Means, B. (2006). Prospects for transforming schools with technology-supported assessment. In R. K. Sawyer, (Ed.), *The Cambridge handbook of the Learning Sciences* (pp. 505-519). New York, NY: Cambridge University Press.

CHAPTER 30

Prospects for Transforming Schools with Technology-Supported Assessment

Barbara Means

The last two decades have been marked by great expectations – and some important progress – in bringing technology to bear on the process of schooling. True, the more dramatic predictions of two decades ago concerning the impact of Information Age technology on schooling (see for example, Milken Family Foundation, 1999; Office of Technology Assessment, 1995) have not come to pass. But even so, the cup is half full. Today the average student:computer ratio is 5:1 in American schools (National Center for Education Statistics, 2003), and the use of technology for student research and report preparation has become commonplace. But arguably the biggest "buzz" in the educational technology community today is around the use of technology to improve assessment. School reformers, technology enthusiasts, and business interests have all identified assessment as an area with great potential for increased classroom use of technology (Bennett, 2002; CEO Forum on Education and Technology, 2001; CoSN, 2005; Education Week, 2003).

Although enthusiasm for a marriage of classroom assessment and technology is widespread, there are two competing visions of the purpose and nature of effective classroom assessments, each with different implications for the role of technology. One vision calls for connecting classroom assessment practices more closely to state-mandated content standards and accountability systems. The other vision, which draws heavily on recent advances in the learning sciences, calls for using technology to develop and deliver assessments that are integrated into day-to-day instruction and that enable teachers to gain deeper insights into their students' thinking and to adapt their instruction accordingly. Proponents of both visions claim the term "formative assessment," but it means quite different things in the two visions.

In this chapter, I first introduce the vision of technology-supported assessment as an adjunct to standards-based accountability systems. I then discuss the concept of formative assessment and the weakness of accountability-based classroom assessments for formative purposes from a learning sciences perspective. Next, I turn to the main focus of the chapter – the second vision of

technology-supported assessments, with its roots in recent learning sciences research, and I discuss the more deeply diagnostic formative function of this vision. The remainder of the chapter provides descriptions of some notable examples of prototype and research-based systems implementing this second vision.

Technology-Based Classroom Assessment in Support of Accountability

One vision of the transforming role that technology-based assessment could have on schools involves tighter linkage between classroom practice and the standards set by district and state education offices, as embodied in accountability systems. If teachers are regularly and efficiently testing students on state standards, the reasoning goes, they will be able to use these test results to see where they need to focus instruction, both for a class as a whole and for individual students. By making it easy for teachers to assess students on the standards applicable to the subject and grade they teach, technology-based systems can encourage an appropriate focusing of instruction (and greater standardization across classrooms, schools, and districts within a state).

This vision precedes, but was certainly given strong impetus by the No Child Left Behind (NCLB) legislation. Since the passage of No Child Left Behind, the (1) heightened emphasis on tying instruction to specific content standards and (2) increasing stakes tied to students' performance on statewide tests (i.e., the requirement that schools demonstrate Adequate Yearly Progress on test scores for all student subgroups) have created a demand for assessment tools that teachers can use to gauge their students' progress, and that administrators can use to identify potential trouble spots early on. Commercial entities have been quick to see the market potential offered by school, district, and state administrators who are nervous about their students' likely performance vis-à-vis NCLB-required Adequate Yearly Progress. Market trend analysts note that there has been stagnation in recent years in the market for instructional software, but dramatic growth in sales of computer-based assessment systems (Dyson, 2004). Accountability-related testing systems have been part of what the 2003 special issue of *Education Week* – devoted to technology – referred to as "the greatest precollegiate testing boom in history" (Education Week, 2003, p. 10).

In some cases, states and districts themselves are developing computer-based assessments that students can take to practice for the state's tests. An *Education Week* survey of state departments of education in 2003 found that twelve states offered such practice exams. Notable among the state systems are those of Texas, with its online Texas Math Diagnostic System, and Florida, where the FCAT Explorer provides test items and skills practice keyed to the Florida Comprehensive Assessment Test (Borja, 2003; Olson, 2003). Among the available commercial products is Pearson's Progress Assessment Series. A February 2005 ad for this product (Pearson Education, 2005, p. 9) opened with the headline, "Measure Their Success Before They Achieve It." The advertisement explains that formative assessments, taken throughout the year, forecast performance on state-specific proficiency standards.

Pinnacle Plus, from Excelsior Software, offers computer-based "real-time assessment feedback for truly informed instruction" (eSchool News, September 2004, p. 31). HOST Learning claims that its LearnerLink product "simplifies standards-based instruction in the classroom" by providing "formative assessments and prescriptive lesson planning" that are "aligned with state and local standards as well as large-scale standardized assessments" (http://www.hosts.com/products/learnerlink.html, accessed 7/31/05). Students take the computer-based assessments, and the LearnerLink System presents the teacher with a list of instructional resources geared to the standards that the students have not yet mastered. SOLAR

System offers pre-assessments that pinpoint achievement gaps. PLATO's *eduTest* (formerly distributed by Lightspan) offers online assessments linked to state standards for classroom use. The PLATO eduTest Brochure, available at (http://www.plato.com/products.asp?cat=Assessment&ID=83, accessed 7/31/05) describes the product as "a comprehensive standards-based online assessment program for classroom formative and district benchmark assessments and reporting." A selling point used for these products is that they enable a principal or district administrator to obtain midyear information on how well students are doing with respect to the requirements for annual improvement in the state's standards for NCLB.

The primary advantage touted in promoting these systems is the ability to identify specific standards a student has not yet attained at a point during the school year when there is still time to provide additional instruction. Some of the systems, like the FCAT Explorer, include an instructional component; others provide teachers and administrators with the assessment results and leave it to them to provide appropriate instruction.

A technology-enabled advantage of these systems is the capability for customization. Commercial vendors have large banks of test items mapped to specific skills and content areas. One can use the system to develop tests that fit a specific state's or district's standards for a particular grade level. Many of the systems offer teachers the option of reviewing test items keyed to standards and then selecting a subset of items that they deem most appropriate. Some systems allow teachers to modify test items or add their own. The testing formats used in these systems are mostly multiple choice and short answer.

Although marketing materials often describe such systems as formative assessment – a type of assessment that is conducted during an instructional unit or sequence with the purpose of making instruction more effective – it should be noted that they provide limited (though important) information to guide instruction. The reports typically provide information on those standards on which the student performs at a level commensurate with state or district expectations, and those standards on which the student falls below the criterion level. The systems reflect a *mastery learning* approach to instruction (Bloom, 1976; Keller, 1983): the content to be learned is subdivided into discrete topics or skills, and individual students work on a topic or skill until it is "mastered." Instruction and learning are characterized in terms of exposure or "time on task" rather than in terms of the quality or nature of interaction with the material. A student failing to meet a standard typically receives more instruction of the same type received before, or even repetition of the very same practice exercises. This view of learning is very different from that underlying modern work in the learning sciences (see Sawyer introduction, this volume).

Rising Interest in Formative Assessment

While the increased emphasis on accountability testing is undoubtedly the most obvious assessment trend in recent years, there has been a concurrent rising interest in the concept of genuinely formative assessment. Formative assessment is contrasted with *summative assessment*, the testing that occurs at the end of a unit of instruction to document or certify what has been learned. An influential review of over 250 studies by Black and Wiliam (1998) concluded that the use of formative assessment techniques was one of the most powerful classroom-level interventions documented in the research literature:

> *There is a body of firm evidence that formative assessment is an essential feature of classroom work and that development of it can raise standards. We know of no other way of raising standards for which such a strong* prima facie *case can be made on the basis of evidence of such large learning gains. (p. 19)*

Formative assessment may be receiving more lip service than serious adoption. Its use is widely advocated by researchers (National Research Council, 2001; Pellegrino, Chudowsky, & Glaser, 2001; Shepard, 2000) and it is finding its way into professional development activities (Black & Harrison, 2001; Koch & Sackman, 2004; McTighe & Seif, 2003; Shepard, 1997). As described earlier, commercial entities distributing test preparation software tout their products as tools for formative assessment. However, the activities they call formative assessments are often simply the same end-of-year summative assessments administered before the end of the school year, not assessments that are designed specifically to inform future instruction.

This critical component – that the assessment provide information that shapes further instruction – was emphasized by Black and Wiliam (1998). If an assessment is used merely to assign grades, and there are no further learning opportunities on the content, the assessment is not really formative. Only if the assessment reveals specifics about students' thinking in ways that can inform further instruction, *and* additional learning opportunities are provided that make use of that diagnostic information, is the assessment "formative" in the sense that Black and Wiliam applied the concept in reviewing classroom studies. (See also the concept of "informative assessment" in Bass & Glaser, 2004.) *The formative nature of an assessment thus lies not in the assessment per se but rather in the intersection between the assessment and its role in the classroom.* (This is similar to a recent shift in the concept of test validity – validity should be evaluated based on how a test's results are used, rather than being inherent to the test instrument itself independent of the context of use.)

Many of the systems that are designed to support classroom assessments that are linked to standards and accountability systems lack this ability to inform instructional decisions. They tend to provide information on whether a student has achieved mastery, but not to provide insights into the way the student is thinking. The items in these systems tend to stress facts, name recognition, and discrete procedures, rather than deeper understanding or the relationships among concepts. Although nearly every curriculum standard–setting body makes statements about the importance of emphasizing depth in key content areas rather than mere breadth of coverage, the proliferation of standards across multiple content areas has produced the notorious "mile wide–inch deep" approach to content in American classrooms (Schmidt et al., 1997). A concrete illustration of how this works was provided by the comments of a math teacher whom I interviewed recently. The teacher showed me the pages of math standards that his large, urban district has stipulated need to be taught in each quarter of the academic year. This requirement is given teeth through district testing on the standards every nine weeks. When I asked the teacher what he did about students who needed more time to attain a standard, he reported that he had to follow the district's guidance, which he characterized as "touch upon it and move on." Technology can make this process more efficient, but when used to generate tests focusing on standards coverage rather than on student understanding, technology will only reinforce a bureaucratic approach to education.

Assessment in the Service of Cognitive Diagnosis

Genuinely formative assessment is an integral part of the vision of researchers who are trying to bring insights from the learning sciences to bear on classroom practice. In drawing educational implications from the body of research summarized in *How People Learn* (Bransford, Brown, & Cocking, 1999), a committee established by the National Research Council (Donovan, Bransford, & Pellegrino, 1999) concluded, "Teachers must draw out and work with the preexisting understandings that their students bring with them" (p. 15) and "Formative assessments – ongoing assessments designed to

make students' thinking visible to both teachers and students – are essential" (p. 21).

Cognitive research on how people learn and acquire expertise in various content areas has given rise to new perspectives on assessment. From this standpoint, the purpose of a formative assessment is to provide insight not so much into a student's level of performance as into the nature of the student's understanding and reasoning in the domain being assessed. Mastery learning approaches decompose learning goals into discrete skills or bits of knowledge, and seek to assess whether each of these is in a state of mastery or nonmastery. In contrast, cognitive learning theorists put greater emphasis on assessing the way in which knowledge and skills are organized in the minds of students. These theorists describe knowledge as a hierarchically organized conceptual framework that predisposes the individual to see problems in terms of meaningful patterns, and to apply approaches that have been successful in dealing with these patterns to new situations. Furthermore, there is a tradition of cognitive research conducted within specific subject domains – contrasting the thinking and problem solving behaviors of novices and experts, or of students at different stages of development. These studies suggest that there are often different ways of "not knowing" something. The younger student's thinking is described not as random guesses produced in the absence of knowledge, but rather as consistently off the mark in a predictable way that makes sense given a particular misconception (diSessa, this volume). Common misconceptions have been identified for numerous science topics, including classical mechanics (Champagne, Klopfer, & Anderson, 1980), motion (Carmazza, McCloskey & Green, 1981; McCloskey, 1983; McCloskey & Kohl, 1983), and understanding of seasons and the movement of the earth (Sadler, 1987). Often students have models of some aspect of a subject area that are partially correct and that lead to correct answers or adaptive responses in some situations but not others (Linn, this volume). In some cases, there is a fairly typical developmental sequence of ways of thinking. For example, most young children can answer questions about "more" and "less" and can reliably count a set of objects before they can integrate these two competencies to answer questions about whether one number is more or less than another (Griffin & Case, 1997). In other cases, there are different common misconceptions and no known dominant sequence (Chi & Slotta, 1993). Learning theorists stress the need for teachers to understand how individual students think about the phenomenon being studied:

> *Students come to the classroom with preconceptions about how the world works. If their initial understanding is not engaged, they may fail to grasp new concepts and information presented in the classroom, or they may learn them for purposes of a test but revert to their preconceptions outside the classroom. (Donovan, Bransford, & Pellegrino, 1999, p. 2)*

In this view, the goal of formative assessment is not so much to ascertain whether a student "has got it" as it is to reveal the way in which the student is thinking about the topic or problem. Two students may both lack a scientifically correct understanding of some phenomenon, for example, but may think about it in very different ways. If a teacher is to focus teaching at each student's level of prior understandings, the teacher needs to know the nature of individual students' thinking. In those curricular areas where research has uncovered detailed models of cognition, those models provide a basis for the development of formative assessments that can inform instruction.

Using Cognitive Research to Design Assessment Items

Phil Sadler's work on students' understanding of the movements of the earth and the solar system provides a good example of how assessments can provide a diagnosis of student understanding. Earlier research in which students of various ages were questioned about the earth's orbit and the reason why temperatures vary with the season revealed that many students think that the

earth's orbit is shaped in such a way that the earth is physically closer to the sun in summer (Sadler, 1987). Other students retain the notion that distance from the sun is the critical factor, but describe differences in distance caused because the earth "leans" toward the sun in summer and away from it in winter (from a Northern Hemisphere perspective). Only a minority of students ever attains the understanding that the earth's spherical shape coupled with the earth's tilt means that the angle at which the sun's rays strike the earth is larger in summer, resulting in longer days and the sun being higher in the sky. Sadler used this research on misconceptions and the development of understanding of seasons as the basis for developing a series of test items (Sadler, 1998). In addition to a correct answer option, each item had incorrect choices ("distractors") which were the answers one would give on the basis of the various misconceptions found in the research literature. Similarly, items were developed to probe students' reasoning about what causes night and day, about the relative distance of stars outside our solar system, about the phases of the moon, and so on. Sadler's astronomy assessment is used to identify the ways in which students think about each of these areas, not just whether or not they give scientifically correct accounts. By giving his assessment to large numbers of students, Sadler was able to demonstrate several instances in which the proportion of students choosing the correct answer appeared to fall off as student age increased, until it began to rise among the oldest grade levels tested. Sadler used these data to argue that in these cases a decrease in "correct" answers to some kinds of questions is in fact a marker of cognitive development. An immature conceptual framework often produces correct answers to some questions for the wrong reasons; as students' concepts mature, they may start giving wrong answers to questions they previously answered correctly for some period, until a new, more comprehensive understanding has solidified.

This description of Sadler's work shows that building an assessment based upon empirical research on how students think about a complex topic is a labor-intensive undertaking. Learning sciences research provides us with a foundation for this work in the form of cognitive studies of reasoning in areas such as early mathematics, biology, calculus, economics, physics, and history (Bransford, Brown, & Cocking, 1999; Donovan & Bransford, 2005). Even so, the systematic development of assessments that are sensitive to important differences in understanding within a domain requires not only careful crafting of assessment tasks but also trying them out with examinees and determining their technical characteristics (e.g., the distribution of examinee responses and the interrelationships among responses to different assessment tasks or portions of tasks). Creating research-based assessments for all areas of the K–12 curriculum would require a huge effort. Certainly, developing research-based assessment items is not something that we would expect every teacher, or even every school district, to take on independently. Although this task is too big for an individual teacher or district, it can be tractable if we pool our efforts, and technology supports make this kind of pooling possible. By developing Web-accessible versions of assessments capable of cognitive diagnosis, researchers are starting to provide resources for an increasing number of classroom teachers.

Using the Web to Make Diagnostic Assessments Widely Available

Jim Minstrell, Earl Hunt, and their colleagues have developed of technology-based diagnostic assessment items in a content area, and have disseminated them over the Web. A former high school physics teacher himself, Minstrell began compiling a set of student conceptions about force and motion based upon both the research literature and the observations of teachers. Some of these ideas, or "facets" in Minstrell's terminology, are considered scientifically correct (or at least correct to the degree one would expect at the stage of introductory physics). Others are partially incorrect, and still others are

seriously flawed. The goal of assessment in this work is to elicit student responses that reveal the underlying thinking, or *knowledge facets*, of each student. Having developed an inventory of knowledge facets, Minstrell and his colleagues proceeded to develop assessment items that would elicit different responses depending on which facets a student held (Minstrell, 1999). For example, when asked to reason about the weight of objects totally or partially submerged in a liquid, one set of facets concerns separating the effect of a fluid or other medium from the effect of gravity. A student might think that surrounding forces don't exert any pressure on objects. Alternatively, he might think that fluid mediums produce an upward pressure only or that the weight of an object is directly proportional to the medium's pressure on it. Some students may have memorized the mathematical formula for net buoyant pressure and may be able to apply it to some problems in order to obtain a correct answer, but might nonetheless lack the facet for a qualitative conceptual understanding (net upward push is a result of differences in pressure gradients).

Minstrell and his colleagues (see http://www.facetinnovations.com, accessed 7/31/05) have developed a computer-based assessment system to get at students' facets. The student is presented with a problem situation (e.g., a solid cylinder is hung by a long string from a spring scale. The reading on the scale shows that the cylinder weighs 1.0 lb. About how much will the scale read if the cylinder which weighs 1.0 lb. is submerged just below the surface of the water?) and a set of multiple-choice answers, each of which is associated with a specific facet. After choosing an answer to the original question, the student is asked to provide the reasoning behind the original answer. The system compares the facet associated with the student's explanation to that associated with the original answer choice. Over multiple problems, the system diagnoses the student's probable facets and the consistency between student predictions and explanations. The system presents the teacher with reports of this diagnosis and with an instructional prescription appropriate for the diagnosed facets.

The first technology-based version of the system, called DIAGNOSER, was developed using Apple's HyperCard environment. More recently, the FACETS system offers teachers Web access to diagnostic assessments in a variety of content areas in middle school science and mathematics. The system also provides guidance to teachers on how to analyze their students' open-ended explanations in terms of facets. After student misconceptions are identified through the FACETS assessments, "benchmark lessons" are suggested to challenge student beliefs. Their purpose is to encourage students to apply their beliefs to new situations, examine their own reasoning, and see where their expectations are confirmed and where there are discrepancies between their beliefs and what they observe actually happening.

Using Technology to Perform Complex Diagnoses

The FACETS assessments illustrate technology's contribution to formative assessment in providing readily accessible banks of research-based diagnostic assessment items and associated instructional strategies and materials. Technology also has the potential to perform complex analyses of patterns of student responses that would be time-consuming and difficult for teachers to perform. An example of the use of technology to analyze student performance with complex problems in ways that would be difficult or impossible for a human teacher is provided by the IMMEX (Interactive Multimedia Exercises) system at UCLA (Vendlinski & Stevens, 2000). The system was originally developed to teach diagnostic skills to medical students, but has been extended to teach problem solving in grades kindergarten through twelve in topics in earth science, chemistry, social studies, mathematics, and language (see http://www.immex.ucla.edu, accessed 7/31/05).

The IMMEX system presents students with a dramatic, complex problem scenario such as spilling of a chemical in a stock

room after an earthquake, or identification of a child's true parents after a potential incident of baby switching at the hospital. The learner is challenged to investigate the problem, and is given both background information and the option to perform various simulated tests (such as submitting the mystery chemical to a flame test) and to receive the results of the tests before proposing a solution. The system maps out the information that each student examined, the sequence in which it was examined, and the amount of time spent in each portion of the problem (Underdahl, Palacio-Cayetano, & Stevens, 2001). Using neural network technology, the IMMEX system compares the sequence of steps taken by a learner to the problem-solving steps of more and less skilled learners working with the same type of problem. The teacher receives a search path map showing each student's problem-solving strategy. Student solutions can be classified in terms of problem-solving strategies which vary in the kinds of information examined, the incorporation of strategic elements (such as choosing information most likely to rule out alternatives), and the likelihood of reaching a correct solution for a given problem set (Stevens et al., 2004).

The IMMEX team has developed additional supports for teachers to implement IMMEX problem solving units in their classrooms. An authoring system allows teachers to author their own IMMEX problems. Professional development around the use of IMMEX provides teachers with suggestions on when and how to provide additional instruction on problem solving strategies, based on students' performance on IMMEX problems (Underdahl, Palacio-Cayetano, & Stevens, 2001).

Technology Supports for Student Self-Assessment

The ThinkerTools Inquiry Project of Barbara White and John Frederiksen has developed technology-supported classroom assessment practices that involve students in assessing their own understanding. ThinkerTools is a middle school curriculum that helps students learn the concepts of force and motion, as well as the processes of scientific inquiry. Each "inquiry cycle" begins with the identification of a question about a set of phenomena (e.g., how a puck moves on surfaces with different amounts of friction) that students do not yet understand. The class is subdivided into small research groups to discuss their intuitions about these situations and to develop hypotheses or models. Students then test their models, using both computer-based simulations and real-world materials, and record their observations. Each group reasons about what they have learned from testing their predictions and models, and then all the groups reassemble in a research symposium to present their models, findings, and interpretations. The class discusses the various groups' models and findings, coming to consensus on the best model, and then tries to apply that model to new situations. White and Frederiksen (2000) tested the effects of adding a formative assessment component to ThinkerTools in which the students themselves engaged in what the researchers called "reflective assessment." A computer-based system provides students with a set of criteria characterizing good scientific inquiry, and asks them to evaluate their own and each other's work according to those criteria. The inclusion of this formative assessment component led to greater gains on a science inquiry test, especially for those students who began the project with lower scores on a standardized achievement test. Thus, the incorporation of reflective assessment appeared to reduce the achievement gap in the area of science inquiry.

White and Frederiksen argued that such self-assessment activities address an important goal for instruction: Students need strong models of good performance and also need help internalizing the criteria for high-quality performance. The ThinkerTools work exemplifies the way in which students can be actively involved in classroom assessment practices – not just responding

to examination questions, but rather applying assessment criteria themselves. This kind of involvement highlights both the essential qualities of the performance or product students are expected to achieve and the process whereby such products are created (Frederiksen & Collins, 1989).

Combining Technology Supports and Human Judgment

While the IMMEX system described above illustrates the capability of neural net software to perform pattern diagnoses that would be difficult and time consuming for humans, there are many scholars in the learning sciences community who would argue that in most cases, human observers can do better instructional diagnoses than can today's artificial intelligence software. Even if that were true, technology could still play an important role by delivering problems that elicit a systematic set of student responses for the human observer to use in instructional diagnosis.

In ThinkerTools, for example, the technology-based environment provides simulations for use in testing students' ideas, and a structure reflecting the inquiry cycle, which can both be related to assessment activities. The reflective assessment itself, however, is done not by the system but by the learners themselves and their peers. In a similar vein, research on an intelligent tutoring system for high school geometry (Schofield, Eurich-Fulcer, & Britt, 1994) found that teachers could use students' interactions with computer-based geometry problems as a basis for providing feedback and assistance that was better tuned to student needs, according to the students. Schofield et al. concluded that teacher assistance was more articulate and relevant to student needs than the feedback provided by the computer-based tutoring system, but the tutor was able to reveal areas of individual need in ways that normal whole-class instruction was not. Hence, the tutoring system and the teacher together provided formative assessment in a way that neither could do alone.

Integrating Technology-Based Assessment with the Act of Teaching

In contrast to the approaches described earlier, which have stressed the diagnosis of individual students' thinking, are a set of emerging technology-supported practices that provide instructors with a nearly real-time snapshot of the thinking of entire classes (Means et al., 2004). The challenge of teaching large college classes triggered the development of classroom communication systems (CCS). A CCS consists of a networked set of computers, personal digital assistants, or small wireless input devices that look like TV remote control. Every student in the class can answer the question using their personal device, and the wireless network aggregates and presents their responses on a screen (usually as a histogram).

Eric Mazur, a Harvard physics professor, has been one of the major proponents of using classroom communication systems to support teaching for conceptual understanding (Crouch & Mazur, 2001; Fagen, Crouch, & Mazur, 2002; Mazur, 1997). After finding that many of his students still did not understand basic physics concepts at the end of his course, Mazur switched from his accustomed practice of delivering content through lecture alone to using formative assessment and a technique he called "peer instruction." Using this new approach, Mazur would lecture for a short while and then pose a conceptual question (e.g., "Imagine holding two bricks below water. Brick A is just below the surface, while Brick B is at a greater depth. How does the force needed to hold Brick B in place differ from that needed for Brick A?"). Students used the classroom's communication system to register their responses to the question, and then Mazur would invite them to work with one or two other students, discussing their answers and providing explanations to try to convince each other. After these discussions

among small groups, Mazur would had the class answer the same question a second time. Typically, the proportion of correct responses rose dramatically. At this point, after the students had been actively engaged via the initial visible clash of opinions, Mazur then explored the topic more deeply, for example, by challenging students to think about the limits of the rule or explanation they had converged on, or the relationship of that explanation to underlying principles.

Mazur (1997) offered concrete advice on the strategic planning and classroom practice required for a teacher to implement peer instruction. He made it clear that there is an art to designing good tasks and questions; in particular, the task must get to the heart of the conceptual matter and be neither too easy (or there is no need for discussion) nor too hard (which would result in an insufficient distribution of the correct answer among the class population). Mazur believed that technology's contribution resides in prompting students to think deeply enough about the question initially to commit to a response to the question, and in making students feel comfortable in arguing for their response by making it clear that the class as a whole holds a range of opinions (as reflected in the projected histogram of student responses). Students' initial thinking is made visible, both through their responses to the instructor's question and through the arguments and explanations they offer to their peers.

Classroom communication systems are starting to be applied in K-12 settings as well. Hartline (1997) described the practices of an elementary reading teacher who used a classroom communication system to check students' comprehension of reading passages. After having the fifth graders in her inner-city school read text passages, the teacher had them use the classroom communication system to answer comprehension and inference questions about the passages. When the students finished, she would open up discussion around conceptual issues by projecting a histogram of class responses to the first question. If students had different responses, she asked students to volunteer "clues" from the reading passage that could help explain or justify their particular answer choices. As students called out clues, she would write them on the blackboard next to the answer. Then students were invited to talk about which was the best set of clues and why one set of clues was more persuasive than another. After discussion, students could change their answers, and then the teacher projected a new histogram and introduced another cycle of discussion.

Both the college physics class and the elementary reading class illustrate ways in which a classroom communication system provides a means of eliciting student understandings so that rich classroom discussions that connect with and build on that initial understanding can occur (also see Pea & Maldonado, this volume). Although students do not think of these experiences as tests, they do have all the earmarks of formative assessments.

System Supports for Assessment Design

Developing formative assessments based on cognitive research in specific domains is a resource-intensive activity. How could we possibly develop enough of the kinds of cognitive formative assessments described here to have a significant impact on K–12 education? Fortunately, technology, when coupled with advances in cognitive research and assessment theory, holds potential for supporting the development of formative assessments with cognitive diagnostic value (Pellegrino, Chudowsky, & Glaser, 2001). The ongoing Performance Assessment Designs for Inquiry (PADI) Project, led by Geneva Haertel and Bob Mislevy, is attempting to bring together advances in cognitive science – which provide insights into how students reason in particular content domains – with advances in psychometric models, which can now handle more complex, multipart items than was possible in conventional psychometric models. PADI is implementing

a process of evidence-centered design (Mislevy, Steinberg, et al., 2003) in which assessment items are constructed systematically, beginning with a *student model* of proficient performance in the area to be assessed, then moving to a *task model*, which specifies the assessment situation to be presented to the learner, and an *evidence model*, which specifies the evidence needed to support the inferences that can be made from various possible elements of task performance. This approach is being applied in the development of *design patterns* that characterize learning goals common in middle school science curricula (e.g., using data to support a scientific argument, designing and conducting a scientific experiment, evaluating the quality of scientific data) in a narrative form that lays out the logic chain linking the conclusions one wants to make about student thinking to the observable behaviors that would support or refute those conclusions and the situations that could elicit those behaviors.

The assessment developer using the computer-based PADI tools has a structure for laying out the logic that leads from specific student responses on an assessment task to inferences about the student's understanding and skill. The design patterns lay the groundwork for the more technical specification required for the design of particular assessments (Mislevy, Hamel, et al., 2003). The design patterns apply across different science specialties (i.e., they are applicable to genetics, chemistry, and physics); currently, the PADI design patterns are being used by science curriculum development projects at the University of Michigan and the University of California, Berkeley's Lawrence Hall of Science. Because assessments built on the PADI framework are designed around a cognitive model of proficient task performance, they are capable of supporting both formative and summative assessment practices, with teachers using assessment findings to draw inferences about their students' thinking and about the specific kinds of learning opportunities their students need.

Shall the Twain Ever Meet?

This chapter has described two very different visions of technology supported assessment. Both visions call for frequent assessment and a closer connection between instruction and assessment, but they have very different views of the nature of good assessments and of what teachers and students should learn from the assessment activities. The first vision in essence makes the standards-driven accountability system part of day-to-day classroom activity – practice tests pinpoint areas where an individual student is likely to fare poorly on the end-of-year test that is used for accountability purposes, so that classroom instruction can focus on those topics or skills. The second vision arises out of concern with understanding the preconceptions and problem solving strategies that students bring to the classroom. The notion here is that teachers must address not only students' mastery of content, but also their preconceptions and problem-solving strategies, to bring about deep and lasting changes in student thinking.

The two visions reflect very different learning theories. The first is closely associated with large-scale testing, and thus inherits its testing's relationship to the goal of differentiating among students with varying levels of achievement or "potential," and reflects behavioristic theories of learning (Shepard, 2000). The second derives from modern conceptions in the learning sciences (Bransford, Brown, & Cocking, 1999) and the goal of understanding human intellectual performance as a precursor to enhancing it. Interestingly, proponents of both visions borrow some formative assessment arguments and jargon from the other, but the two remain quite different at their core.

Two questions remain:

- Are the visions really in conflict?
- Does either or both have the potential to transform schools?

The two visions are not necessarily mutually exclusive. A given classroom could

exercise both types of technology-supported assessment at different times. Adoption of one approach does not preclude adoption of the other, and the two could, in fact, share the same technology infrastructure. But even so, classrooms always face a trade-off between covering a small number of important concepts in depth versus covering a large amount of content with less time per topic, and the two assessment visions come down on two different sides of the depth versus breadth trade-off. One could also characterize this trade-off in terms of focus: Assessment practices geared to the school's accountability system will tend to focus teaching and learning on the content covered in the state or district test; practitioners, researchers, and the general public differ in whether they view this focus as a positive or a negative result. Each combination of school system and classroom teacher will resolve this trade-off somewhat differently.

The two visions also suggest somewhat different roles for the teacher. Accountability-oriented assessments provide the teacher with a set of data designed to predict performance relative to state or district standards. But the data teachers receive from these assessments are typically either at a level of aggregation that is too gross to provide instructional insights beyond a focus topic (e.g., spend more time on fractions).[1] This situation is a natural consequence of the fact that these assessments link to standards rather than to students' thinking. The learning sciences approach, in contrast seeks to provide insight into how students actually think about the area being assessed. Cognitively oriented formative assessments provide teachers with more information concerning a starting point for instruction, but by themselves do not tell teachers how to help their students move from that starting point to a more complete understanding. Either approach to formative assessment, then, depends on teachers bringing considerable pedagogical content knowledge to the task of designing instruction to support further learning (Shulman, 1987).

With respect to the potential for significant change, the two visions have rather different prospects. Technology-supported accountability-focused classroom assessment is an increasing reality. Many districts are purchasing or considering systems that integrate classroom testing with their larger student information systems. One market analyst (Dyson, 2004) estimated the sales at $645 million and growing. In addition to giving teachers earlier feedback about areas where students need more instruction if they are to achieve state standards, these assessment systems, because of their linkage to student data systems, provide more detailed, inspectable information for administrators at higher levels of the education system (U.S. Department of Education, 2004). A principal or district curriculum specialist can direct a teacher to spend additional time teaching fractions or to give more emphasis to spelling. For this reason, individuals at higher levels of the education system are promoting use of these computer-based assessments. Widespread implementation of these systems is just beginning. We know that districts are buying them, but we do not know how and to what extent teachers are using the assessment features of the systems with their students. To the extent that districts adopting accountability-related assessment systems experience increases in test scores, this trend is likely to gather momentum.

The alternative – the cognitive diagnosis vision of technology-supported assessment – faces longer odds for having a real impact on schooling. The vision has the potential to bring about the blurring of the distinction between assessment and instructional practices – as many researchers advocate (Bransford, Brown, & Cocking, 1999; Pellegrino, Chudowsky, & Glaser, 2001; Shepard, 2000). It has strong appeal within the learning science research community and among some teachers. The challenges of producing research-based assessments for cognitive diagnosis is much higher than that of producing practice versions of standardized reading and mathematics tests, however. Without the commercial

interests and policy imperatives that are propelling the accountability-oriented classroom assessment systems, these research-based systems face many more barriers to adoption. Moreover, such approaches make heavy demands on teachers to develop deep expertise, both in the subjects they teach and in the ways students think about and problem solve in those content areas. If the theory underlying this approach is correct, though, instruction with cognitively diagnostic formative assessment should produce learning that is more long lasting and more likely to be brought to bear in new learning or problem solving situations (Bransford & Schwartz, 1999). Research demonstrating these advantages could well be a critical enabler of acceptance of technology-supported cognitively diagnostic assessments.

Acknowledgments

I am indebted to my Center for Technology in Learning colleagues Geneva Haertel and Bill Penuel for their advice and suggestions provided in response to an earlier draft.

Footnote

1. It could be argued that these systems do provide detailed information since many of them are capable of generating performance reports at the individual item level. When assessment items are developed and sampled simply as exemplars from the larger topic covered by a subtest, however, item-level performance information adds little except the temptation to teach the correct answers to missed items (rather than addressing a conceptual misunderstanding).

References

Bass, K. M., & Glaser, R. (2004). *Developing assessments to inform teaching and learning.* CSE Report 628. Los Angeles: National Center for Research on Evaluation Standards, and Student Testing, University of California, Los Angeles.

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment*, 1(1). Available from http://www.jtla.org, accessed November 22, 2005.

Black, P., & Harrison, C. (2001). Feedback in questioning and marking: The science teacher's role in formative assessment. *School Science Review*, 82(301), 55–61.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment and Education*, 5(1), 7–74.

Bloom, B. S. (1976). *Human characteristics and school learning.* New York: McGraw-Hill.

Borja, R. R. (2003). Prepping for the big test. *Technology Counts 2003*, 22(35), 23–24, 26.

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

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (Vol. 24, pp. 61–100). Washington, DC: American Educational Research Association.

Carmazza, A., McCloskey, M., & Green. B. (1981). Naïve beliefs in "sophisticated" subjects: Misconceptions about trajectories of objects. *Cognition*, 9, 117–123.

CEO Forum on Education and Technology. (2001). *School technology and readiness – Key building blocks for achievement in the 21st century: Assessment, alignment, access, analysis.* Available at www.ceoforum.org/downloads/report4.pdf, accessed November 22, 2005.

Champagne, A. B., Klopfer, L. E., & Anderson, J. H. (1980). Factors influencing the learning of classical mechanics. *American Journal of Physics*, 8, 1074–1075.

Chi, M. T. H., & Slotta, J. D. (1993). Ontological coherence of intuitive physics. *Cognition and Instruction*, 10(2&3), 249–260.

CoSN (Consortium for School Networking). (2005). *From vision to action: How school districts use data to improve performance.* Washington, DC: Author.

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *The Physics Teacher*, 69, 970–977.

Donovan, M. S., & Bransford, J. D. (2005). *How students learn history, mathematics, and science in the classroom*. Washington, DC: National Academy Press.

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

Dyson, E. (2004, October). Held back: The market for software in our schools. *Release 1.0. Esther Dyson's Monthly Report*. Available at http://www.release1-0.com, accessed November 22, 2005.

Education Week. (2003). Pencils down: Technology's answer to testing. *Technology Counts 2003*, 22(35), 8, 10.

Excelsior Software. (2004, September). Advertisement for Pinnacle Plus Assessment Management System. *eSchool News*, p. 31.

Fagen, A. P., Crouch, C. H., & Mazur, E. (2002). Peer instruction: Results from a range of classrooms. *The Physics Teacher*, 40, 206–207.

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

Griffin, S., & Case, R. (1997). Re-thinking the primary school math curriculum: An approach based on cognitive science. *Issues in Education*, 3(1), 1–49.

Hartline, F. (1997). *Analysis of 1st semester of Classtalk use at McIntosh Elementary School*. Yorktown, VA: Better Education.

Keller, J. M. (1983). Motivational design of instruction. In C. Reigeluth (Ed.), *Instructional-design theories and models: An overview of their current status*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Koch, M., & Sackman, M. (2004). Assessment in the palm of your hand. *Science and Children*, 33(9), 33–37.

Mazur, E. (1997). *Peer instruction: A user's manual*. Upper Saddle River, NJ: Prentice Hall.

McCloskey, M. (1983). Naïve theories of motion. In D. Genuner & A. I. Stevens (Eds.), *Mental models* (pp. 299–324). Hillsdale, NJ: Lawrence Erlbaum Associates.

McCloskey, M., & Kohl, D. (1983). Naïve physics: The curvilinear impetus principle and its role in interactions with moving objects. *Journal of Experimental Psychology Learning, Memory, and Cognition*, 9, 146–156.

McTighe, J., & Seif, E. (2003). *A summary of underlying theory and research base for understanding by design*. Unpublished manuscript.

Means, B., Roschelle, J., Penuel, W., Sabelli, N., & Haertel, G. (2004). Technology's contribution to teaching and policy: Efficiency, standardization, or transformation? In R. E. Floden (Ed.), *Review of Research in Education* (Vol. 27, pp. 159–181). Washington, DC: American Educational Research Association.

Milken Family Foundation. (1999). *Transforming learning through technology: Policy roadmaps for the nation's governors*. Santa Monica, CA: Author.

Minstrell, J. (1999). Facets of student understanding and assessment development. In J. W. Pellegrino, L. R. Jones, & K. Mitchell (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP*. Washington, DC: National Academy Press.

Mislevy, R. J., Hamel, L., Fried, R., Gaffney, T., Haertel, G., Hafter, A., et al. (2003). *Design patterns for assessing science inquiry*. PADI Technical Report 1. Menlo Park, CA: SRI International.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., Haertel, G. D., & Penuel, W. (2003). Improving educational assessment. In G. D. Haertel & B. Means (Eds.), *Evaluating educational technology: Effective research designs for improving learning* (pp. 149–180). New York: Teachers College Press.

National Center for Education Statistics (NCES), U.S. Department of Education. (2003). *Internet access in U.S. public schools and classrooms, 1994–2002*. Washington, DC: Author.

National Research Council (2001). *Classroom Assessment and the National Science Education Standards*. Washington, DC: National Academy Press.

Office of Technology Assessment, U.S. Congress. (1995). *Education and Technology: Future Visions*. OTA-BP-HER-169. Washington, DC: U.S. Government Printing Office.

Olson, L. (2003). Legal twists, digital turns. *Technology Counts 2003*, 22(35), 11–14, 16.

Pearson Education. (2005, February). Advertisement for Progress Assessment Series. *ESchool News*, p. 9.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.) (2001). *Knowing what students know: The science and design of educational*

assessment. Washington, DC: National Academy Press.

Sadler, P. M. (1987). Alternative conceptions in astronomy. In J. D. Novak (Ed.), *Second international seminar on misconception and educational strategies in science and mathematics* (Vol. 3, pp. 422–425). Ithaca, NY: Cornell University Press.

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching*, 35(3), 265–296.

Schmidt, W. H., Raizen, S., Britton, E. D., Bianchi, L. J., & Wolfe, R. G. (1997). *Many visions, many aims: Volume II: A cross-national investigation of curricular intentions in school science*. London: Kluwer.

Schofield, J. W., Eurich-Fulcer, R., & Britt, C. L. (1994). Teachers, computer tutors, and teaching: The artificially intelligent tutor as an agent of classroom change. *American Educational Research Journal*, 31 (3), 579–607.

Shepard, L. (1997). *Insights gained from a classroom-based assessment project*. CSE Technical Report 451. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Shepard, L. A. (2000). The role of assessment in a learning culture. Presidential address at the annual meeting of the American Educational Research Association, New Orleans, April 26. Available at aera.net/pubs/er/arts/29–07/shepo2.htm.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Education Review*, 57, 1–22.

Stevens, R., Soller, A., Cooper, M., and Sprang, M. (2004). Modeling the development of problem solving skills in chemistry with a web-based tutor. In Lester, J. C., Vicari, R. M., & Paraguaca, F. (Eds.), *Intelligent Tutoring Systems*. (pp. 580–591). Heidelberg, Germany: Springer-Verlag.

Underdahl, J., Palacio-Cayetano, J., & Stevens, R. (2001). Practice makes perfect: Assessing and enhancing knowledge and problem solving skills with IMMEX software. *Learning and Leading with Technology*, 28, 26–31.

U.S. Department of Education, Office of Educational Technology. (2004). *Toward a new golden age in american education: How the internet, the law and today's students are revolutionizing expectations*. Washington, DC: U.S. Department of Education.

Vendlinski, T., & Stevens, R. (2000). The use of artificial neural nets (ANN) to help evaluate student problem-solving strategies. In B. Fishman & S. O'Connor-Divelbiss (Eds.), *Proceedings of the fourth international conference of the learning sciences* (pp. 108–114). Mahwah, NJ: Lawrence Erlbaum Associates.

White, B., & Frederiksen, J. (2000). Metacognitive facilitation: An approach to making scientific inquiry accessible to all. In J. Minstrell and E. van Zee (Eds.), *Inquiring into inquiry learning and teaching in science*. (pp. 331–370). Washington, DC: American Association for the Advancement of Science.