The Secretary's Conference on Educational Technology 2000

Developing Assessments for Tomorrow's Classrooms

Barbara Means, Bill Penuel, and Edys Quellmalz
Center for Technology in Learning
SRI International

In an inner-city high school physics class in Chicago, students are examining computer images captured by automated telescopes. To identify the types of galaxies represented in these images, pairs of students use software tools to enhance the images on their computer screens so that patterns are easier to detect. They change the colors and brightness of the images, zoom in to look at specific features, and zoom out to get an impression of overall shape; they rotate an image to see it from multiple perspectives. These student activities are part of a technology-supported project called Hands-On Universe. Automated telescopes now capture many more images from outer space than professional astronomers have time to analyze. Developed at UC Berkeley's Lawrence Berkeley Lab with support from TERC, the Hands-On Universe project involves students in reviewing images from space. In the course of these activities, students learn basic concepts and skills of research astronomy and help search for super novas and asteroids. (Two Hands-On Universe student groups have in fact discovered previously unknown super novas and had their work published in scientific journals.) Hands-On Universe enables students to use the same kinds of software tools that scientists use (albeit with more user-friendly interfaces) to examine and classify the downloaded images.

Technology-Supported Activities to Support Meaningful Learning

How People Learn, a recent report from the National Research Council (Bransford, Brown, & Cocking, 1999), applies principles derived from research on human learning to issues of education. The report explores the potential of technology to provide the conditions research indicates are conducive to meaningful learning. The report illustrates how technology can be used to help supply five key conditions for learning:

- real-world contexts for learning
- connections to outside experts
- visualization and analysis tools
- scaffolds for problem solving
- opportunities for feedback, reflection, and revision.

Hands-On Universe exemplifies many of these features, as do other technology-supported interventions such as GLOBE, UC Berkeley's WISE, the Quests from Classroom Connect, Vanderbilt University's Scientists in

Action, and many of the projects you will see exhibited here over the next two days.

Anyone who has reviewed items from the kinds of standardized tests given in most states will note the striking contrast between the kinds of activities students undertake in these projects and the content of the test items (Popham, 1999). This becomes a problem for classroom practice not only because teachers may feel anxious about devoting precious instructional minutes to technology-based activities that are not preparing students to do well on mandated multiple-choice tests but also because teacher-produced tests and other assessment practices are so strongly influenced by conventional practice in large-scale assessment.

Teachers (and our experience suggests, university faculty as well) tend to think in terms of multiple-choice and short-answer test items that put a premium on learning definitions for new terms, memorizing numbers, and distinguishing correct statements of facts or relationship from plausible-sounding distractors. The kinds of complex investigations, deeper understanding, and ability to apply concepts to new situations fostered by technology-supported programs like Hands-On Universe are difficult to capture with conventional test formats.

In many cases, teachers are attracted to approaches that actively engage students and hold promise for enhancing learning with understanding. Lacking familiarity with ways to test deeper understandings or higher-order skills, however, teachers often implement the activity without assessing what students are learning from it. The drawbacks of this omission are two-fold. First, students may not be acquiring the kinds of understanding the activity is intended to promote, and with no assessment of their level of understanding, the teacher is unaware of what they do not know. Barron et al. (1997) examined students' learning when their classrooms collected data on the quality of the water in a local stream. They found that when the teacher did not conduct assessment activities during the course of the water quality project, students went through the motions without understanding basic concepts. In an examination of classroom implementations of the inquiry-oriented Global Lab Curriculum, Young et al. (1998) found that many of the teachers who had students work in small groups to conduct the Global Lab investigations assigned only participation scores for that part of the class. Science grades were based on tests of the factual content in the textbook rather than on what students did in the course of investigations of the air, land, and water in their Global Lab study site. When this happens, students are receiving an implicit message about what is important–that which is graded–and the lack of teacher evaluation for their inquiry work suggests that it is "fun" rather than substance.

Lack of availability of assessments for higher-order and inquiry-oriented activities is a problem for researchers and evaluators as well. What is a project evaluator to do with a Technology Innovation Challenge Grant whose mission is to "encourage student inquiry" or "teach students to be change agents"? Scores on the Stanford Achievement Test (SAT-9) are unlikely to be affected, at least in the short run, by experiences supporting these goals. Thus, researchers and classroom teachers have a common stake in the development of assessment techniques and instruments more appropriate for the kinds of student-centered, inquiry-oriented teaching and learning we hope to support with technology.

Planning for a New Research Agenda

Despite the major investment of federal, state, local, and private funds in school technology, questions remain both about technology's impact on student learning and achievement and about how to implement technology within schools to maximize the learning benefits. The President's Committee of Advisors on Science and Technology (PCAST, 1997) called for an ongoing federally supported research program with complementary studies conducted by dozens of research organizations to provide "rigorous, well-controlled, peer-reviewed, large-scale empirical studies to determine which [technology-supported] educational approaches are in fact most effective in practice."

In 1999 SRI International received a grant from the U.S. Department of Education to support planning for a major program of rigorous, systematic educational technology research. In addition to commissioning research design papers from leading experts in research methodology, assessment, and learning technology, we have been involved in developing prototype technology-based assessments to help address the dearth of appropriate student learning measures available to inquiry-oriented, technology-supported projects. The remainder of this paper is a description of these assessment prototypes and of what we are learning from trying them out in classrooms.

We wanted our prototype assessments to capture skills that are not easily assessed with more conventional standardized tests and to demonstrate the capabilities provided by technology for doing more flexible, in-depth assessments. This is very much a look at work-in-progress as we are just half-way through the development and piloting process. Last year we did the initial design, development, and pilot work on the two technology-based assessments. Based on this early work, we are in the process of revising the assessments for further field testing this fall. While our assessments tap skills intentionally selected for their wide applicability, we have structured the two assessment prototypes as "templates," which will be modifiable to support the development of additional assessment tasks exemplifying the same approach.

Internet Research Task

Although word processing remains the most common application of technology in U.S. schools, Internet research is probably the fastest growing. On the national survey conducted by Becker and Anderson in the spring of 1998, 30% of teachers (and over 70% of those with direct high-speed Internet connections in their own classrooms) said that they had assigned Internet research tasks that school year (Becker, 1999). Given the fact that the proportion of U.S. classrooms with Internet connections rose from 51% to 63% between 1998 and 1999 (based on NCES statistics), we can extrapolate a commensurate increase in the frequency with which students' teachers are asking them to perform Internet research.

But what do students get out of these on-line research activities? And how do we know what they are learning? Our observations of classes conducting Internet research suggests that the nature of the research assignments, student skill requirements, and grading criteria vary markedly from class to class. We have seen classrooms engaged in long-term problem solving where the students figure out how to frame a problem, decide that they need a certain kind of data to support their problem solving, identify on-line sources of relevant data, analyze the quality and relevance of alternative data sources, and then pull down data sets for analysis.

We have also seen classes where the Internet task is more on the order of "find five facts about this country." Sometimes any fact and any source will do. Other times the task is so constrained, with teachers providing a small set of URLs and asking for fill-in-the blank type information, that students have little opportunity to exercise skills other than typing and copy-and-paste functions. Students may get graded on how many sites they accessed rather than on the judicious choice of information sources or important information.

The products students are asked to produce based on their Internet research are equally various. They run the gamut from lists of facts to conventional term papers to student's own interactive multimedia presentations or Web sites.

Standards-setting bodies—both those concerned with technology per se and those dealing with academic content areas—place an emphasis on research and communication skills. The standards of the National Council of Teachers of Mathematics (NCTM), for example, start with four process skills important at every

grade level: problem solving, communication, reasoning, and connections (linking different subfields of mathematics and linking mathematics to other disciplines and real-world problems). The *Benchmarks for Science Literacy* (AAAS, 1993) assert that all students should possess the critical response skills of being able to judge the quality of claims based on the use or misuse of supporting evidence, language used, and logic of the argument. as well as communication skills. The International Society for Technology in Education (ISTE) asserts that "capable information technology users" are skilled at information seeking, analysis, and evaluation as well as communicating, collaborating, publishing, and producing. Similarly, the National Research Council (NRC) report *Being Fluent with Information Technology* (1999) argues that intellectual capabilities such as organizing and navigating information structures and evaluating information and communicating to other audiences are just as much a part of technology fluency as are basic information technology concepts and skill at using contemporary technologies.

Because we wanted our prototype assessments to be potentially useful to a wide range of classrooms and research settings, we concentrated on technology-supported research and communication skills. Unlike knowledge standards that are necessarily different from subject to subject and grade to grade, these "new basics" are widely applicable.

Approach. The assessment prototype presents an engaging, problem-based learning task that integrate technology use with investigations of an authentic problem. The student outcomes assessed include technology use (Internet and productivity tools), reasoning with information, and communication. The assessment prototype builds on technology assessments developed for the WorLD Links program evaluation (Quellmalz & Zalles, 1999). The current assessment prototype extends that earlier work by probing students' Internet skills more deeply and exploring issues related to administering the assessments on the Internet and scoring student work online.

The prototype assessment task poses the problem that a group of foreign exchange students wants to come to the U.S. for the summer and needs to choose one of two cities. (See Figure 1.) The middle-school version of the assessment specifies that the foreign students are most concerned about recreational opportunities and public transportation. The secondary-school version adds the health of the city's economy as a third criterion. Examinees are asked to research information about the cities, decide which city the foreign students would prefer, and then write a letter to the foreign students recommending that city. Students are given URLs for a set of real, complex Web materials for the cities of Knoxville, TN, and Ft. Collins, CO. In addition to finding information on the specified dimensions, students were instructed to evaluate the credibility of information on particular Web pages and to formulate a search query for finding additional, relevant information. Students were asked to compare and weigh all of this information in making their selection and to present the reasons for their choice when writing their letter to the foreign exchange students. The assessment was designed to require approximately two hours to complete.

Analytic scoring rubrics were drafted to rate students' technology use, reasoning with information, and communication. A generic scoring rubric was then tailored to develop item-specific scoring criteria for each version of the assessment. Exhibit 1 presents an example of question-specific scoring rules.

Pilot Testing. The middle school prototype was first tried out with four students. SRI assessment staff observed each of the students, encouraged them to think aloud as they responded to each question, and debriefed each student when the tasks were completed. The prototype was refined based on the "think alouds" and prepared for pilot testing with a class of 31 middle-school students from an urban school in which students had been using the Internet in class projects.

The secondary-level version was first tried with two secondary students. The "think aloud" procedures and debriefing informed revisions. The secondary prototype was then administered on line to 62 high school students drawn from four Virtual High School courses. Students logged on from 26 schools to take the assessment. In general, students were able to complete the task in the two-hour time frame.

Results. For the secondary-school version of the assessment, two raters scored the student responses independently, with agreement levels ranging from 84-96%. Results indicated that in general, students handled the search and word processing easily. Particularly revealing were the varieties of queries the students generated, the ways they interpreted the request to identify questionable information on the Web sites they visited, and their explanations of why they found certain information questionable. In general, students demonstrated greater proficiency at finding topically appropriate information than at reasoning with the information or communicating conclusions in a well-organized and thoughtful manner. Not surprisingly, we are finding relationships between students' prior experience with technologies and their scores.

Next Steps. Upon completion of the analyses, the prototypes will be revised and further field tested. In addition, we plan to use the Internet research template to develop a prototype on a new topic and to design additional questions related to Internet searching and organization of information. The on-line scoring function will be further developed to permit examination of student work by question or for the entire task. The scoring rubric will be refined and sample student work representing differing levels of quality will be selected to illustrate the scoring levels.

Palm-top Collaboration Assessment

One of the trends observed in many technology-using classes is a move toward more student collaboration and more teacher coaching as opposed to direct instruction (Sandholtz, Ringstaff, & Dwyer, 1996; Penuel et al., 2000). In student-centered classrooms, whether using technology or not, it’s common to see students working in small groups to solve a problem or produce a product. The teacher in these classes can’t be everywhere at once. How does he or she assess the quality of students’ collaborative work? Our observations in Global Lab Curriculum classrooms, cited above, suggests that many teachers do relatively little to assess the quality of student collaboration in small-group work. Although the ability to work in teams is often cited as a critical workplace skill for the 21st century, most classrooms do little to enhance student skills in this area or help students become more reflective about their collaboration skills. When teachers do assess collaboration, the assessment is often simply a global measure of participation (or the absence of serious disruption) rather than a more nuanced assessment that includes the cognitive dimensions of working together to build a knowledge base or generate a plan, design, or product.

The challenge of assessing what is happening in multiple small groups within a classroom struck us as a suitable problem for taking advantage of the palm-top computer’s portability. We wanted to explore the feasibility of having teachers performing "mobile real-time assessments" of collaboration skills as they move from group to group performing observations and offering suggestions. Initially, we envisioned our assessment as a tool for teacher use, but as our work unfolded, we decided to explore its usability for student self assessment as well.

Approach. We began with a review of the academic research on collaboration. The research base yielded a large number of dimensions of collaboration, as shown in Exhibit 2. We realized that the limitations of the size of the palm-top computer screen as well as the mental workload imposed by the requirement for monitoring the functioning of multiple student groups all interacting at the same time meant that we would have to be selective in terms of the number of dimensions that teachers rate. At the same time, we wanted to offer

teachers enough options so that they felt the collaboration assessment would get at those features they think are most important. Accordingly, we began with a Web interface for teacher use in constructing an assessment tailored for his or her class. Figure 2 shows a screen shot of the prototype assessment-building interface. Under each dimension of collaboration, the teacher is presented with multiple potential assessment items (e.g., under the category of Forming Arguments, the teacher could select the item "Did group members back up their theories or ideas with supporting evidence?") that could be included on the tailored palm-top assessment. The teacher chooses those items he or she want to use and the resulting item set is down-loaded to the palm-top. (For our prototype assessment, we imported the assessment items into an off-the-shelf piece of software called *Survey Mate*.) It was necessary to pare down the item labels for the palm-top to avoid an overly cluttered screen. For each item, the teacher can observe each student group, rate the group on a simple three-point scale, and input the rating onto the palm-top computer. Exhibit 3 provides an example of a real classroom interaction and the way the students' behavior gets scored using the collaboration rubric.

Aggregated ratings for the whole class can then be uploaded onto the teacher's PC for classroom display (using the same Web site that supports assessment construction). Figure 3 shows a portion of such a display.

Pilot Test. We worked with a teacher of a nearby fourth/fifth-grade class to refine and pilot our assessment prototype. Working in an alternative public school, this teacher organizes most of his instruction around long-term projects and stresses the importance of student collaboration and students' ability to manage their own learning.

Before and after the use of the prototype assessment in this class, researchers administered a questionnaire concerning opinions regarding collaboration skills to the teacher and the students. We observed the teacher's use of the prototype assessment and then the use of the same assessment tool by the students themselves. Afterwards, we interviewed the teacher and the students concerning the usability of the palm-top assessment.

Results. Both the teacher and the students were able to use the palm-top tool. They felt comfortable both with the concepts in the items they were rating and with the palm-top interface. The teacher's ratings and those of the students generally followed the same pattern, but with groups earning higher scores on average from their teacher than they gave themselves. The teacher's interpretation of this difference was in terms of students' greater consciousness concerning effective collaboration processes when they knew he was nearby.

The pre- and post-questionnaire results suggested that there was some movement toward greater valuing of cognitive aspects of collaboration above sheer participation on the parts of both students and teacher. (See Figure 4.)

While providing the encouraging findings described above, the pilot test also revealed several limitations of the prototype. Although the teacher had picked the particular collaboration dimensions and items to score, once he actually tried to use them in the classroom, he found that they were not necessarily relevant to what students were doing at the time he was ready to rate their interactions. He wanted the capability to be able to modify the items to be scored "on the fly." Further, the research-based dimensions we used to organize the assessment items did not correspond well to the way the teacher thought about class activities. He indicated a strong preference for organizing assessment items by the type of activity (e.g., group research, planning, or design review) rather than by psychological dimension (e.g., developing social norms, assigning roles, or forming arguments). Finally, we found that some of the students were uneasy about being observed so closely. The challenge of attending to collaboration without raising the specter of "Big Brother" will need further attention. Our hope is that having students do more self assessment will mitigate this problem.

Next Steps. This fall we are redesigning the collaboration assessment with an organization based on student activity rather than psychological dimensions. We are also developing our own software shell to replace the off-the-shelf program that proved cumbersome for our purposes. The revised assessment will be pilot tested with five teachers to gain broader feedback on its usability.

Conclusion

While these assessment prototypes are still under development, they do offer illustrations of the way that technology supports can make classroom assessment of complex skills more feasible. One major advantage of embedding assessment within learning activities is the heightened focus on learning outcomes. Through the act of developing or choosing formative assessment measures, teachers must think about the kinds of skills and knowledge they are trying to impart through learning activities, and this reflection in turn supports better activity design and better articulation of learning goals to students. Research shows that the use of formative assessment as part of instruction increases learning (Black & Wiliam, 1998). Technology can make assessments of the kinds of skills needed for the 21st century knowledge economy more feasible–providing assessment tasks that mimic the features of real-world problems and providing portable, easy-to-use templates for collecting and storing classroom assessment data.

References

American Association for the Advancement of Science, (1993*). Benchmarks for science literacy: Project 2061*. Oxford University Press, New York.

Barron, B., Schwartz, D.L., Vye, N.J., Moore, A., Petrosino, A., Zech, L., Bransford, J.D. & CTGV. (1998). Doing with understanding: Lessons from research on problem and project-based learning. *Journal of Learning Sciences*.

Becker, H.J. (1999). Internet use by teachers: Conditions of professional use and teacher-directed student use. Teaching , Learning and Computing: 1998 National Survey of Schools and Teachers, Report #1. Irvine, CA: Center for Research on Information Technology and Organizations, University of California, Irvine and University of Minnesota.

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

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80* (2), 139-148.

ISTE (International Society for Technology in Education). (1998). *National Educational Technology Standards for Students.* Eugene, OR: Author.

NCTM (National Council of Teachers of Mathematics). (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Research Council. (1999). *Being Fluent with Technology*. Washington, DC: National Academy Press.

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

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

Penuel, B., Golan, S., Means, B., & Korbak, C. (2000). Silicon Valley Challenge 2000: Year 4 Report. Menlo Park, CA: SRI International.

Popham, J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership, 56*(6) 8-15.

Quellmalz, E.S. and Zalles, D. (1999). WorLD Student Assessment 1998-1999 Report. Menlo Park, CA: SRI International.

Sandholtz, J., Ringstaff, C., & Dwyer, D. (1996). *Teaching with Technology: Creating Student-Centered Classrooms*. San Francisco: Jossey-Bass.

Young, V. M., Haertel, G., Ringstaff, C., & Means, B. (1998). *Evaluating Global Lab Curriculum: Impacts and Issues of Implementing a Project-Based Science Curriculum.* Menlo Park, CA: SRI International.

Exhibit 1
Excerpts from Internet Research Task Scoring Rubric

URL Scoring

1: URL goes to a page from the wrong city, or from something unrelated, or has been entered incorrectly

2: URL goes to a page about the city in question, but not on the correct topic;

3: URL goes to a page on the same topic but not directly to the evidence (e.g., recreational opportunities, economy, public transportation) – (Note: use this if the student cites a topically-relevant URL but neglected to put in the evidence)

4: URL goes right to the page that contains the evidence, or provides a listing of the evidence

Evaluating Questionable Information

1: States that he cannot find questionable text, or the text he finds appears to be factual and he has not explained what he finds questionable about it

2: The student has found questionable text but has not tried to explain why

3: The student has found questionable text and has tried to explained why, but the explanation has some shortcomings

4: The student has found questionable text and has adequately explained why

Exhibit 2
Dimensions of Collaboration

Analyzing the Task

Developing Social Norms

Assigning and Adapting Roles

Explaining/Forming Arguments

Sharing Resources

Asking Questions

Transforming Participation

Developing Shared Ideas and Understandings

Presenting Findings

Exhibit 3
Scoring Classroom Interactions with the Collaboration Rubric

In one classroom session where we tested the assessment, we observed collaboration as groups of fourth- and fifth-grade students used their history textbooks to generate a list of causes of the American Revolutionary War. Members of one particular group were taking turns reading individual passages from their text aloud. When they finished, one student grabbed a pencil and asked the other students to write down what they read. Two students were particularly active in providing answers, and as an answer was given, the student with the pencil wrote it down. At one point, one of the students attempted to involve a boy who was not participating actively by telling him he had to give one answer. He reluctantly provided a cause for the Revolutionary War, but it was an answer the group had already generated. The three collaboration items the teacher had selected for the assessment can be applied

to the behavior of this group. All but one member of the group appeared to take responsibility for getting the assignment done, so the "yellow" choice, "some students are invested" would be selected as the answer to the question "How much do group members feel accountable for the success of the task?" For the question, "Do group members give explanations for concepts or phenomena they are studying?" the students would be scored lower (be given the "red" score, indicating that no explanations were given that elaborated on the content of the text). None of the students elaborated on the text they were reading; they simply summarized the text out loud, and the student with the pencil recorded each answer as it was called out.