than the Year 2 gain of 5.2, this difference was likely to have occurred by chance [F (1, 71) < 1].

Conclusions

Our efforts to revise the formative feedback routine showed continued progress in improving students test scores. The average gains on the far-transfer multiple-choice test across Mr. N's students in Year 3 was over one full standard deviation, and almost three times as large as the gains in the comparison classes at the same school. As the test was more closely aligned with the curriculum in the

comparison classroom, we conclude that the significant learning that was documented on the NewWorm transferred to student performance on the multiple-choice test.

Looking across our three annual implementations suggest that our efforts to iteratively refine classroom assessment practices is promising way of obtaining long-sought gains on externally developed, multiple choice achievement measure. To reiterate our success in this regard, Figures 10 and 11 display Mr. N's students' learning gains on the two outcome measures across the three study years. Both outcome measures show continuous improvement and increasingly large gains relative to the comparison classroom. We reiterate that the far-transfer measure included items that were more likely to be directly presented in the comparison classroom, and were never directly presented in the GenScope classroom.

OVERALL CONCLUSIONS

We conclude that our study offers a promising approach that instructional innovators can use to obtain heretofore-elusive outcomes on externally developed high-stakes tests. Our approach uses a deliberately comparative perspective and design-based research methods to "tune" classroom assessment practices to maximize scores on external assessments. While the use of a comparative approach is innovative, we believe its use around carefully aligned classroom and external assessments makes our study unique. We found the comparative approach was particularly useful for understanding the difference between different types of outcome measures, and reminding us of the features of the particular learning environments that compromise the validity of particular types of outcome measures.

We believe that our study also illustrates the power of design-based research methods. We used scientific methods and our assumptions about learning to meet clearly defined expectations, across three annual implementation cycles. In doing so, we developed and refined 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" (DBRC, p. 10), targeting an "intermediate" theoretical scope (diSessa, 1991). We acknowledge that our study was not explicitly proposed or conceptualized as design-based research. Indeed, there are several key areas where our failure to heed the basic premises of design-based approaches cost us substantial time and effort. In particular, we clearly should have started with a more well articulated model of how we expected learning to occur during the classroom formative feedback session. In addition to reminding us to clearly specify one's presumed "developmental trajectory of expertise", Cobb, Confrey, diSessa, Lehrer, & Schauble, (2003) point to the value of clearly articulating the presumed starting point of that trajectory. If we had heeded such advice from the outset of our project, it probably would have not required three years to devise a seemingly effective curricular context for the formative feedback activity.

We also believe that our effort also illustrates how the application of contemporary notions of assessment can enhance the scope of design-based educational research. In retrospect, we found that the framework advanced by Ruiz-Primo, Shavelson, Hamilton, and Klein (2002) was useful for characterizing the distance between our various measures and the enacted GenScope curriculum (i.e., *immediate, close, proximal, distal* and *remote*). We ultimately organized our effort around the matrix that results when crossing distinctions between internal curriculum-oriented and external standards-oriented assessment with the three different views of knowing and learning. We believe that others will find this a useful framework for organizing their own efforts. For example, this framework helps illuminate the common practice of "cherry-picking" items from existing high-stakes. Once selected from a larger "distal" instrument, such items become more proximal, limiting claims about the generalizability of resulting scores. We believe that such extensions can help design-based studies of curricular innovations address concerns of critics (e.g., Levin & O'Donnell, 1999) and skeptics (e.g., Shavelson, Phillips, Towne, & Feuer, 2003) of these approaches.

We conclude by expressing our enthusiasm for design-based studies of assessment practices around promising instructional innovations. Design-based methods seem ideal for refining the

alignment of innovative curriculum, classroom assessments and external assessments, and maximizing the impact of formative feedback at the various levels. 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, 2002b; Pellegrino & Goldman, 2002).

References

Bereiter, C. & Scardamalia, M. (1989). Intentional learning as a goal of instruction. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 361-385). Hillsdale, NJ: Lawrence Erlbaum Associates.

Black, P. & Wiliam, D. (1998). Asssessment and classroom learning. *Assessment in Education,* 5, (1).

Bloome, D., Puro, P., & Theodoru, E. (1989). Procedural display and classroom lessons. *Curriculum Inquiry, 19*, 265-291.

Bransford, J. B., Brown, A. L. Cocking R. (1999) (Eds.). *How people learn: Brain, mind, experience, and school.* Committee on Learning Research and Educational Practice. Washington, DC: National Academy Press.

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

Cameron, J. & Pierce, W. D. (1994). Reinforcement, reward, and intrinsic motivation. A meta-analysis. *Review of Educational Research, 64,* 363-423.

Case, R. (1996). Changing views of knowledge and the impact on educational research and practice. In D. R. Olson and N. Torrance (Eds.). *The handbook of education and human development*, (pp. 75-99). Blackwell: Cambridge.

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-Verlag.

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

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics. In In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453-494). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Detterman, D. K., & Sternberg, R. J., (Eds.) (1993). *Transfer on trial: Intelligence, cognition, & instruction.* Norwood, NJ: Ablex.

Donovan, M. S., Pellegrino, J. W., & Bransford, J. D. (Eds.) (1999). *How people learn: Bridging research to practice.* Washington, DC: National Academy Press.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84,* 287-312.

Duschl, R. A. & Gitomer, D. H. (1997). Strategies and challenges to changing the focus of assessment and instruction in science classrooms. *Educational Assessment, 4,* 37-73.

Fredrick, L. D., Deitz, S. M., Bryceland, J. A., & Hummel, J. H. (2000). *Behavior analysis,*

education, and effective schooling. Reno, NV: Context Press.

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

Feuer, M. J., Towne, L. & Shavelson, R. J. (2003). Scientific culture and educational research. *Educational Researcher, 31* (8), 4-14.

Gipps, C. (1999). Sociocultural aspects of assessment. *Review of Research in Education.*

Graham, S. (1994). Motivation in African Americans. *Review of Educational Research, 64*, 55-117.

Graue, M. E. (1993). *Integrating theory and practice through instructional assessment. Educational Assessment, 1 (4),* 283-309.

Greeno, J. G. (1998). The situative of knowing, learning, & research. *American Psychologist, 53,* 5-26.

Greeno, J. G., Collins, A. M., & Resnick, L. (1996). Cognition and learning. In D. Berliner and 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., Lauman, B., & Linde, C. (1999). 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 Publishers.

Gruber, H. Law, L. C., Mandl, H., & Renkl, A (1995). Situated learning and transfer. In P. Reimann & H. Spada (Eds.), *Learning in humans and machines: Towards an interdisciplinary learning science* (pp. 168-188). Oxford: Pergamon.

Hickey, D. T. (1997). Motivation and contemporary socio-constructivist instructional perspectives. *Educational Psychologist, 32,* 175-193.

Hickey, D. T. (in review). Engaged participation vs. marginal non-participation: A stridently sociocultural approach to achievement motivation. In review, *Elementary School Journal.*

Hickey, D. T. (forthcoming). A pragmatic, situative framework for evaluating innovative science learning environments. *Science Education.*

Hickey, D. T., & Holbrook, J. (2000, April). *PALS-supported performance assessments for the Learning by Design project.* Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA.

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., Moore, A. L., & Pellegrino, J. W. (2001). The motivational and academic consequences of two innovative mathematics environments: Do curricular innovations and reforms make a difference? *American Educational Research Journal, 38* (3).

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., Kindfield, A. C. H., Horwitz, P., & Christie, M. A. (2000). Integrating instruction, assessment, & evaluation in a technology-supported genetics environment: The *GenScope* follow up study. In B. Fishman and S. O' Conner (Eds.), *Proceedings of the Fourth International Conference of the Learning Sciences.*

Hickey, D. T., Kindfield, A. C. H., Horwitz, P., & Christie, M. A. (in press). Integrating curriculum, instruction, assessment, and evaluation in a technology-supported genetics environment. *American Educational Research Journal, 40* (2) (Summer 2003)

Hickey, D. T., & McCaslin, M (2001). Comparative and sociocultural analyses of context and motivation. In S. Volet, S. & S Järvelä (Eds.), *Motivation in learning contexts: Theoretical and methodological implications.* (pp. 33-56). Amsterdam: Pergamon/Elsevier

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.

Hickey, D. T, & Zuiker, S. (in press). A new perspective for evaluating innovative science learning environments. *Science Education, 87,* (3).

Horwitz, P. & Christie, M. (2000). Computer-based manipulatives for teaching scientific reasoning: An example. In M.J. Jacobson & R.B. Kozma, (Eds.), *Learning the sciences of the Twenty-first century: Theory, research, and the design of advanced technology learning environments.* Hillsdale, NJ: Lawrence Erlbaum & Associates.

Horwitz, P. (1999). *BioScope: Linked computer-based manipultives for biology.* National Science Foundation Grant REC 975524 to the Concord Consortium.

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

Jiménez-Aleixandre, M. P., Rodríguez, A. B., & Duschl, R. A. (2000). "Doing the lesson" or "Doing science": Argument in high school genetics. *Science Education, 84,* 757-792.

Kellaghan, T., Madaus, G. F., & Raczak, A. (1996). *The use of external examinations to improve student motivation.* Washington, DC: American Education Research Association.

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

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

Kruger, A.C., & Tomasello, M. (1996). Cultural learning and learning culture. In D.R. Olson & N. Torrance (Eds.), *Handbook of education and human development: New models of learning, teaching, and schooling* (pp. 369-387). Cambridge: Blackwell.

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 Publishers.

Lave, J. (1988). *Cognition in practice.* Cambridge: Cambridge University Press.

Levin, J. R., & O'Donnell, A. M. (1999). What to do about educational research's credibility gap? *Issues in Education, 5* (2), 177-229

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

McCaslin, M. & Hickey, D. T. (2001a). Self-regulated learning and academic achievement: A Vygotskian view. In B. Zimmerman and D. Schunk (Eds.), *Self-regulated learning and academic achievement: Theory, research, and practice, Second Edition* (pp. 227-252) Mahwah, NJ: Erlbaum .

McCaslin, M., & Hickey, D. T. (2001b). Educational psychology, social constructivism, and educational practice: A case of emergent identity. *Educational Psychologist, 36,* 133-140.

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

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

Micheal, M., Zuiker, S., Hickey, D. (2003, April). *Motivating students to engage in meangingful assessment conversations: Effects of extrinsic recognition on learning and motivation.*

Presentation at the Annual Meeting of the American Educational Research Association, Chicago.

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

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

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

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

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

National Research Council (2002a). Committee on Assessment in support of instruction and learning: Bridging the gap between large scale and classroom assessment. J. M. Atkin (Chair).

National Research Council Board on Testing and Assessment: Washington, DC.

National Research Council (2002b). Scientific inquiry in education. R. J. Shavelson & L. Towne (Eds.). Committee on Scientific Principles for Educational Research. Washington, DC: National Academy 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 (PCAST) (1997, March). *Report to the president on the use of technology to strengthen K-12 education in the United States.* Author.

Ryan, R. M., & Deci, E. L. (2000). *When rewards compete with nature: The undermining of intrinsic motivation and self-regulation.* In C. Sansone & J.M. Harackiewicz (Eds), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 14-48). San Diego, CA: Academic Press.

Schauble, L., Glaser, R., Duschl., R., Schulze, S., & John, J. (1995). Students' understanding of the objectives and procedures of experimentation in the science classroom. *Journal of the Learning Sciences, 4,* 131-166.

Shavelson, R. J., Phillips, D. C., Towne, L., & Feuer, M. J. (2003). On the science of educational design studies. *Educational Researcher, 32* (1), 25-28.

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,

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.

Stoakes, D. E. (1997). *Pasteur's quadrant" Basic science and technological innovation.* Washington, DC: Brookings Institution Press.

Tudge, J. R. H. & Winterhoff, , P. A. (1993). Vygotsky, Piaget, and Bandura: Perspectives on the relations between the social world and cognitive development. Human Development, 36, 61-81.

Turnstall, P, & Gipps, C. (1996). Teacher feedback to young children in formative assessment: A typology. *British Educational Research Journal, 22,*

Vygotsky, (1978). *Mind in society.* Cambridge: MIT University Press.

Zimmerman, B. J. (1993). Commentary. Human Development, 36, 82-86.

Table 1. Primary dimensions of reasoning in introductory genetics.

		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

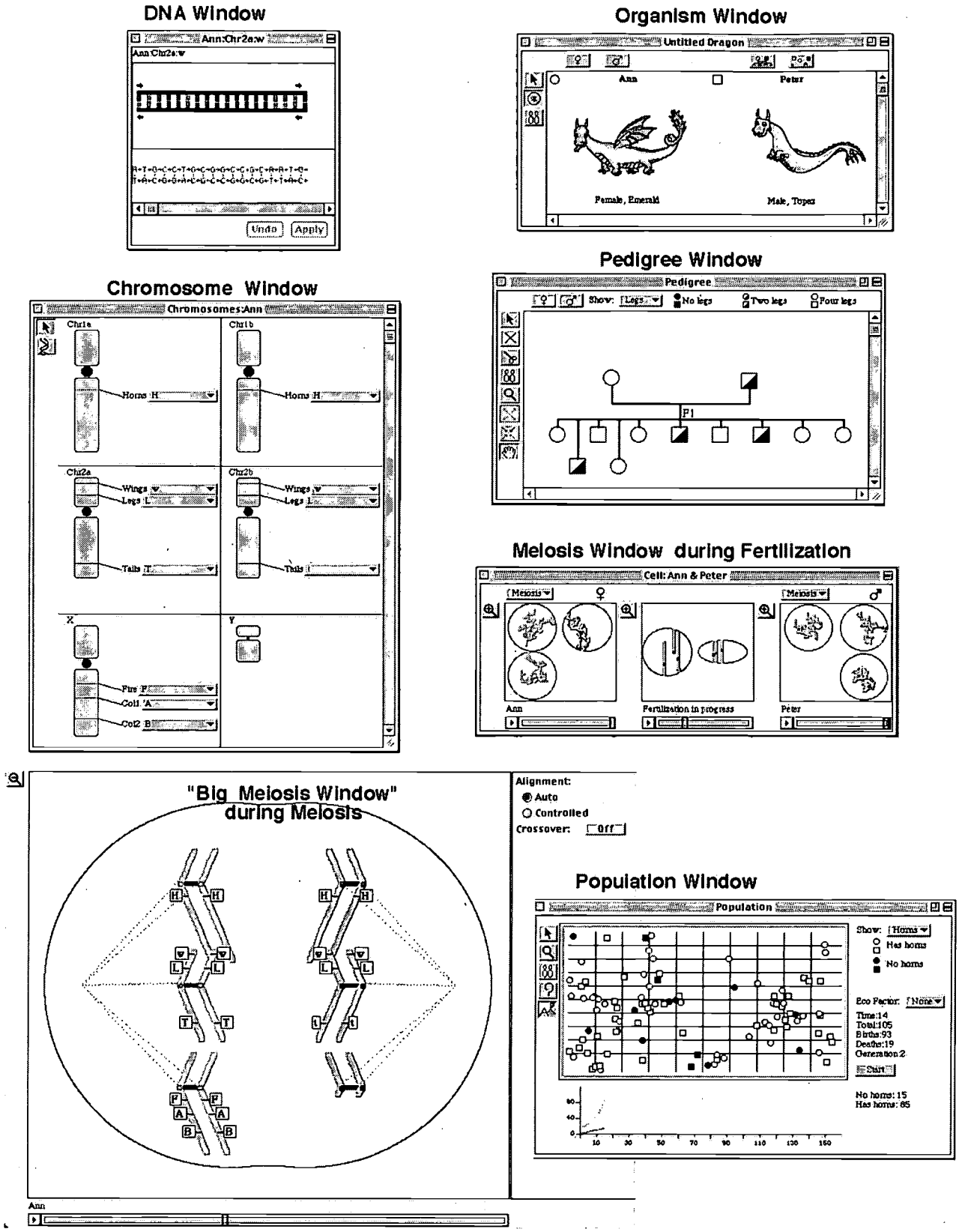


Figure 1. Examples of Screens in the *GenScope* Software.

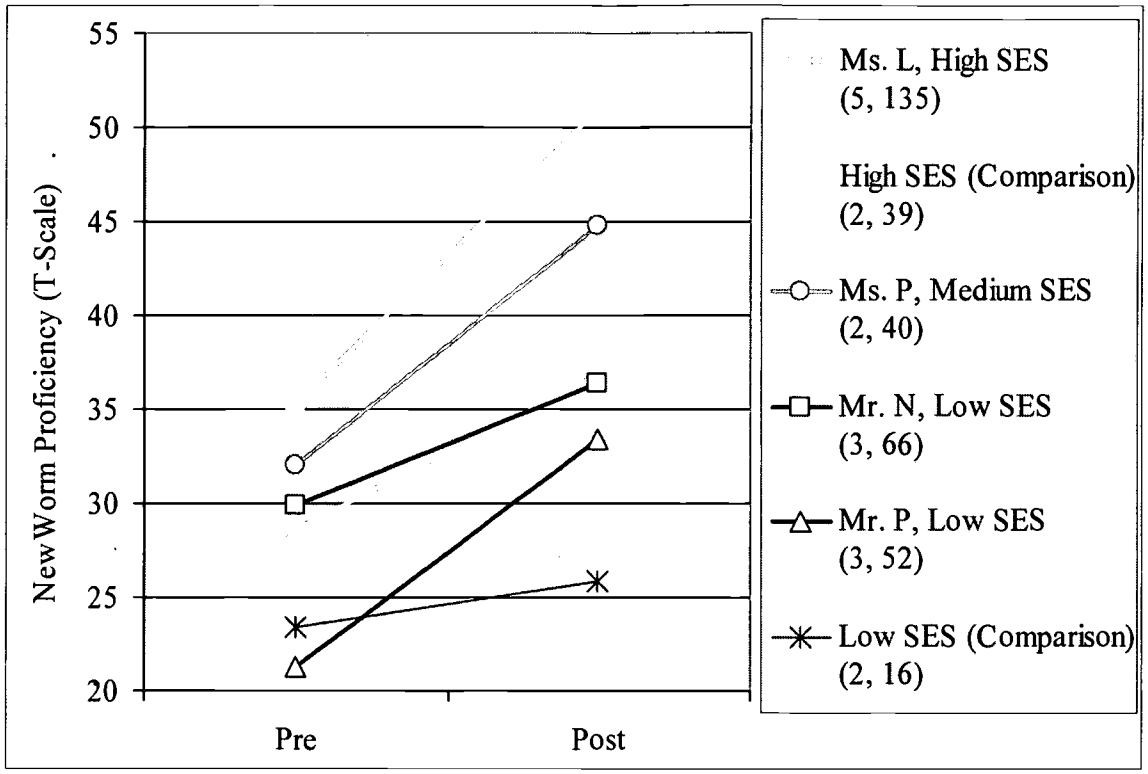


Figure 2. Proficiency gains on *NewWorm* "near-transfer" performance assessment during Year One

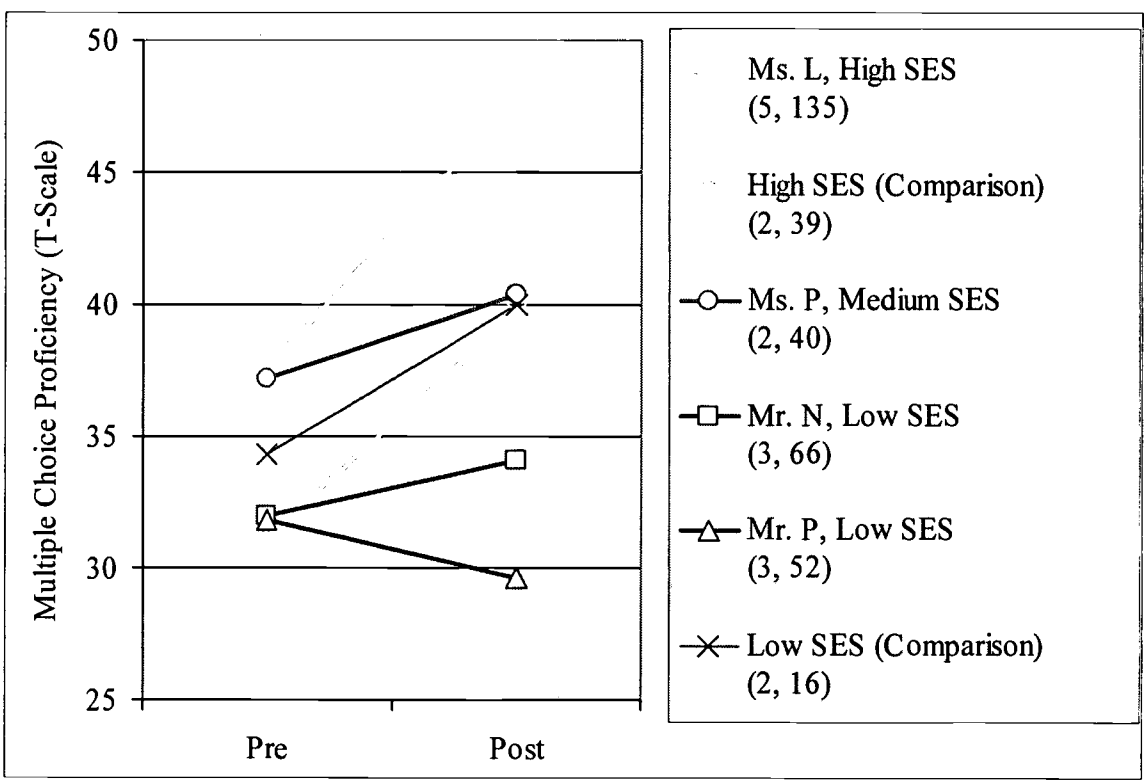


Figure 3. Proficiency gains on multiple-choice "far-transfer" test during Year 1

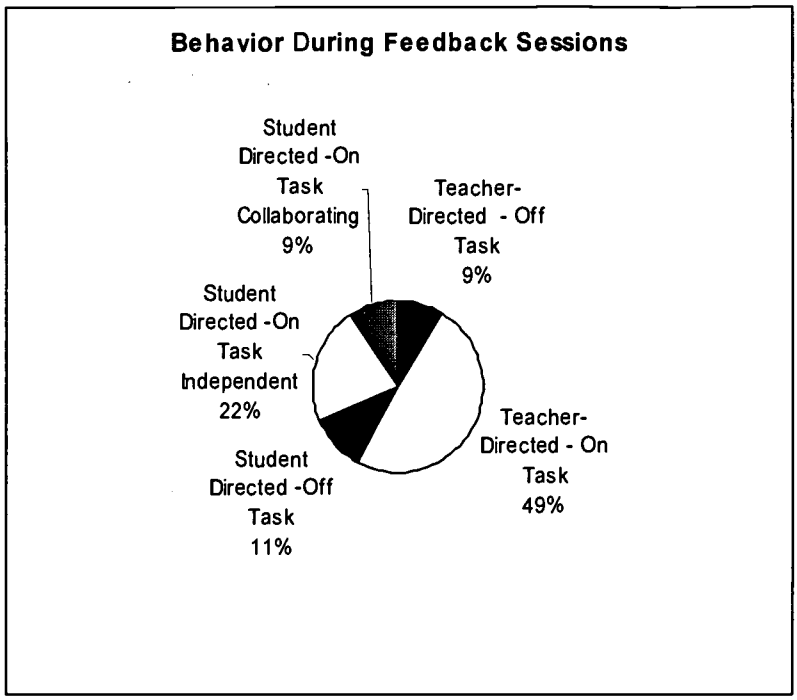


Figure 4. Distribution of behavior during formative feedback sessions during Year Two.

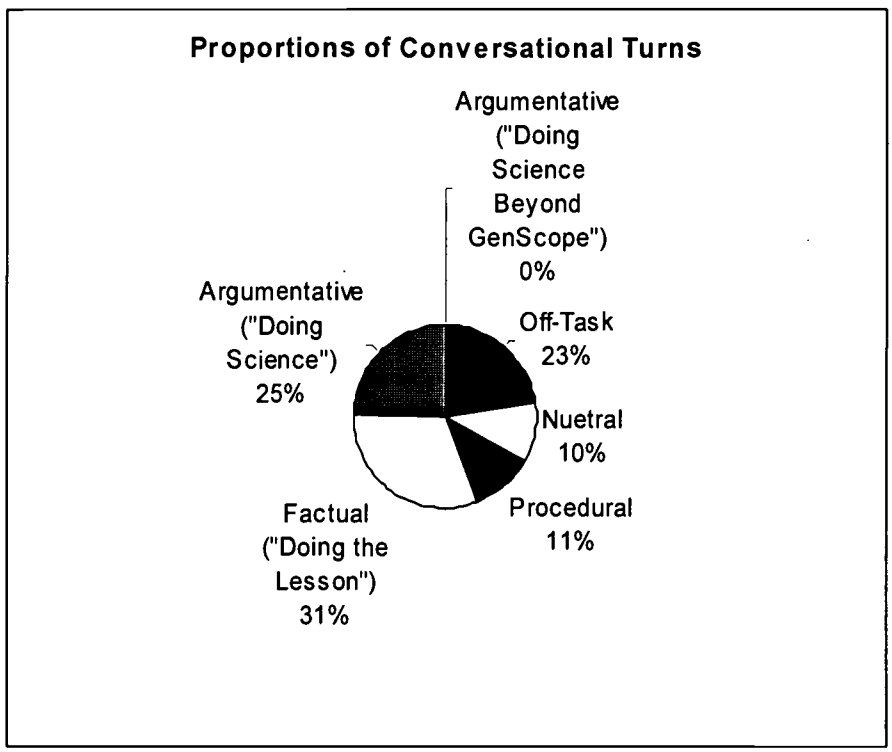


Figure 5. Distribution of types of assessment discourse transactions during Year Two.

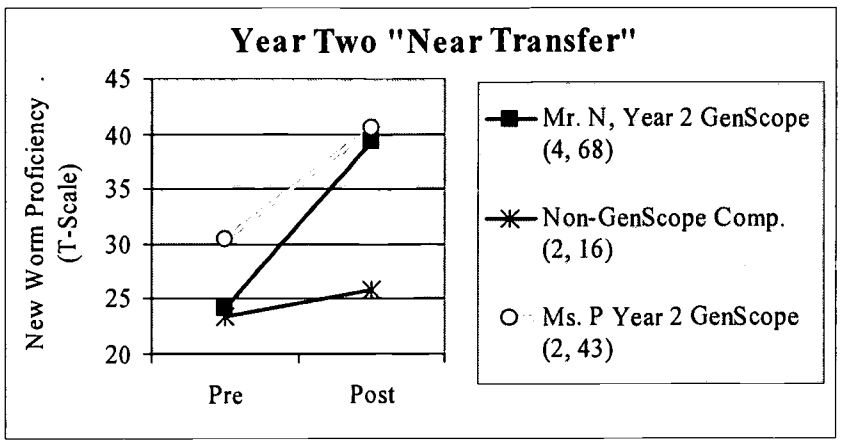


Figure 6. Gains across classes for GenScope teachers during Year 2 and comparison teacher on near-transfer performance assessment.

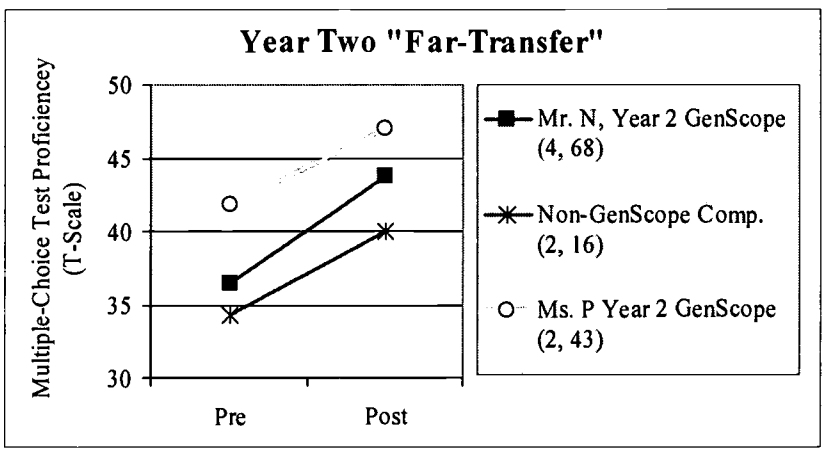


Figure 7. Gains across classes for GenScope teachers during Year 2 and comparison teacher on far-transfer multiple choice test.

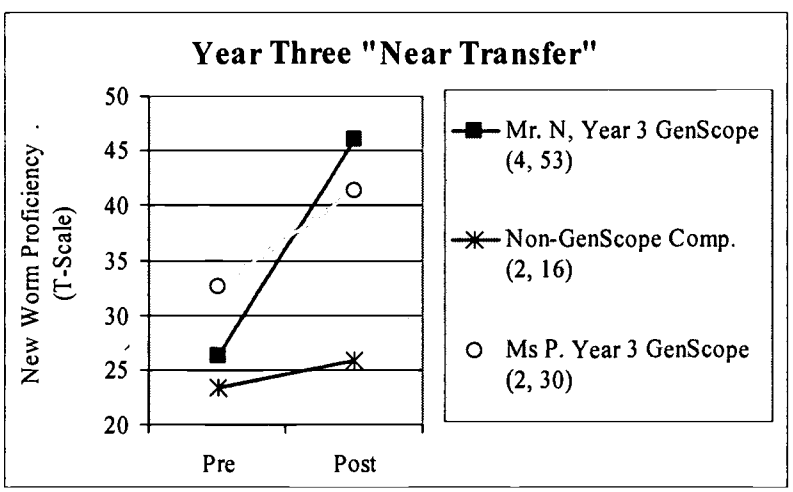


Figure 8. Gains across classes for GenScope teachers during Year 3 and comparison teacher on near-transfer performance assessment.

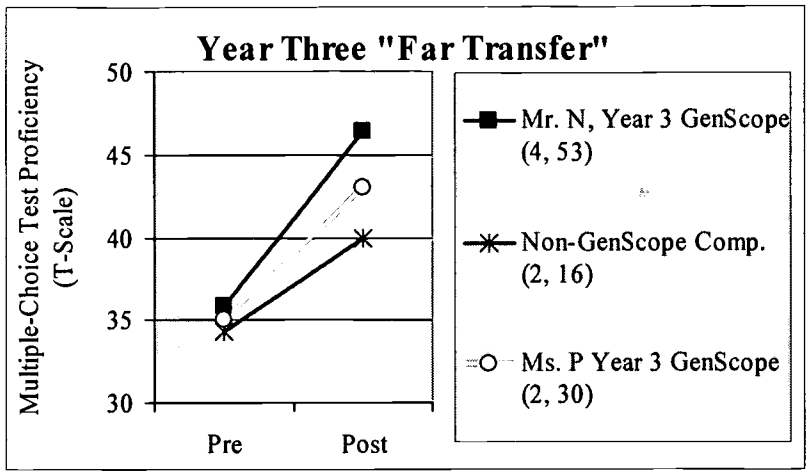


Figure 9. Gains across classes for GenScope teachers during Year 2 and comparison teacher on far-transfer multiple choice test.

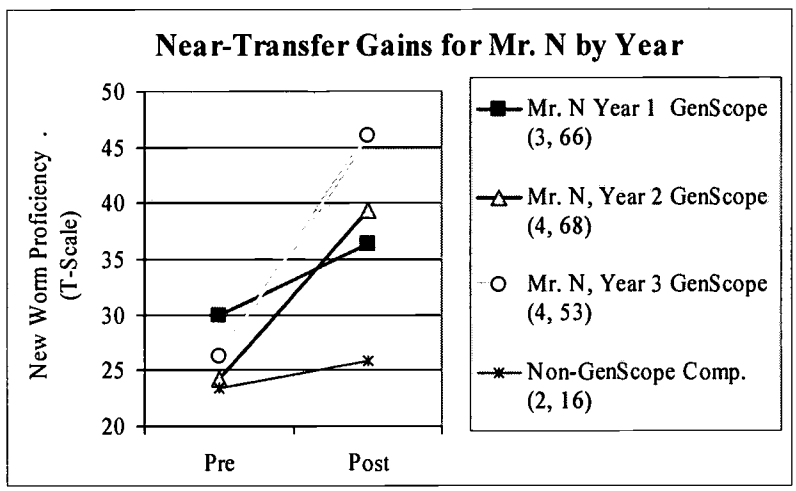


Figure 10. Gains across Mr. N's GenScope classes across years and a non-GenScope comparison teacher at the same school on the near-transfer performance assessment.

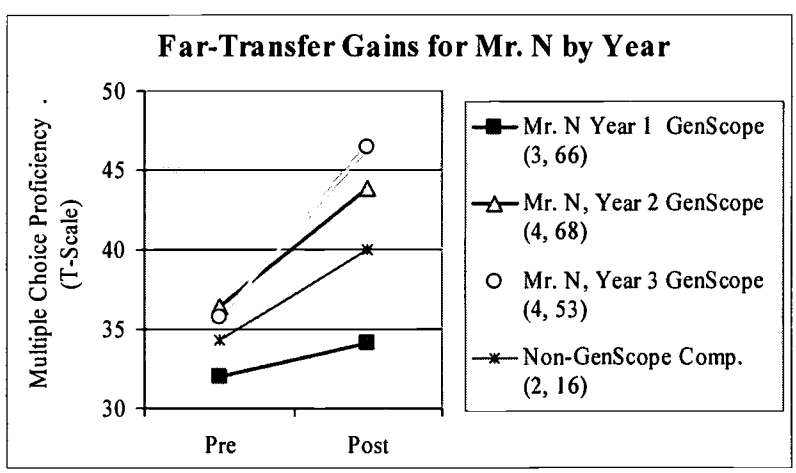


Figure 10. Gains across Mr. N's GenScope classes across years and a non-GenScope comparison teacher at the same school on the far-transfer multiple-choice test.

Appendix A: Sample Formative Assessment

Section 2B: Assessment

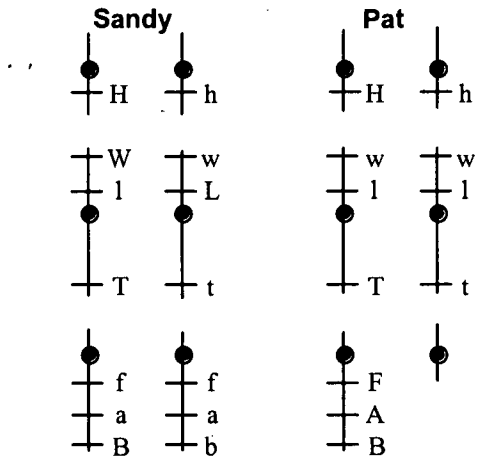
From Parent to Offspring III: Dihybrid Inheritance I

Sometimes it is useful to figure out inheritance for more than one characteristic at a time. Working with two characteristics at a time is called *dihybrid* inheritance.

DRAGON GENETICS

Horns:	Horns dominant to no-horns.
Wings:	Wings recessive to no-wings.
Legs:	4-legs incompletely dominant to no-legs; 2-legs intermediate. (LL= 4-legs)
Tail:	Fancy-tail dominant to plain-tail.
Fire:	Fire-breathing recessive to not-fire-breathing.
Sex:	Females are XY. Males are XX.
Note:	The — indicates that the gene is **not** present in the Y-chromosome

TWO DRAGON GENOTYPES



Questions 1-3:

Q1 & Q2: **Finish or make & fill in the Punnett square for each problem. Then use the information to answer the questions about the possible offspring (The first one is started for you.)**

1a. Horns & Wings (HhWw X Hhww)

Sandy / Pat	HW	Hw	hW	hw
Hw	HHWw horns/ no wings	HHww	HhWw	
hw	HhWw horns/ no wings	Hhww	hhWw	

1b. If Sandy & Pat have one baby, will it have **no horns** and **no wings**?

Definitely yes_____ Maybe_____ Definitely no_____

1c. What are the chances that Sandy & Pat's baby will have **no horns** and **no wings**?

0 ___ 1/8 ___ 1/4 ___ 3/8 ___ 1/2 ___

5/8 ___ 3/4 ___ 7/8 ___ 1/1___

2a. Horns & Legs (HhLl X Hhll)

Sandy / Pat	HL	Hl	hL	hl
Hl				
hl				

2b. If Sandy & Pat have one baby, will it have **four legs** and **no horns**?

Definitely yes_____ Maybe_____ Definitely no_____

2c. What are the chances that Sandy & Pat's baby will have **two legs** and **horns?**

(Continues with one more item where students have to draw Punnett)

Appendix B: Formative Feedback ("Key Points")

Section 2B: Key Points
From Parent to Offspring III: Dihybrid Inheritance I

These activities deal with **dihybrid inheritance,** where you pay attention to the inheritance of two single-gene characteristics at a time. In addition, these crosses include examples of **complete dominance, incomplete dominance**, and **X-linked inheritance**.

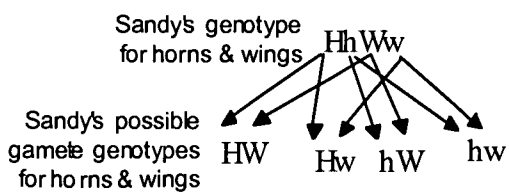
As with monohybrid inheritance, in dihybrid inheritance a **Punnett square** is used to determine offspring possibilities and probabilities. The Punnett square is a tool that helps you keep track of the gametes that each parent can produce and the possible ways to combine the gametes from each parent to produce offspring genotypes. The Punnett square does not show the actual offspring, only the possible genotypes that can be found in any given offspring

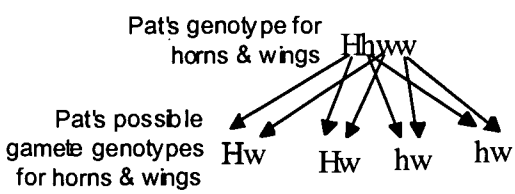
There are 11 steps to determining the genotypes of offspring in dihybrid inheritance using a **Punnett square**.

1. Determine the characteristics that you are interested in examining. In this example, we will use the horns and wings characteristics.
2. Use the genotypes to determine what alleles you will be crossing. Sandy is heterozygous for horns (**Hh**) and for wings (**Ww**) and Pat is heterozygous for horns (**Hh**) and homozygous recessive for wings (**ww**). So in order to figure out the possible horns and wings phenotypes of their babies, you first need to set up the following cross:

HhWw	X	**Hhww**
Sandy's genotype	Crossed with	Pat's genotype

3. Body (somatic) cells of parents and offspring contain two copies of each autosomal gene like the Horns gene or the Wings gene. Gametes contain only one copy. Since Sandy is **Hh** for horns, he can produce gametes that contain either **H** or **h**. Since he is **Ww** for wings, he can produce gametes that contain either **W** or **w**. Since Pat is **Hh**, she can produce gametes that contain either **H** or **h**. Since she is **ww**, all of her gametes will contain **w**.
4. Since each gamete produced by Sandy or Pat contains one copy of the Horns gene and one copy of the Wings gene, you need to figure out how to combine Horns and Wings alleles to produce all possible gamete combinations of horns and wings for each dragon. The diagram below shows how to do this.





5. In the diagram, you can see that Sandy produces four different gamete genotypes (**HW**, **Hw**, **hW**, **hw**) and Pat produces two different gamete genotypes (**Hw**, **hw**). Given these gamete genotypes, you can now draw a dihybrid Punnett square

(Continues for 1.5 more pages)

Appendix C: Formative Feedback ("Answer Explanation")

Section 2B: Key Points
From Parent to Offspring III: Dihybrid Inheritance I

1. This Punnett square is done for you in the **Key Points**. By examining the inner squares, you can see that four different offspring types are possible: horns/no-wings, horns/wings, no-horns/no-wings and no-horns/wings. This means that **any particular baby** can have any combination of horns/no-horns and wings/no-wings. It is not possible to say that any particular baby will definitely have a specific combination of horns/no-horns and wings/no-wings. This is different than the **chance** of having a particular combination of phenotypes. In this cross there is a 37.5% chance for an offspring to be horns/wings, a 37.5% chance for an offspring to be horns/no-wings, a 12.5% chance for an offspring to be no-horns/wings, and a 12.5% chance for an offspring to be no-horns/no-wings.

2. In this square, Sandy is heterozygous for both horns and legs (**HhLl**) and Pat is heterozygous for horns and homozygous for legs (**Hhll**) so the cross will be **HhLl** x **Hhll**. Because Pat is homozygous for legs, the Punnett square will only require **two rows** to account for the two different gamete types (**Hl, hl**). Sandy will require **four columns** (**HL, Hl, hL, hl**). This cross results in four different offspring phenotypic possibilities: horns/2-legs, horns/no-legs, no-horns/2-legs and no-horns/no-legs. This means that any particular baby can have any combination of horns/no-horns and 2-legs/no-legs. It is not possible to say that any particular baby will definitely have a specific combination of horns/no-horns and 2-legs/no-legs. This is different than the **chance** of having a particular combination of phenotypes. In this cross: there is a 37.5% chance for an offspring to be horns/2-legs, a 37.5% chance for an offspring to be horns/no-legs, a 12.5% chance for an offspring to be no-horns/2-legs, and a 12.5% chance for an offspring to be no-horns/no-legs. Note that it is not possible for any of the offspring from these two parents to have 4 legs as there is only one available **L** allele.

3. In this square, Sandy is heterozygous for fancy-tail and homozygous for breathing-fire (**Ttff**). Sandy has two X-chromosomes and is, therefore, a male. Thus, he carries two alleles for the fire breathing characteristic. Pat is heterozygous for fancy-tail and contains the allele for not-fire-breathing in her X chromosome (**TtF—**). Her Y chromosome does not contain the Fire gene, which is indicated by the — in the genotype. In this case, Sandy will produce only two gamete genotypes (Tf, tf) while Pat will produce four (**TF, T—, tF, t—**). This means the Punnett square will have 2 columns and 4 rows. This cross results in four different offspring possibilities: male/fancy-tail/no-fire, male/plain-tail/no-fire, female/fancy-tail/fire & female/plain-tail/fire. This means that any particular baby can have any combination of fancy-tail/plain-tail and fire/no-fire. It is not possible to say that any particular baby will definitely have a specific combination of fancy-tail/plain-tail and fire/no-fire. This is different than the **chance** of having a particular combination of phenotypes. In this cross there are many factors to consider. First, there is a 50% chance that a given offspring will be female and a 50% chance that it will be male. Next, there is a 37.5% chance for an offspring to be fancy-tail/no-fire, a 37.5% chance for an offspring to be fancy-tail/fire, a 12.5% chance for an offspring to be plain/no-fire, and a 12.5% chance for an offspring to be plain/fire. When you combine gender, tail and fire, you end up with a 37.5% chance for an offspring to be male/fancy-tail/no-fire, a 37.5% chance for an offspring to be female/fancy-tail/fire, a 12.5% chance for an offspring to be male/plain/no-fire, and a 12.5% chance for an offspring to be female/plain/fire.

Appendix D: Sample Student Understanding Rubric

Section 2B: Standards-Based Scoring Rubric

From Parent to Offspring III: Dihybrid Inheritance I

This assessment looks at your understanding of **Cause-to-Effect** problems in a **Between-Generation** setting. This means you are able to look at a **Cause** (in this case, the dihybrid genotypes of two dragon parents) and figure out its **Effect** (the phenotype of the dragon offspring) for **two generations** (parents & offspring) of dragons.

In addition, this assessment looks at **three main concepts**:

1. **Completing Punnett squares**: filling in the Punnett square by using the parent genotypes to determine the offspring phenotypes. **Questions 1a, 2a & 3a**
2. **Offspring Possibilities:** determining the possibility of a particular offspring phenotype by using parent genotypes in a dihybrid cross. **Question 1b, 2b & 3b**
3. **Offspring Probabilities:** determining the probability (chances) of a particular offspring phenotype by using parent genotypes in a dihybrid cross. **Questions 1c, 2c & 3c**

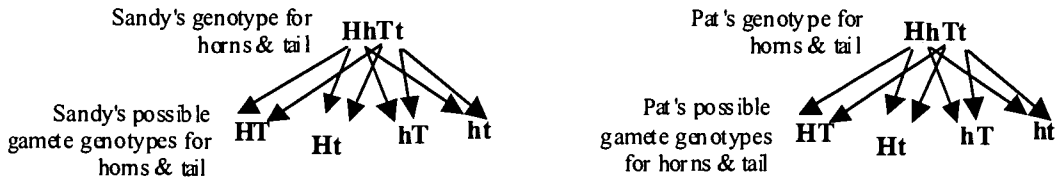
If your understanding of these concepts is ...	You probably should have solved:
EXEMPLARY, you probably understand how to use genotypes of parents to determine the possible genotypes and phenotypes of offspring for most problems in dihybrid inheritance. You were probably able to solve problems in **all three** of the main concepts: • Completing Punnett squares • Offspring Possibilities • Offspring Probabilities	Most Questions
ACCOMPLISHED, you probably understand how to use the genotypes of parents to determine possible genotypes and phenotypes of offspring for some problems in dihybrid inheritance. You were probably able to solve problems in **two** of the three main concepts: • Completing Punnett squares • Offspring Possibilities • Offspring Probabilities	Two of these three groups: Q. 1a, 2a & 3a Q. 1b, 2b & 3b Q. 1c, 2c & 3c
DEVELOPING, you probably understand how to use the genotypes of parents to determine possible genotypes and phenotypes of offspring for a few problems in dihybrid inheritance. You were probably able to solve problems in **one** of the three main concepts: • Completing Punnett squares • Offspring Possibilities • Offspring Probabilities	One of these three groups: Q. 1a, 2a & 3a Q. 1b, 2b & 3b Q. 1c, 2c & 3c
BEGINNING, you are not really able to understand how to use genotypes of parents to determine the possible genotypes and phenotypes of offspring for most problems in dihybrid inheritance. You may have been able to solve **a problem or two** in **one** of the three main concepts: • Completing Punnett squares • Offspring Possibilities • Offspring Probabilities	2 or fewer Questions
UNKNOWN, because you did not answer any questions. You probably don't understand the concepts at all, but it is impossible to tell because you did not even try to guess.	No Answers

Appendix E: Revised Formative Feedback ("Answer Explanation")

Answer Explanation 2B
Dihybrid Inheritance I

1. **Dihybrid inheritance** concerns two characteristics. This problem concerns horns and tail—both autosomal, complete dominance characteristics. Sandy is heterozygous for horns (**Hh**), so he can produce gametes that are **H** or **h.** Likewise he is heterozygous for tail (**Tt**), so he can produce gametes that are **T** or **t.** Pat is also heterozygous for both horns (**Hh**) and tail (**Tt**). Therefore she can produce gametes that are **H** or **h** and **T** or **t.**

   Each gamete produced by Sandy or Pat contains one horns allele and one tail allele. This diagram shows how to figure out all of the possible gamete phenotypes for each.



   Both Sandy and Pat produce four different gamete genotypes (**HT, Ht, hT, ht**). This means the Punnett square looks like this:

Sandy / Pat	**HT**	**Ht**	**hT**	**ht**
HT	**HHTT** horns/ fancy tail	**HHTt** horns/ fancy	**HhTT** horns/ fancy	**HhTt** horns/ fancy
Ht	**HHTt** horns/ fancy tail	**HHtt** horns/ plain	**HhTt** horns/ fancy	**Hhtt** horns/ plain
hT	**HhTT** horns/ fancy	**HhTt** horns/ fancy	**hhTT** no horns/ fancy	**hhTt** no horns/ fancy
ht	**HhTt** horns/ fancy	**Hhtt** horns/ plain	**hhTt** no horns/ fancy	**hhtt** no horns/ plain

   This shows that Sandy and Pat **might** have an offspring with no horns and a plain tail (**1.2**). But you need to do a little work to figure out the chances that this or any other phenotypic combination will occur. Because there are 16 possibilities, each combination represents $1/16^{th}$ of the possible outcomes. Only 3 of the 16 possibilities have no horns and a fancy tail, so the chance is $3/16^{th}$ (**1.3**).

   You should understand that this **DOES NOT MEAN** if they had 16 offspring, that three of them would necessarily have no horns and a fancy tail. It means that there is a $3/16^{th}$ **chance** of a given offspring having no horns and a fancy tail.

2. This dihybrid inheritance problem involves autosomal complete and incomplete dominance. Sandy is heterozygous for both horns and legs (**HhLl**) and Pat is heterozygous for horns and homozygous for legs (**Hhll**). Because Pat is homozygous for legs, the Punnett square will only require **two rows** to account for the two different gamete types (**Hl, hl**). This cross results in two possibilities for each of four different offspring phenotypes: horns/2-legs, horns/no-legs, no-horns/2-legs and no-horns/no-legs. .

Sandy / Pat	HL	Hl	hL	hl
Hl	HHLl horns/ 2 legs	HHll horns/ no legs	HhLl horns/ 2 legs	Hhll horns/ no legs
hl	HhLl horns/ 2 legs	Hhll horns/ no legs	hhLl no horns/ 2 legs	hhll no horns/ no legs

Note that since Pat is homozygous for no legs (**ll**), it is not possible for any of their offspring to have 4 legs. None of the possible offspring possibilities include no-horns/4-legs (**2.2**), while approximately 3/8 include horns/2-legs (**2.3**).

3. This dihybrid inheritance problem involves autosomal and X-linked complete dominance. Sandy is heterozygous for wings and homozygous for breathing-fire (**Wwff**), and can therefore produce only two different gametes (**Wf** and **wf**). Pat, the female (XY) is homozygous recessive for wings and hemizygous dominant (has one X-linked dominant allele) for fire (**wwF—**), and can produce two different gametes (**wF** and **w—**).

Sandy / Pat	**Wf**	**wf**
wF	**WwFf** no wings/ no fire	**wwFf** wings/ no fire
w—	**Wwf–** no wings/ fire	**wwf–** wings/ fire

This cross results in four different offspring possibilities: male/no-wings/no-fire, male/wings/no-fire, female/no-wings/fire & female/wings/fire. So it **is** possible for them to have a baby that has wings and does not breathes fire (**3.2**). Approximately 1/2 of the offspring are male and approximately 1/2 of the males have wings and do not breathe fire so approximately 1/4 [= (1/2) X (1/2)] of the offspring can be male/wings/no-fire (**3.3**). Or, you could just look at how many of the four possible offspring fit that category, 1 of 4. Of male offspring, approximately 1/2 can be wings/no-fire (**3.4**).

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

TM035082

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Design Experimentation with Multiple Perspectives: The GenScope Assessment Project

Author(s): Daniel T. Hickey, Ann C. Kruger, Laura D. Fredricks, Nancy J. Schafer, Steven Zviker

Corporate Source: University of Georgia

Publication Date: April 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: Daniel T. Hickey

Printed Name/Position/Title: Daniel T. Hickey, Asst Prof

Organization/Address: University of Georgia LPSL 30603
611 Aderhold Hall, Athens GA

Telephone: 706 542 3151

FAX: 706 542 4321

E-Mail Address: dhickey@coe.uga.edu

Date: July 1 '03

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com