Accountability Bottom Up

Joan L. Herman

ACCOUNTABILITY systems in education typically are designed top down, with the expectation that they will stimulate action bottom up, from the school level. We all no doubt can recite the chain of logic in our sleep: the assessments communicate the standards and establish performance expectations. Through rewards and sanctions, schools and the educators within them are motivated to pay attention to student learning and use performance results to improve what they are doing, engaging in a continuous improvement process that will enable all children to achieve the standards. This is the view from the top down, but what of the chain from the bottom up?

Current accountability systems in fact mark a significant change in the traditional focus of evaluation responsibility and underscore important insights about where information must be applied if student learning is to be improved: local schools and their communities are the ones who must engage in evaluation and use the data to improve. It is at the school and classroom levels where students’ needs can best be understood and served.

While this is hardly a revolutionary idea in school reform, it is unprecedented in its expectations for local assessment and schools’ capacity to collect and use data to inform their instructional programs, especially in the sense of its distance from educators’ current assessment capabilities and practices. The limits of educators’ knowledge in the assessment and information uses area, in fact, have been fairly well documented, and despite periodic hand-wringing, assessment literacy has yet to be added as a staple of educators’ preparation and development programs.

How does this reality comport with expectations that undergird standards-based assessment and reform? In an op-
timal world, schools use the results of external or accountability assessments to obtain an accurate reading of how their students are doing relative to established content and performance standards and a measure of the progress students have made. In the process, schools acquire a picture of students’ relative strengths and weaknesses and a view of the relative strengths and weaknesses of the school program. But to truly understand why student performance is as it is and to get to the root of whatever teaching and learning issues may exist, schools really need to move to a more detailed level of assessment and analysis. They need to be able to supplement the external assessment results with other local data. After all, there is a definite limit to the amount of information that once-a-year assessments of limited duration—a typical mode for accountability assessments—can provide. No matter how well aligned and how sensibly crafted, these assessments can only offer limited perspective on what children really know and can do relative to standards and what factors may be working against their progress.

This is one of many important rationales for “multiple measures”—the need to view student performance from multiple perspectives in order to better understand and improve it. Moreover, if we really want to know how students are doing relative to the actual standards we have established, rather than just knowing how they did on a test(s), it takes multiple assessment perspectives. Take an obvious example: standards related to students’ communication. Certainly multiple-choice items and open-ended items cover part of the terrain, but don’t we really need to assess students’ writing? Different types of assessment reveal different aspects of performance that are important to a comprehensive measure of standards and important to understanding the learning of individual students.

Using a single measure for planning and improvement is not only insufficient, it can be dangerous. Absent a rich, local standards-based assessment system, with only one external assessment available for analysis, what’s likely to occur? First, some important standards are missed, because it is simply not possible to fully assess all our standards with a single measure. Furthermore, educators tend to focus on the test, not the standards, a trend well established in the literature. Worse yet, the school-based inquiry that current reform holds as so essential can devolve into a microanalysis of how students perform relative to the specific knowledge, skills and formats covered by the test. Implications for action become a curriculum of test preparation. Little wonder that if this happens, questions are raised about the validity of gains observed on high-stakes, external measures. Standards-based assessment as implemented in these circumstances falls far short of standards-based assessment as intended—a variant of the intended versus enacted curriculum issue.

How do we assure a picture closer to the vision as intended? The answer lies at least partially in coordinated systems of local assessments: district, school and/or classroom assessments that are aligned with standards and that can provide educators with the diverse forms of evidence that they need to understand and improve their students’ learning. Moreover, integrated with classroom curriculum and/or administered periodically over the course of the year, such local assessments are also necessary to provide teachers with essential ongoing information to gauge student progress and adjust teaching and learning opportunities accordingly. Ultimately, these are the “multiple measures” that really can make a significant difference in student learning. Good teaching is a process of continual assessment and adjustment—waiting until the external results to show up annually or even semiannually just is not enough.

There is no single model or easy solution for accomplishing such local systems, nor a single recipe for what they should look like. Depending on local context, history and philosophy, as well as the talents and dispositions...
APEC 2000 International Assessment Conference
New Directions in Student Testing and Technology
Anne Lewis

The assessment of students, and often of schools, is in an international process of change, agreed most representatives attending the APEC 2000 International Assessment Conference, New Directions in Student Testing and Technology. Depending on the goals in each economy, assessment reform may lean toward decentralization or away from it, integrate different goals for learning or retain traditional ones, or be at different stages of using new technologies. Nonetheless, as Alan Ginsburg, U.S. Department of Education, noted, “the Asia-Pacific economies attending this conference hold many common issues and a desire to learn from each other’s efforts.”

Ginsburg welcomed representatives from 12 economies as far away as Singapore and Chile and as nearby as Mexico and Canada. Held at the University of California, Los Angeles, from October 23-25, 2000, the conference was co-sponsored by the U.S. Department of Education, the Ministry of Education of China, and the Center for Research on Evaluation, Standards, and Student Testing (CRESST) at UCLA. By meeting with colleagues from around the world, “we learn that some of the problems we thought were unique are shared among us, but also that people in other economies come at issues in different ways,” Robert Linn, CRESST co-director, observed at the opening of the conference. Among the common issues discussed by representatives of the 12 economies attending the conference:

- Resolving the tension between education reforms and traditional testing. The former often focus on problem solving and interdisciplinary learning, while the latter is still used in many economies for college selection and thus leads to a narrowing of the curriculum.
- The tension between national goals for assessment and desires for decentralized control of instruction.
- The extent to which new technologies can be used to assess learner outcomes and school performance.
- An uncertainty as to whether current assessment practices can “capture” evidence of desirable traits, such as creativity.

As a leadoff to the conference, Linn presented some of the assessment challenges CRESST is addressing in the United States. Beyond the diagnosis of instruction and feedback for students and teachers, assessment is being used to certify that students ought to move on to the next grade or graduate from high school and to hold schools and teachers accountable. This requires that assessment systems meet three criteria—validity, fairness, and credibility.

As in other economies, educators in the United States face the fact that “what you test is what you get,” Linn said. This is why CRESST recommends that standards be developed first, then curriculum, professional development, instruction and assessment, all aligned to the standards. Putting assessment systems in place before the standards are complete may create unintended consequences such as a narrowing of instruction, inflated gains of assessments, higher rates of student academic failure and school dropouts.
Some of the issues raised by Linn might be helped by emerging technologies, according to CRESST Co-director Eva Baker. A searchable database can store test content domains and thus facilitate improved alignment between tests and intended content. Technology also can help design tests that build in cognitive, linguistic, and content sampling requirements. Computer-supported test design, Baker said, can be fast, provide better sampling and make test development rules so explicit that they are accessible to teachers. Showing examples of computer-generated knowledge maps in different languages, she said teachers would be able to evaluate students’ declarative and procedural knowledge, allowing teachers to align what they are doing with what is assessed.

Given the issues outlined by Linn, “some of the computer technology as it exists now could provide deeper assistance in improving the quality of assessments and in helping people see that assessment is a productive part of the education system, not the most feared,” Baker said. Computer-assisted assessment can present data uniformly and enhance the validity of assessments, she added.

“…some of the computer technology as it exists now could provide deeper assistance in improving the quality of assessments and in helping people see that assessment is a productive part of the education system, not the most feared,” Baker said.

Largely because of technology, educators can anticipate having even greater sources of data about student and institutional performance, noted Marshall Smith, former under secretary of the U.S. Department of Education. Elaborating on assessment activities in the United States, Smith discussed three uses for data—to improve teaching and learning; to hold students, teachers, and schools accountable for performance; and to inform policymaking.

Data from cognitive sciences research on how people learn will help frame the curriculum and provide diagnostic tools, Smith told the conference. Moreover, benchmarks such as student work will affect instruction. Data that drive high-stakes accountability in the United States, however, “need to be brought under control,” he said, because often the testing policies do not meet the criteria of validity, reliability, and fairness. As to informing policy, data from sources such as the Third International Mathematics and Science Study (TIMSS) may not be definitive but can provide ideas and concepts useful in different economies.

Referring to the link between assessment and technology, Smith used several examples such as distance learning, which can provide feedback on both student achievement and effectiveness of instruction, and computer tutoring, which can provide data useful for teachers such as the frequency of certain errors. Smith urged the conference participants to consider “different sets of lenses” to view student success. He suggested that the United States has sustained an enviable economic boom despite an education system described as “at risk” because the system encourages second or more chances, not a single event like a test, to determine a person’s direction in life. The “tremendous creativity” in American society is due to multiple chances for people to succeed, noted Smith.

This attention to the creative spirit echoed in remarks to the conference by Alan Kay, vice president of research and development at the Walt Disney Company and designer of the prototype for the desktop computer. Kay said that learning is a product of innovative creativity, something that is not captured in traditional tests. Nor are technology and its vast resources for acquiring knowledge guaranteed to produce breakthrough thinking. The Internet, now with 350 million users, can at best provide access and networking for the handful of people who can do more than solve a math problem, the few “who can answer the why of a question,” he said.

The “tremendous creativity” in American society is due to multiple chances for people to succeed, noted Smith.
For Randy Bennett of the Educational Testing Service, however, the Internet is the technology that will do the most to encourage radical improvements in assessments. It will soon allow assessment to be interactive, networked, and standards-based, producing quantities of information through broadband transmission. He envisions the business side of the Internet’s use in assessment to be able to develop tests, deliver them, and present them through dynamic stimuli such as audio, video, and animation. “Assessment will go directly to the desktop,” Bennett predicted, and eventually will be embedded in instruction, serving both institutions and individuals. Ultimately, assessment will be customized for the user.

Technology may harbor great changes in assessment, but meanwhile, most APEC economies are designing assessment systems that respond to their current education reforms.

**People’s Republic of China**

The return of the National Entrance Exam after the Cultural Revolution brought back intensive focus on a single test in the People’s Republic of China, according to Chen Qi of Beijing Normal University. Changes in the entrance exam focus on improving education in secondary schools and on giving campuses greater autonomy. A subsequent national policy has directed schools to pay more attention to the quality of education all students receive rather than their rank on the national exam. “We have achieved a certain amount of progress,” Chen Qi said, “but old habits are slow to die, and reforms must address changing the philosophy and pedagogy among policymakers, exam paper makers, and teachers.” The People’s Republic of China is using technology to make test item banks available and to facilitate the university admission system in isolated areas.

**Indonesia**

Indonesia’s new national exam system is a school-exit test given at the end of the primary grades, Grade 9 and Grade 12. It is used both as an indicator of quality at the school level and as a screening tool for schools with limited space, explained Bahrul Hayat of the Indonesia Ministry of National Education. The national exam is public, with data available on school performance by subject each year. Teachers participate in developing test item banks and help decide what knowledge and skills will be covered on the national exam each year. Hayat said that feedback data from the exam have expanded in the past two years, and now include content area analysis such as simple tables describing the percent correct per item at school, district, province and national levels. The economy has a national curriculum that guides instruction in core subjects and about 50% of other subjects, although implementation in classrooms is not precise.

**Canada**

The province of Alberta, Canada, charts progress at Grades 3, 6, and 9 through an assessment that is linked to standards. Using 1996 data as the baseline, an important aspect of the system is the feedback to schools, explained James Dueck, assistant deputy minister of education. He said that analysis of data from the assessment has led to a different perspective, one that holds schools responsible for learning. The focus on accountability also applies to the ministry, where every department has a performance plan and progress on the achievement tests “is a significant proportion of the bonus plan for employees at the provincial level.”

**Australia**

In Australia, the eight state and territory governments have primary responsibility for schooling, including assessment. Rather than push for a national test, the Commonwealth has built upon existing collaboration among the states and territories to frame new national goals and establish national literacy and numeracy standards. These performance standards, known as benchmarks, represent important and essential elements of literacy and numeracy.
Di Weddell of the Australia Commonwealth Department of Education and Peter Titmanis of the Western Australia Education Department described how a sampling procedure has been used to equate the tests from the different jurisdictions. Expert judges, using an item pool from all of the tests, developed a common scale based on level of difficulty and a cut score on the scale. This allows each jurisdiction to calculate the proportion of their students who had a score at or above the benchmark on their own scale. The sample approach won over state officials, according to Titmanis.

**Chili**

A new assessment in Chile measures higher order thinking skills, using both multiple-choice and extended response formats. Given for the first time to fourth graders in 1999, it will expand to eighth grade soon and is intended to assess school performance, according to Claudia Matus of the Chili Ministry of Education. The first year’s data revealed that student performance in public schools and in subsidized private schools was not significantly different. Officials are dealing with issues of reliable scoring on open-ended questions, generating appropriate items for higher order skills, and informing the public.

**Korea**

Korea has three types of assessments: school activity records that evaluate students; a National Assessment of Educational Achievement, which controls the quality of the system by providing information on overall educational achievement; and a college scholastic ability test administered nationally. The reforms advocate replacing a norm-referenced system with a criterion-referenced system and using student records for diagnostic purposes as well as a cumulative evaluation.

**Singapore**

Unlike Korea’s homogeneous population, Singapore is multi-cultural with four official languages; English is the primary language of administration. Singapore has gone through several education reforms with the latest reform characterized by five features: bilingualism (English and a mother tongue); streaming, the placement of students in different levels at primary, secondary and pre-university points; special curriculum for each stream; curriculum and assessment in two languages plus math up to Grade 10; and compulsory but nontested moral, physical and national education through Grade 12. According to Quek Choon Lang and Toh Hoon Sin of the Singapore Ministry of Education, the assessment system consists of school-based assessment by teachers and national exams at Grades 6, 10, and 12. Assessment data are used to determine promotion to the next grade and streaming (placement), and as feedback to pupils on their readiness for the national exams.

**Thailand**

Based on 1999 reforms, Thailand’s new assessment framework will be instituted in 2002. The reforms decentralize authority and shift assessment responsibility to schools,
teachers, parents, and students. The revisions introduce multiple assessments and alternative assessments, such as portfolios, and allow credit transfer between formal and informal education settings. According to Boonchoo Chalassathien of the Thailand Ministry of Education, officials have identified five areas for future research: assessment of moral education, which is taught in many Asian economies; grading of essays, especially online; methods to discourage excess test preparation; item bank development across several economies; and improved school and student data reporting to the public. The focus for this last area is on how reporting can provide important information but avoid comparisons and stratification given that each school has its own characteristics.

**Closing Remarks**

The conference produced sufficient information and interest to justify follow-up activities, in the opinion of Martin Orland, Office of Educational Research and Improvement at the U.S. Department of Education. “There is enough commonality of interest and enough unsolved problems to continue the work,” he said. International perspectives are important not as ways to copy ideas but to infuse policies “with new ways of thinking.”

Beyond immediate activities such as a clearinghouse and networking that promotes a regular exchange of information among researchers, developers, and private companies, Orland discussed longer term research needs focusing on standards, assessment, accountability, and teacher quality. In the area of assessment, international discussions could elaborate on whether existing tests are measuring what really matters, the relationships among multiple measures, issues of validity and data use. Economies could share their ideas on best uses of technology. Finally, the economies could benefit from knowing how effective their various reform efforts have been.

The participants acknowledged that some activities, such as item banks, may not be fully useful because of the differences in education systems, but they did endorse continuing the conversation. “Even though every economy here has its own advantages and experiences, we found that we have a lot in common,” observed Chen Qi (Beijing Normal University). “Further exchanges are needed,” she said.

In addition to a clearinghouse of information, Ginsburg said the discussions suggested the need for the creation of communities of practice across the economies represented, including practitioners, government officials, and researchers, who are working on common areas. He announced that Singapore had offered to take the lead on a clearinghouse for instructional technology.

The United States and China will present a report and recommendations on follow-up activities at an APEC meeting in Mexico City next spring, announced Ginsburg. A priority will be to continue much of the valuable work begun at this meeting.

---

**CRESST Conference 2001**

**Save the Date**

**September 13-14, 2001**

Look for more details soon on the CRESST Web site: www.cse.ucla.edu
of teachers, some districts or schools might take a rather top-down approach. For example, after suitable dialogue and consensus building, such a district might mandate specific curriculum-embedded or other assessments that teachers should be using quarterly in their classrooms to assess student progress. Such mandates could include portfolios with specified kinds of tasks in them, and on-demand performance assessments to supplement what’s available from the state system.

Another district or school might take a more bottom-up approach. Here the approach might be to have teachers work with standards to develop standards-based curriculum and assessment units, then share them with other teachers, or let teachers choose from a number of alternatives that fit the individual priorities of the district. Or such a district might provide a general structure for a portfolio and then let teachers decide what goes in and what meets the standards.

In one district, the measures might be largely formal ones, in the sense of being able to be aggregated up and summarized numerically. In another district, the mix might include more qualitative measures—observation measures of students during discussions, or presentations of student projects.

These are just a few of the alternatives. The bottom line is that local assessments really matter and should count too. They need to be standards based, compatible in approach with their state assessment systems, and developed through a process that builds consensus and agreement on what students need to know and be able to do. That local assessment also should utilize multiple measures is axiomatic—for the reasons noted above, but also to respond to the reality of individual differences and provide all students a real opportunity to show what they can do.

Local assessment systems represent an ambitious agenda in their development and in the teacher assessment capacity necessary to bring them to fruition. We hope that CRESST research over the next five years will contribute importantly to that agenda.