ED 471 664 TM 034 693

AUTHOR	Hickey, Daniel T.; Kruger, Ann Cale; Fredrick, Laura D.; Schafer, Nancy Jo; Kindfield, Ann C. H.
TITLE	Balancing Formative and Summative Science Assessment Practices: Year One of the GenScope Assessment Project.
SPONS AGENCY	National Science Foundation, Arlington, VA.
REPORT NO	REC-0196225
PUB DATE	2002-04-00
NOTE	26p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).
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; *Science Instruction; Secondary School Teachers; *Summative Evaluation
IDENTIFIERS	*GenScope Computer Program

ABSTRACT

This paper describes the GenScope Assessment Project, a project that is exploring ways of using multimedia computers to teach complex science content, refining sociocultural views of assessment and motivation, and considering different ways of reconciling the differences between these newer views and prior behavioral and cognitive views. The project is refining a computer-supported introductory genetics curriculum and the associated formative and summative assessment practices. Researchers then manipulate motivational conditions in an effort to maximize participation in formative feedback. In the first of three planned annual implementations, four secondary school life-science teachers implemented a month-long curriculum. A within-teacher/between class design contrasted three different assessment conditions: grade-oriented, standards-oriented, and accountability-oriented. Self-report and observational assessments of learning, engagement, and motivation were collected in 16 implementation classrooms and 2 comparison classrooms. The implementation classrooms show significantly larger gains on the new transfer measures and comparable gains on the far transfer measures. In the 26 students whose participation in formative feedback was videotaped and analyzed, engagement was strongly related to learning gains. However, overall engagement in formative feedback was disappointingly modest, and there were no noteworthy effects of the motivational manipulation on learning or motivation. Further modifications to the curriculum and formative assessment practices are being undertaken in response to these findings. Four appendixes contain sample assessments and scoring rubrics. (Contains 5 figures and 50 references.) (Author/SLD)

Balancing Formative and Summative Science Assessment Practices: Year One of the GenScope Assessment Project

Daniel T. Hickey*
Learning & Performance Support Laboratory
University of Georgia, Athens, GA USA

Ann Cale Kruger
Laura D. Fredrick
Nancy Jo Schafer
Georgia State University, Atlanta, GA USA

Ann C. H. Kindfield
Educational Designs Unlimited, Neshanic Station, NJ USA

Abstract

This project is (1) exploring ways of using multimedia computers to teach complex science content, (2) refining sociocultural views of assessment and motivation, and (3) considering different ways of reconciling the differences between these newer sociocultural views and prior behavioral and cognitive views. This paper summarizes these lines of inquiry and describes a project that explores the issues that emerge in their convergence. We are refining computer-supported introductory genetics curriculum and associated formative & summative assessment practices; we are then manipulating motivational conditions in an effort to maximize participation in formative feedback. In the first of three annual implementations, four secondary school life-science teachers implemented a month-long curriculum. A within-teacher/between class design contrasted three different assessment conditions (*grade-oriented, standards-oriented* and *accountability-oriented*). Self-report and observational assessments of learning, engagement, and motivation were collected in sixteen implementation classrooms and in two comparison classrooms. The implementation classrooms showed significantly larger gains on the near transfer measures and comparable gains on the far transfer measures. In the 26 students whose participation in formative feedback was videotaped and analyzed, engagement was strongly related to learning gains. However, overall engagement in formative feedback was disappointingly modest, and there were no noteworthy effects of the motivational manipulation on learning or motivation. Further modifications to the curriculum and formative assessment practices are being undertaken in response to these findings.

Balancing Formative and Summative Science Assessment Practices: Year One of the GenScope Assessment Project

[*] Presented at the annual meeting of the American Educational Research Association, New Orleans, April, 2002. This research is supported by the USA 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 key contributions of many graduate assistants and educators on this project, including Novella Abbott, Bryon Davis, Jessica DeCuir, Bryon Hand, Brandon Kyser, Simona Laprocina, Monkieta Lucas, Morgan Nolan, Marcus Norman, Annette Parrott, John Price, Art Russell, Christine Yang, and Steve Zuiker. 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.

This multifaceted project represents the convergence of three lines of inquiry. One line of inquiry concerns the use of multimedia computer technology to teach introductory genetics. The second line concerns the pragmatic implications of contemporary sociocultural views of assessment and motivation, particularly as they relate to motivating students to participate in formative assessment. The third involves the complex theoretical issues that emerge when attempting to reconcile the relationship between these newer sociocultural views and prior behavioral and cognitive views. Following is a summary of each of these lines and the research goals associated with each.

Curricular Inquiry.

The first line of inquiry concerns the use of 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. These and other 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; 2000). 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.

Our primary curricular goal is refining the formative assessment tools developed in previous research. Our efforts to meet this goal draw directly from the emerging consensus around the value of formative assessment (e.g., Black & Wiliam, 1998, Gipps, 1999; Graue, 1993, Turnstall & Gipps, 1996). This goal is in essence a search for modest, scaleable assessment practices that motivate learners to engage in effective "assessment conversations" (Duschl & Gitomer, 1997). Such activity promises to dramatically enhance student learning. This admittedly idealized activity is characterized by authentic scientific argumentation (e.g., Driver, Newton, & Osborne, 2000) in which 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). While not central to the project goals, we are also devoting substantial resources to refining the curricular activities that structure the student and teacher activities around the software.

Pragmatic Inquiry

Another line of inquiry underlying the present investigation concerns 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). As outlined in Hickey (1997) three articles published in 1989 provided key ideas for this line of inquiry. The first of these articles is Fredriksen & Collins, in which they present the notion of *systemic validity,* advancing a fundamental reconceptualization of assessment:

> 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 second article was Collins, Brown, and Newman's (1989) *cognitive apprenticeship* paper. This landmark paper set out the practical implications of situative/sociocultural theories for contemporary instructional practice. In doing so it highlighted how new views of learning and instruction provided a framework for rethinking motivational issues; this complex relationship to assessment was often acknowledged but seldom investigated. While acknowledging that competition (an intentional and unintentional byproduct of many forms of assessment) "raises difficult emotional issues for many students" (p. 490), they argued that many negative effects of the competition are actually caused by inadequate instruction:

> Under many forms of teaching, students lack the means, in the form of an understanding of the underlying processes, strategies, and heuristics involved in solving problems, for improving their performance. In these cases, the motivation to improve that might be engendered by competition is blocked, leaving students frustrated and discouraged (p. 490).

With clear links to the systemic validity issues, Collins, Brown, & Newman argued "For competition to be effective, comparisons must be made not between the products of students problem solving, but between the processes…" (p. 490). One of the objectives of the present study is considering whether the well-documented negative consequences of competition on motivation and learning are minimized or reversed in systemically valid learning environments.

While it is exciting to think that systemically valid assessment practices might reverse the negative effects of competition, a third publication hinted at the challenges facing investigations of this issue. Berieter & Scardamalia's (1989) chapter advanced an *intentional learning* framework that was consistent with the cognitive apprenticeship model, and provided further consideration of the motivational issues these new approaches presented. Precluding ready application of prevailing models of motivation, Berieter & Scardamalia concluded "the conventional distinction between intrinsic and extrinsic motivation is too crude to be of much service in studying the intentional aspects of learning…" (1989, p. 366). One of the longer-term goals of this study is illustrating and overcoming the inadequacy of prevailing cognitive/rationalist models for addressing the motivational issues that emerge when considering assessment from situative/socio-constructivist perspectives. Specifically, we are attempting to further develop the situative/sociocultural model of motivation outlined in Hickey and McCaslin (2001), and McCaslin & Hickey (2001a).

Overall, this line of inquiry is attempting to provide a scientific basis for understanding the consequences of increasingly common accountability-oriented educational practices. Modern motivation theory suggests that such practices will have generally lead to *decreased* engagement in learning and lower achievement overall. As summarized by Kelleghan, Madaus, & Raczak, (1996), there is abundant empirical evidence that the presumably-unmotivated lower-achieving students will engage in ego-protective task disengagement when facing increased standards without increased resources. Meanwhile behaviorists (e.g., Cameron & Pierce, 1994) analyze much of the same body of research to reach a very different conclusion.

This prior research was based on *acquisitory* views of knowing and learning, and therefore educational practice. Our study is exploring motivational practices in a manner that is more consistent with newer *participatory* view of knowing, learning, and practice.

Theoretical Inquiry

The third line of inquiry concerns the complex tensions between the situative/sociocultural views of knowing and learning and prior behavioral/empiricist views and cognitive/rationalist views. These goals are perhaps best understood as an application of the ideas set forth by Greeno (e.g., Greeno et al, 1998; Greeno, Collins, & Resnick, 1996) to the tensions raised by the pragmatic goals relative to the prevailing empiricist models of assessment and rationalist models of motivation. To this end, our project is studying learning and motivation from each of these three perspectives. We are then comparing different ways of reconciling the tensions that are manifested in the inevitably contradictory conclusions that follow.

As outlined in some detail in Hickey and McCaslin (2001) and Hickey (in review, forthcoming), the "levels of aggregation" approach to reconcile these tensions that follows logically from the earlier views is problematic. A fundamental goal of the present study is to consider the value of a "dialectical" reconciliation. Following quite directly from Greeno et al. (1998), this approach treats 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 builds on the strengths and weakness of the prior approaches, therefore subsuming them. A dialectical approach may help advance that seemingly intractable debate over competition and extrinsic rewards. These unresolved issues have allowed policy makers to ignore relevant educational research while driving dramatic changes in educational practice via accountability-oriented reforms.

METHODS

Our investigation features rigorous outcome measures, sophisticated statistical analyses, comparison groups, complex quasi-experimental designs, and other characteristics of conventional empirical studies of classroom learning. However, the investigation is perhaps better understood as a parallel series of "design experiments" (i.e., Brown, 1992, Collins, 1999) around modest variations in classroom assessment practices. While the study is designed to yield many relevant empirical outcomes, we are also exploring more fundamental interpretive questions that transcend conventional methods of conducting and describing educational research.

Participants

Central to our investigation is the key role of classroom teachers. The graduate student managing the project and key members of the curriculum and assessment development teams are current or former classroom teachers. Furthermore, two of the implementation teachers are currently working on the project as graduate assistants or curriculum consultant. Their input and insights are central to the goal of studying practices that are readily scalable. This is particularly important given that we are exploring the common-but-controversial practice of publicly acknowledging individual proficiency. Among the teachers working on this project, we found the entire range of opinions about this practice. Understanding their reasons for and against is invaluable for advancing practices that are likely to be appropriated by other educators.

Implementation Teachers and Students

In the 2000 implementation, four high school teachers with a total of 18 ninth-grade life sciences classes participated as implementation teachers. Two teachers were at a 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. The first teacher implemented GenScope in one honors life-science and three AC (adjusted-curriculum)[1] life science courses. Students in the three AC courses participated in the videotaped aspects of the study. The second teacher implemented GenScope in three AC life-science courses and one honors life science course.

The third teacher implemented GenScope in two AC life sciences classrooms at 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. The fourth teacher implemented GenScope at 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. This teacher implemented GenScope in three honors life science courses and three non-AC regular life science courses. All three non-AC regular courses participated in the videotaped aspects of the study.

A second teacher at the high-SES school participated as a comparison teacher by administering our genetics proficiency assessment before and after genetics instruction in two of her non-AC regular life science classes. This teacher used the fairly typically district-approved genetics curricula to teach genetics and reported spending the same number of days on genetics as the GenScope teacher. Both teachers had less than five years teaching experience. The comparison teacher (but not the GenScope teacher) had a graduate science education degree, and was familiar with the GenScope curriculum because she had completed her graduate teaching practicum in a classroom that had participated in the prior GenScope research.

All but the second teacher at the low SES school were African American, all had 3-8 years experience, and all were credentialed science teachers with undergraduate life science degrees. The teacher at the middle-SES school was a doctoral student in science education who also participated in the curriculum development effort as a graduate research assistant; the other three GenScope teachers did not have graduate degrees, were recruited with the assistance of the district science coordinator, and were paid a non-trivial honorarium for their participation.

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 by the USA National Research Council. In this report, Donovan, Pellegrino, & Bransford (1999) argue that assessment and instruction should more reflect what is known about the

[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.

[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).

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, submitted) 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.

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, 2000, in submission)

Formative Assessment Unit Tests and 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 the assessment team during a 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.

[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/

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 and preparations for the implementation were underway, debate emerged within the project over the difficulty and complexity of the unit assessments, key points, and answer explanations, and understanding rubrics. On one hand, the content expert/curriculum writers (Kindfield and graduate assistant Morgan Nolan) argued (1) that these were 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, and (3) that teachers and students had reportedly found earlier versions of these materials very useful. In support of this position, the principle investigator (Hickey) argued (1) that the formative assessment context would provide scaffolding to help students use and comprehend the materials, (2) that the challenge would highlight difference in motivation to learn the material, and (3) that the earlier versions of the materials had been most effective in the lowest achieving classrooms. On the other hand, former or current classroom teachers on the project (Project manager Schafer, implementation teacher Annette Parrott and graduate assistant Brian Davis) argued that many ninth-graders would still be unwilling and/or unable to use materials as written. In support of this position, the head of the Behavioral Analysis team (Fredrick) argued that it was unethical to present materials to students that so obviously exceeded their reading grade level, particularly given the 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 for this round, and to further address this issue after the first implementation.

Independent Variable (Feedback Manipulation)

We manipulated the feedback conditions in order to explore the consequences of accountability-oriented practices in a more systemically valid assessment environment. At the end of each of the four units, students completed the unit tests, and were informed that their grade for the genetics portion of the life science class would be based on how well they did on each of unit tests. While the teachers held the curricular activities and the completion of the unit tests constant across their similar classes (e.g., three AC life science), a within-teacher design had them implement three different formative assessment models. With the some help from a research assistant, teachers marked the incorrect items on the unit tests. The manipulation that was concerned with the way student performance was characterized, was as follows:

> In the *grade-oriented* classrooms, unit tests were marked with the percentage of items answered correctly, as is convention.
>
> In the *standards-oriented* classrooms, 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.
>
> In the *accountability-oriented* classrooms, performance was characterized just as in the standards-oriented classroom. But we also attempted to induce performance goal orientation by imposing a

salient extrinsic recognition of individual proficiency. We asked teachers to post the names of students in the *exemplary* or *accomplished* categories on each of the parts of each of the unit tests.

On the first day of the next unit, the marked unit tests, along with the answer key, answer explanation, and scoring rubric (in the two conditions) were returned to students after they had assembled into their groups at the computer. Teachers were asked to instruct the students to review their unit tests before beginning the computer activities for the following unit. Our assumption was that high levels of motivation would be needed for students to freely engage in the challenging practice of reviewing their unit tests—particularly when facing a relatively attractive alternative task of working on a computer activity.

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/socioconstructivist views of knowing and learning (as outlined in Hickey & McCaslin, 2001; also Greeno, Collins, & Resnick, 1996; and Case, 1996).

Engagement Measures.

Cognitive/Rationalist motivational experiences survey. Reflecting the prevailing cognitive/rationalist approach to motivation, one of the measures of engagement was a context-specific self-report of motivational goal orientation. Reflecting the mainstream assumptions that meaningful learning involves intrinsic sense-making processes, this measure follows from 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). The 24-item *motivational experiences survey* was adopted from the one used by Hickey, Moore, & Pellegrino (2001), and asked students to rate the strength of different kinds of goals during biology class that day. It was administered at the end of a class period during the first and third GenScope computer activities; Table 2 lists the specific scales and the internal consistencies obtained in the present administration.

Videotape recording. The other two engagement measures were both based on analyses of videotape recording of classroom activity. We recorded 120 hours of videotape, capturing four (of the twenty) class periods for five groups of students in six classes. We simultaneously recorded five groups of students during a GenScope computer activity and a feedback session during the first and third units. Three of the six classrooms (one from each feedback condition) were regular (non-AC) life science classes taught by the GenScope teacher at the high SES school; the other three were the AC life-science classes taught by the first teacher at the low-SES school.

Initial observations revealed few students used feedback information—regardless of feedback condition- and the low-SES teachers did not fully implement the grade-posting aspect of the accountability-oriented condition. Given these observations and the fact that this was essentially a pilot year, 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.

The two teams analyzed only activity around students' use of test feedback. This means that coding stopped when all of the students in the group turned their attention away from their graded/scored unit tests to begin the first GenScope computer activity for the next unit. Because the unit tests, standards based scoring rubric, answer key, and answer explanation sheet were each copied onto paper of different colors, it was usually possible to follow what the students were engaged with. An omnidirectional microphone was placed on the table next to students to record conversation. Coupled with the close quarters and the low ceiling in the low SES students' portable classroom, we were left with mediocre audio.

The videotape was analyzed by two teams who were asked to devise an interpretive methodology 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.

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 teams 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/socio-constructivist 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, we intended for this team to analyze discourse patterns within the triads of students, searching for emerging ritualized use of desirable knowledge practices associated with introductory genetics. However, the 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.

The feedback sessions on each of the 40 videotapes were analyzed by tracking the shifts in social interaction of the three individuals and coding the resulting segments.

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.

Situative/socio-constructivist analysis of transferable knowledge practices. The inherently contextualist world-view of situative/socio-constructivist perspectives precludes the conventional 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 complicated 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 initial finding described below, we choose not to pursue this analysis in the current implementation.

Cognitive/rationalist analysis of transferable knowledge structures. 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" measure described next[4].

Behavioral/empiricist analysis of transferable associations. Implicit in conventional multiple choice assessment practices is the assumption that knowledge consists of cognitive or behavioral representations of

[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, 2001, and Hickey & Holbrook, 2000

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).

RESULTS

Initial observations revealed that few students were engaging in the formative assessment feedback materials, regardless of the feedback condition. This was not entirely surprising, as our teacher-research assistants had expressed concern that the answer explanations were written at too high of a level and that the students had little prior experience using feedback materials to do more than simply verify correct answers. However, because of the scope of the implementation and the substantial investment of time and money in producing curriculum materials for every student in the 20 classrooms, we were unable to make any modifications that might have addressed the problem.

Engagement

The three measures of engagement confirmed that there was disappointingly little use of feedback over all and little difference between the conditions. The videos also confirmed our worries that students would find the feedback materials confusing and frustrating. There were multiple incidents observed on the tapes where 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.

Self-reported goal orientation. The five scales on the *motivational experience survey* were shown to be sufficiently reliable. Table 2 shows that the internal consistencies for the scales ranged from .74 to .86. Means for the different scales appeared to be unrelated to feedback condition, with no statistically significant changes across time (the condition by time interaction) or main effect of condition at either time.

There were main effects of SES, with students at the low-SES school (who were all African American) reporting significantly lower learning orientation, individual cognitive activity, and perceived relevance than the students at the other two (largely non-minority) schools. However main effects of race on such self-report measures are common (they were found with similar measures in studies conducted with elementary schools in the same communities, as reported in Hickey, Moore, & Pellegrino, 2001) and their source and meaning remains a subject of debate (see Graham, 1994). More importantly, because this variable did not interact significantly with time, they appear to be fairly stable individual differences that are not influenced by instruction. Similarly there were main effects of class type, with honors students reporting significantly lower learning orientation and significantly higher perceived relevance. There were also significant interactions with time, with honors students reporting relative increases on all five scales, with most of the interactions statistically significant. Given the overall engagement results, none of the differences appeared worthy of further consideration.

Engagement behavior. For the 84 students whose behavior during the first and the third feedback activity could be scored, the mean time engaged in any behavior involving the unit test was just 157 *seconds*, with a mean of 161 seconds for the 39 students in the grade-oriented condition and 155 seconds for the 45 students in the standards-oriented condition. There was a substantial difference between the two teachers, with a mean of 188 seconds for the 42 high-SES students and 108 seconds for the 32 low-SES students. 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. Encouragingly, almost all of this activity occurred during the third unit, indicating that some engagement increased across time. What little substantive engagement that occurred during unit three did not appear related to the feedback condition, with none observed in the high SES standards-oriented classroom, and none observed in the low-SES grade-oriented classroom.

Sociocultural analysis. The results of the sociocultural coding largely confirmed the findings described 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. While the students in the graded condition were significantly more focused on grades than students in the standards-based condition, this did not appear to yield increased content-oriented engagement. Indeed, on-task/content oriented activity was so infrequent the team found only few brief incidences of it.

Learning

Scores on the NewWorm and the multiple-choice test were scaled separately using *Facets* (Linacre, 1989). This Rasch method makes it possible to (1) directly compare gains 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 for each of the five teachers, broken out by feedback condition. Gains ranged from 5.7 to 15.4, roughly ½ to 1½ standard deviations. We see that there is no systematic effect of feedback condition. In fact, gains within teacher are remarkably consistent across conditions. Given that the NewWorm represents a near-transfer measure of the knowledge assessed in the unit tests, these findings confirm our initial observations that there was little difference in the way students used feedback across conditions.

These gains are in the same range as those observed in 40 GenScope and comparison classrooms between 1996 and 1999. (The largest gain observed so far was from 22.4 to 53.5, over three standard deviations, under relatively unique conditions in a single classroom). 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.

Multiple choice (far-transfer) proficiency gains. Figure 3 shows proficiency scores on the multiple-choice items before and after instruction. As expected, the gains on the far transfer tests were somewhat smaller than the gains on the near transfer performance assessment. The gains across the high-SES classes were about 7.0, while the gains across the two medium-SES classes were 4.3. However, the gains in the two sets of low SES classrooms were somewhat surprising. The gains in the second low-SES teacher's 4 classes were consistent and tiny, just 0.21; the changes in the first teacher's classes went in dramatically different directions. Examination of the data revealed no obvious explanations. But given that the between-class manipulation to feedback conditions did not appear to have been meaningfully enacted, further interpretation of these effects appears fruitless.

Within-school comparison. We were only able to obtain comparison data at the high-SES school, where two regular biology classrooms completed the learning assessments before and after textbook based genetics instruction. The GenScope teacher and the comparison teacher confirmed that the introductory genetics curriculum in the comparison classroom was quite similar to what had been replaced by the GenScope curriculum. Figure 4 shows that the students in the two comparison classrooms (both regular biology) recorded a relatively small gain of 6.9 on the NewWorm. This was about half the 12.8 gain in the matched regular biology GenScope classroom [F (1,111) = 2.87, p = .093], and smaller still than the honors GenScope students at this school (15.6).

Needless to say, because the GenScope unit assessments give GenScope students experience in solving problems similar to those on the NewWorm, the NewWorm assessment is biased against the comparison students. This is not the case for the nine-item multiple-choice test. Figure 5 shows that the gain in the 3 regular biology GenScope classrooms (8.6) were slightly lower than the gain in the matched regular biology classrooms (7.1), but the difference was not statistically relevant ($F < 1$). While the honors GenScope students showed substantially larger gains on the far transfer measures (14.7), these groups are not directly comparable. It is worth noting however that a gain of 1.5 SD on a far transfer measure represents a particularly noteworthy accomplishment in this particular implementation.

Engagement in Feedback and Learning Outcomes.

The apparently limited motivation to engage in the feedback led to very modest engagement in that activity across all conditions. This lack of systematic difference due to the manipulation precludes any causal interpretation of the consequences of the quality or quantity of participation in that activity for learning outcomes. Among the 26 students for whom the videotapes allowed us to code behavioral engagement in feedback during Unit One and Unit Three, we observed a range of time engaged. We summed the number of 5 second intervals where the student's activity was coded as behaviorally engaged in the feedback activity (i.e., reading the feedback materials, reviewing their answers, discussing their answers in their group, etc.). One student was engaged for 82 intervals (6.8 minutes) and one students was not observed engaged at all, and the rest were fairly evenly distributed (mean = 52, SD = 9.2). The partial correlation of engagement in feedback with posttest score on the near transfer NewWorm assessment (partialing out pretest score) 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 the intentional correspondence between the two. However, engagement in feedback was not significantly correlated to scores on the far-transfer multiple-choice assessment.

Results Summary

While the implementation of the GenScope curriculum went reasonably well, initial observations suggested that our intended manipulation of assessment practice was not realized in the students' behavior. Students made very little use of feedback information—regardless of condition. Confirming the worries of several members of the research team, the formative feedback materials appeared entirely too challenging for many of the students. The teachers confirmed that students did not have enough practice using these kinds of materials to engage with them meaningfully in small groups. Exacerbating this problem, the feedback materials and answer explanations were written at too advanced a level to be used by students under these circumstances. Project teachers and researchers had raised these concerns, but the press of preparation for four teachers and twenty classrooms made it impossible to realize their suggested modifications in time for the first-year's implementation.

CONCLUSIONS

The primary conclusion from this first pilot year is that further modifications are needed to support worthwhile assessment conversation within our curriculum. The feedback manipulation will be completely revised, as follows. In response to teachers expressing that they felt "disconnected" from the GenScope computer activities, we have modified the curriculum for the Fall 2002 implementation to allow for more teacher scaffolding of computer activities and assessment feedback sessions. Nine of the eighteen GenScope computer activities will be completed as whole class activities, with the teacher using an LCD panel attached to a computer. The readability of the curriculum and the feedback materials will also be adjusted to meet students' ability. Finally, we have revised the standards-oriented feedback sessions to reflect more of a formative approach to assessment.

With these improvements, especially using the teacher as the leader of meaningful discussion of content at the time of feedback, we expect to improve learning outcomes overall and to have more opportunities to explore the consequences of our assessment manipulation for engagement and learning.

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).

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.

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.

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.

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.

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.

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. (submitted). Integrating curriculum, instruction, assessment, and evaluation in a technology-supported genetics learning environment. Manuscript submitted to *American Educational Research Journal,* March 2002.

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.

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.

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.

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

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.

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.

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.

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.

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

Table 2. Scales on Motivational Experiences Survey

Scale	# Items	α[a]	Example Item[b]
Mastery/Learning Orientation	5	.86	my main goal was learning as much as I could
Performance/Ego Orientation	5	.86	I tried to make others think I did a good job.
Perceived Relevance	4	.85	I felt like I was learning something useful.
Individual Cognitive Activity	5	.78	I always went back over the things I did not understand.
Distributed Cognitive Activity	5	.74	We relied on each other to understand what we were learning.

[a] Internal consistency (Cronbach's Alpha) at the second adminstration

[b] All items were preceded with the stem *In Biology class today...* All items were scored *strongly disagree, agree, both agree and disagree, agree,* or *strongly agree.*

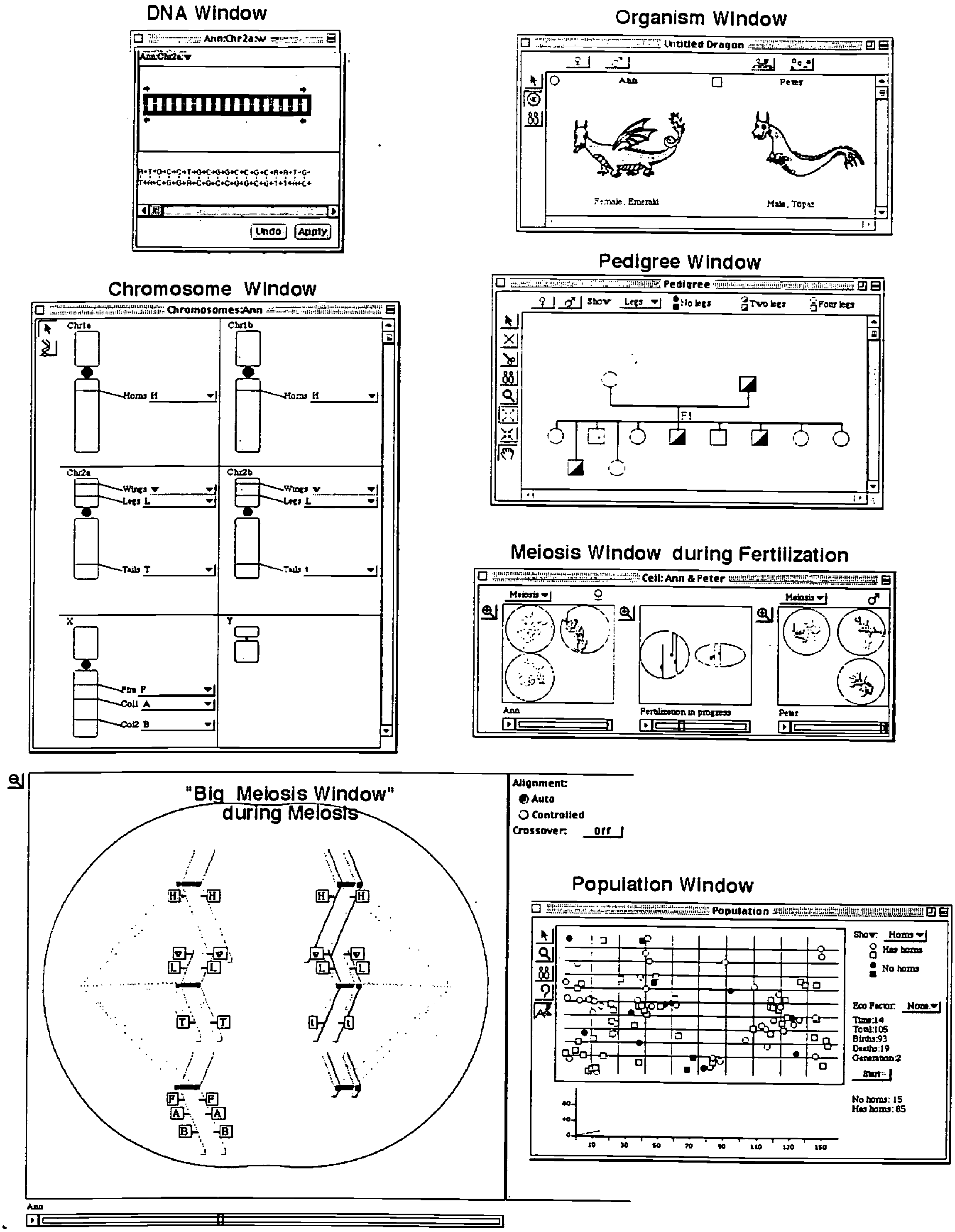


Figure 1. Examples of Screens in the *GenScope* Software.

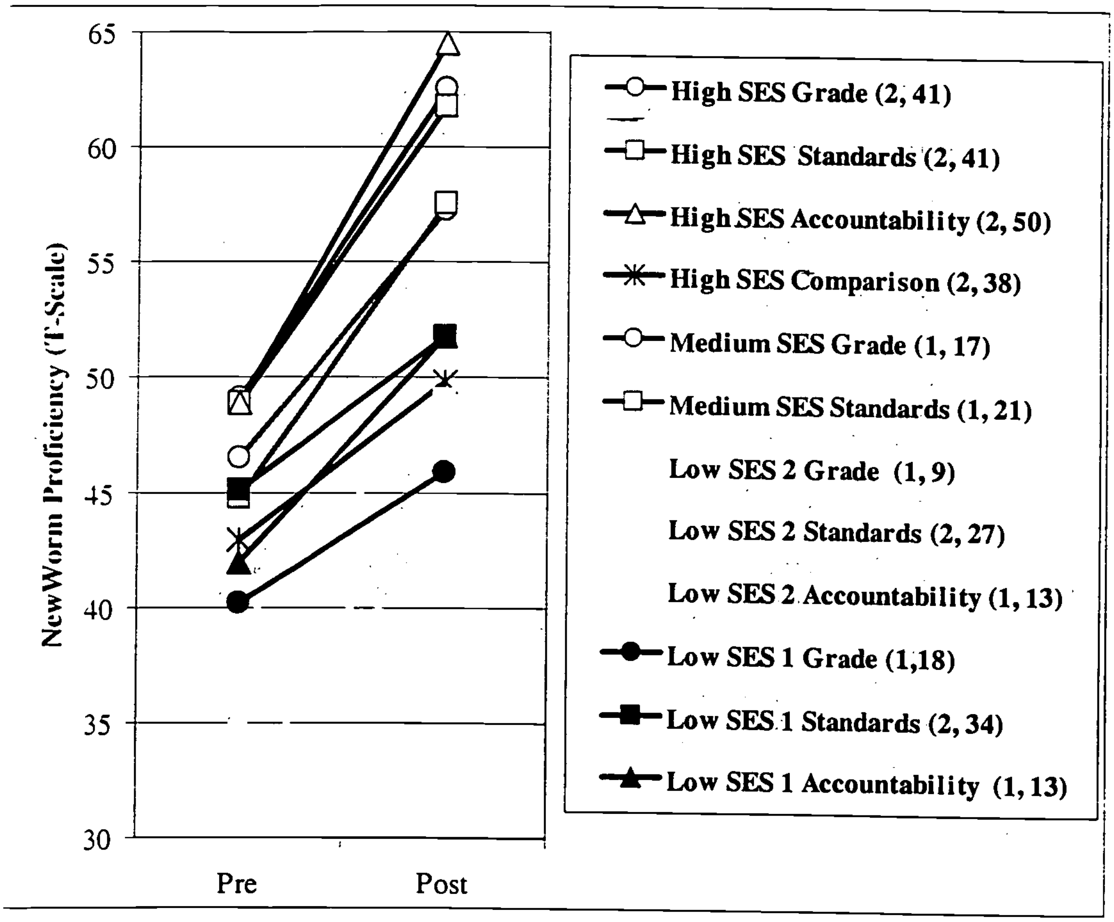


Figure 2. Proficiency gains on *NewWorm* "near-transfer" performance assessment.

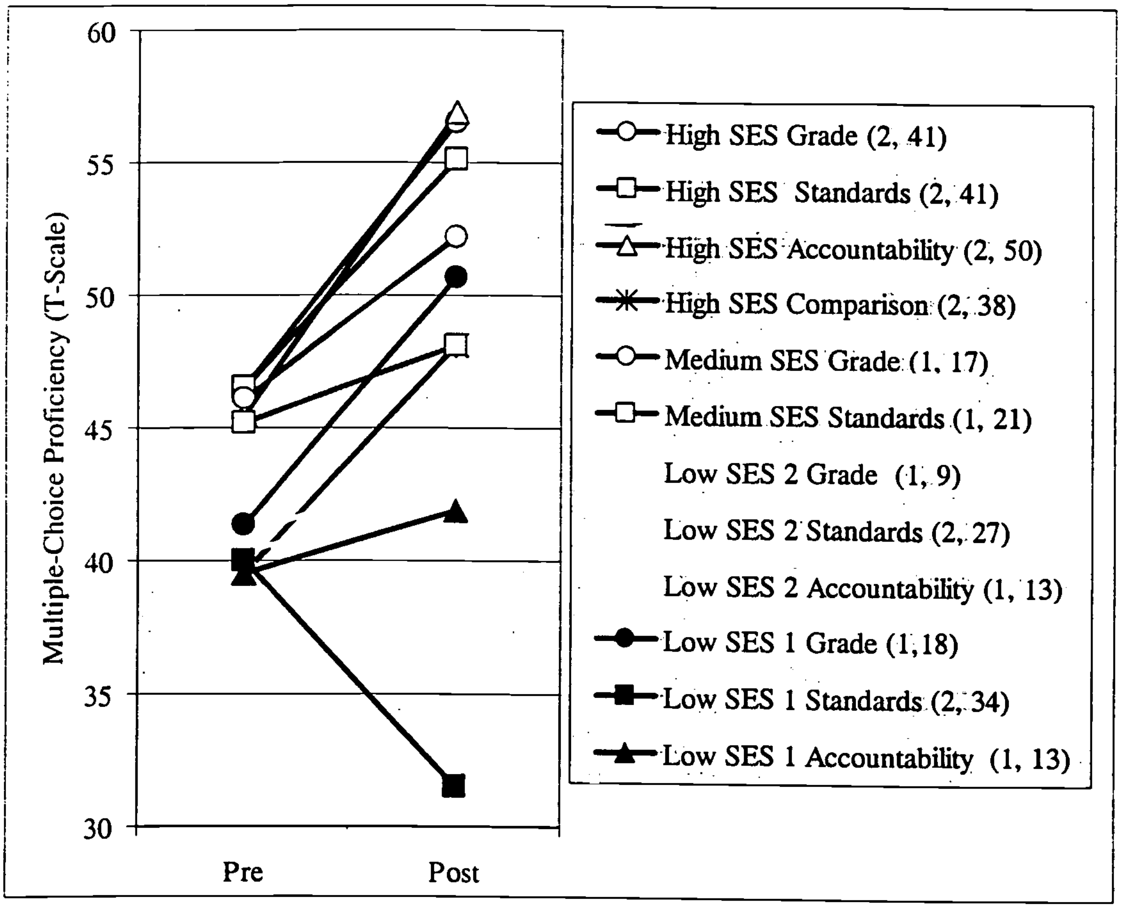


Figure 3. Proficiency gains on Multiple Choice "far-transfer" test.

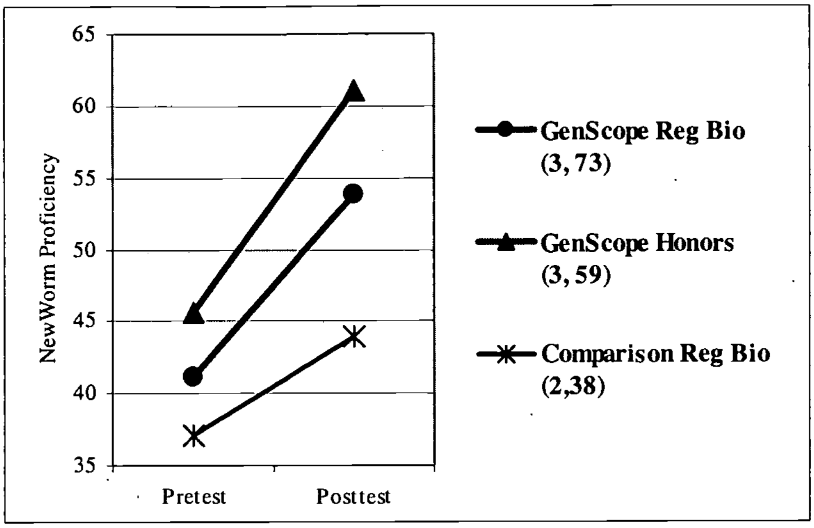


Figure 4. Near-transfer gains in within-school comparison substudy.

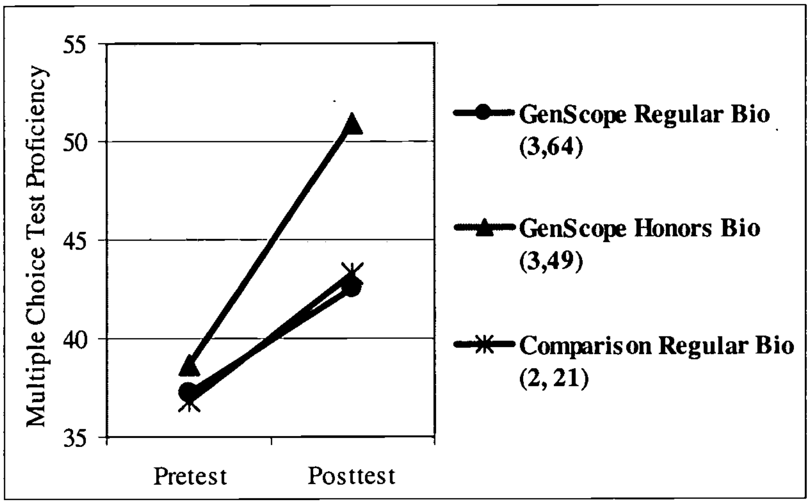


Figure 5. Far-transfer gains in within-school substudy

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)

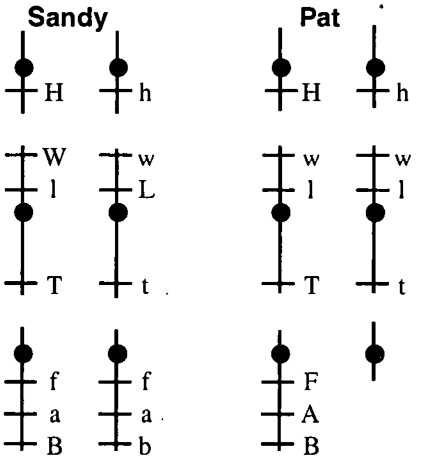
Tail: Fancy-tail dominant to plain-tail.

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**.

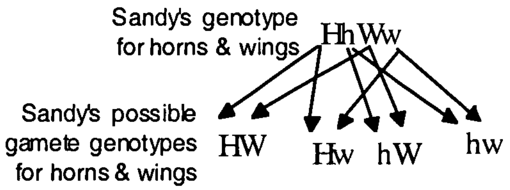
As with monohybrid inheritance, in dihybrid inheritance a **Punnett square** is used to determine offspring possibilities and probabilities. The Punnett square is a tool that helps you keep track of the gametes that each parent can produce and the possible ways to combine the gametes from each parent 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

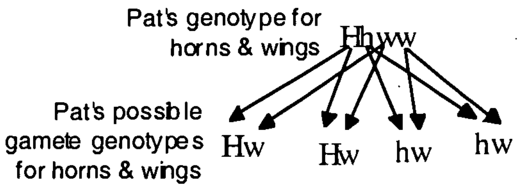
There are 11 steps to determining the genotypes of offspring in dihybrid inheritance using a **Punnett square**.

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.


Sandy's genotype for horns & wings
HhWw
Sandy's possible gamete genotypes for horns & wings
HW
Hw
hW
hw


Pat's genotype for horns & wings
Hhww
Pat's possible gamete genotypes for horns & wings
Hw
Hw
hw
hw

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

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

ERIC
Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM034693

I. DOCUMENT IDENTIFICATION:

Title: Balancing Formative and Summative Science Assessment Practices: Year One of The GenScope Assessment Project	
Author(s): Daniel T. Hickey, Ann Kruger, Laura Fredrick, Nancy Schafer, Ann Kindfield	
Corporate Source: University of Georgia Learning and Performance Support Laboratory	Publication Date: April 2002

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	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	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	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☒	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	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	Printed Name/Position/Title: Daniel T. Hickey / Asst Professor	
Organization/Address: University of Georgia LPSL 611 Aderhold Hall Athens GA 30602	Telephone: 7065423157	FAX: 7065424321
	E-Mail Address: dhickey@coe.uga.edu	Date: Dec 10 02

(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://ericfacility.org