CRESST 2004 CONFERENCE
Research Guidance: Assessment, Accountability, Action!

CSE Report 658

Anne Lewis

July 2005

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was partially supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute for Education Sciences, U.S. Department of Education.

CRESST 2004 CONFERENCE

Research Guidance: Assessment, Accountability, Action!

Anne Lewis

The 2004 CRESST conference, *Research Guidance: Assessment, Accountability, Action*[1], confirmed that the research community is well poised to guide educators and policymakers into the more sophisticated use of test-based accountability. The September meeting of 200 researchers and educators looked more to future possibilities than to past missed opportunities in a program that focused on accumulating knowledge about assessments, especially their effect on teaching, and refinements of accountability systems. The two years of experience in implementing No Child Left Behind across the country provided the undercurrent at the conference for discussions and debate about how research can guide assessment and accountability.

With its "even-handed approach to analyzing problems," CRESST has done a remarkable job "in bringing expertise and engagement with the field into how to make NCLB work," said Aimee Dorr, dean of the UCLA Graduate School of Education & Information Studies at the beginning of the conference. For the next two days, that expertise was on display. The plenary panels set the research in broad contexts of both practice and policy.

The conference began with "the charge and the challenge," a discussion on moving to the next generation of NCLB. Leading off, CRESST Co-director Joan Herman focused on accountability that supports student learning. In some ways, this is the best of times, she said, because "there is unheralded interest in assessment and a belief in the ability of assessment to improve schools." Most state assessment systems now include both multiple-choice tests and some type of performance assessment, oftentimes an essay. Additionally, conversations between legislators and researchers have improved. "We are not yet at the point of having great accountability systems, but we are on the road," she said.

---

[1] Although we have made every reasonable attempt to include all CRESST conference presentations, a few may not be mentioned here due to technical recording problems. Our apologies to those authors whom we may have missed. Please see the CRESST web site, http://www.cresst.org for additional information about this conference and other CRESST conferences. Copies of overheads from many of the conference presentations and abstracts of all the presentations are also available on the CRESST web site.

On the other hand, it is the worst of times because expectations continue to outstrip the capacities of existing accountability systems. Further the rates of improvement expected of schools and districts appear unrealistic.

To be specific, the theory of action behind current policies is to motivate the public education system through high standards and then hold schools accountable. The process is to establish standards; develop measures; set performance goals; leaven with incentives and sanctions; then count on schools to make the grade. To be effective, the theory assumes a technical system where assessment results provide accurate and valid information at multiple levels; data are well used to inform planning and decision making; and the educational system uses the data well to engage in continuous improvement.

The good news is that districts and states have acted to align their curriculum and assessments with their standards. Teachers are listening to the signals sent by accountability and are focusing their instruction on the learning goals that they think are expected of students. The not-so-good news, according to Herman, is that the curriculum is narrowing; there is the possibility of a dual curriculum growing in the schools; and the morale of teachers and principals may be undermined as growing numbers of schools fail to make adequate yearly progress goals.

Herman asked two critical questions: How can we design standards and assessment systems that optimally focus instruction, but don't stifle it? And, what happens to motivation when schools hit a wall and performance rankings level out as they did in California this past year? The latter issue may encourage schools "to give up," but there is hope if the problem is seen as a technical issue, she said.

So, then, the question becomes: Is the technical system working? The federal government is asking states for the right kinds of validity evidence to assure the quality of their tests, such as evidence of alignment, reliability, and accuracy. And some results are promising. For example, David Rogosa's analysis of the accuracy of California data finds a 75 percent probability that students are correctly classified in proficiency level by the state's standards-based assessments. Still, according to Herman, some note that reliability and year-to-year score fluctuations are a continuing challenge, the validity of gains may be suspect, the feasibility of Adequate Yearly Progress (AYP) targets is problematic, and there is a thorny technical issue in some urban schools needing to meet as many as 40 targets of AYP. Two other questions need to be asked: Will the AYP

methods identify the right schools? And, is there a better way to look at data and validate results?

Alignment is the critical lynchpin, in Herman's opinion, but the evidence to date produces concern. In a recent study, for example, 20 content experts rated each item of the now defunct California Golden State exam as to the topic addressed, the depth of knowledge required, and the content centrality. A clear majority agreed on what topic was addressed by 35 of 42 items; but agreed on both the topic and depth of knowledge for only 22 of 42 items; and agreed on the topic, depth of knowledge and content centrality for only a small fraction of the items.

The question arising from such a study is how can teachers teach to "standards" if experts do not agree on what topics and depth of knowledge mean, she asked. Another example—of math items on successive annual tests in New York— illustrated that alignment also is a moving target. If we want teachers to teach to the standards rather than the test, Herman asked, "How can we be sure that developers and teachers share the same understandings?" Also, "how can we specify standards and assessments in ways that clearly communicate expectations without unacceptable curriculum narrowing?" Finally, "can we frame expectations in ways that are feasible for all students?"

Another important research issue relates to teacher assessments, Herman said. "If we believe in the power of assessment to support learning," she concluded, "we must put that power where learning occurs—in the classroom." There are many classroom assessment capacity challenges, including the fact that teachers tend to not have a learning-oriented perspective. But there also are some realistic remedies, added Herman, including better teacher pre-service preparation and an insistence that test developers include high quality materials for teachers. The research community has a lot of work to do to provide helpful research guidance on improved classroom assessments.

From Unrealistic to Doable

Robert Linn of the University of Colorado), and another CRESST co-director, looked at the next generation of accountability under NCLB. NCLB has several praiseworthy aspects, according to Linn, including an emphasis on improving achievement of all students, especially those with the greatest need. It emphasizes

closing the achievement gap, encourages rigorous content standards, and puts an emphasis on qualified teachers.

The NCLB accountability system, however, has fundamental problems that could undermine its achievement goals. Central to the accountability system is the definition of adequate yearly progress, which Linn compared to a standard that everyone must be able to run a five-minute mile. "Some people can do it right away, some are never able to," he said. Linn discussed five problems that he considers serious:

- *Unrealistic expectations.* The AYP targets are set so high that all but the most selective schools will fail to meet goals within a few years. To reach 100 percent proficiency levels on NAEP math by the year 2014, for example, the rate of gain at the 4th grade would have to be 2.3 times what it is now. The needed rate for NAEP reading gains, which has been flat for several years, would be even steeper. In California, projections show that 98 percent of the schools eventually will fail to meet AYP. Linn recommended that Congress set more realistic, but still ambitious, goals. A more realistic goal for instance, would be to use the rate of gains made by the top 10 percent of schools like, for example, 3 percent a year averaged over five years. That rate of gain would then become the AYP target for all schools. This would meet the concept of *existence proof* , said Linn, meaning that at least some schools have reached a target, proving that the goal is attainable.
- *Definition of proficient achievement.* Because states set their own starting points, "they are scattered all over the map," Linn said, explaining that the beginning reading proficiency benchmark is at about 13% in California whereas Colorado's is 77%. Proficiency goals are not in a straight line, and most states face a "balloon payment" in the future. The variation among states makes little sense, he said, because on NAEP, the state scores are very close. Linn also suggested that states should determine the 5% of schools most in need of improvement and focus resources on them. There will not be enough extra money to help every public school in the nation that ends up on a needs improvement list.
- *Disaggregation rules and effects.* Schools with multiple sub groups are at a relative disadvantage compared to schools with more homogeneous

enrollments. A modification might be to combine scores across more than one year.

- *Safe harbor.* Allowing schools to meet AYP if a sub group fails to meet its target provided the percentage of students who score below the proficient level is decreased by 10% "is a nice idea," Linn said. The gains needed, however, are very large, and few schools have benefited from this provision.

The AYP goals will become increasingly unrealistic once high achieving schools miss their targets, Linn said. Consequently, an existence proof system is a reasonable compromise. Other key needs include a common definition of a proficiency standard, sub group reporting that allows for random year-to-year fluctuations, and revisions to the safe harbor criterion to consider the gains actually achieved by high performing schools. Finally, Linn said that AYP should consider gains in achievement as well as status, and should allow states to use longitudinal student data to evaluate gains.

The "real life" of a test developer in Iowa under NCLB has been to establish a "reluctant" statewide testing program, Stephen Dunbar of the University of Iowa, told the conference. Standardized tests in all subjects have been used in the state for about 70 years, but they had never been mandated and were never funded directly by the state. The emphasis was on their use as information to inform and improve instruction.

NCLB has changed much of that, Dunbar said. Iowa is exempted from the federal mandates on state standards, but has a reporting requirement that districts develop their own standards. The assessment system for NCLB is now two-tiered straightforward reporting on the Iowa Test of Basic Skills (ITBS) and the Iowa Test of Educational Development (ITED) for federal accountability and reports from school districts to the Iowa Department of Education on tests aligned to their own standards.

"This means that a 'little' statewide testing program run out of the University of Iowa is now trying to develop test forms useful to each district," Dunbar said. "I have participated in more alignment work lately than I care to talk about, but alignment is critical." A particular challenge is the distribution of test materials, which in Iowa eventually may mean as many as 300 different test booklets for each of 6 or more grade levels.

Another real life experience recounted by Dunbar is his work as part of a team that is providing test materials for other states. The process includes discussions and

negotiations with federal officials, legislators, and others in which "alignment is the big issue." However, "often by the time we get to the end of the process, the rules of the game have changed either because of negotiations with the federal government or state budget cuts," he said. Alignment can then become an afterthought and hence a by-product of negotiated decision-making that involves other criteria for defining the assessment system.

"In all of the assessment development I have participated in," Dunbar said, "where resources are limited, inevitably someone treated standards as a checklist, going down the list to see if the test has one of these or two of those. That process is not a good one."

Because of the magnitude of state assessment systems, seldom can a single contractor can handle all of a testing program. Contractors may be partners in one state and competitors in another state. "This kind of dynamic in the testing industry is not understood, nor do we know its impact on test validity and on kids," Dunbar said. But its implications for timelines and reporting back results are known. Dunbar added that there are only a few states in which test development work is ongoing. "For all intents and purposes, state programs are now fixed in terms of design and instrumentation," he said, "but the successful delivery and valid use of results isn't."

One impact of NCLB on large-scale assessments has been the recycling of test items. Two years ago at a CRESST conference, Dunbar predicted that the law would lead to "no test item left behind," and that is still a concern of his. "If an item is a good item, that's not a bad idea," he said, "but it's not good if the item is not good." Still, because NCLB does not allow test developers to take a test off the shelf and put a state's name on the cover, NCLB may be contributing to better tests over time. NCLB also has focused attention on more cost-effective ways of developing and delivering test materials with innovative technologies. Such work is in its infancy in K-12 assessment, but is an example of how NCLB may move testing well into the 21st century.

Spread the Responsibility

The impact of a fully successful NCLB has not been considered, Scott Marion of the National Center for the Improvement of Assessment, Inc., reminded the conference. The current postsecondary structure could not accommodate the increase in the college bound, the economy cannot absorb many more people competing for high-wage jobs, and the prospect of a capitalistic society in which there is no basis for wage stratification

would require a new economic model. Still, Marion said, "we cannot walk away from school accountability just because we don't like how the discussion is being framed."

His vision for an accountability system is focused on meaningful student learning and attainment, but is based on distributing responsibility for these outcomes among the many institutions that affect learning. Students should be accountable for documenting learning beyond merely passing tests, parents should be persuaded to support learning goals, and schools should engage students in "profound interactions with content and processes," as Phillip Schlechty has written. Unfortunately, according to Marion, NCLB has created testing pressures that "have likely distracted school leaders from pursuing more ambitious performance-based reform efforts." Schools need to be held accountable for promoting deeper learning and for assessments that measure it effectively.

While schools, more than teachers, are responsible for achievement-related accountability, teachers should be responsible for continually improving their craft and accepting evaluations based on learning goals, Marion said. Districts need to be held accountable for student achievement, but Marion also outlined some important intermediate goals for districts—informed curricular focus, research-based professional development, and development of a local assessment system that provides teachers with more timely and relevant feedback. This last item, Marion said, "is the most powerful way of improving teacher practice."

Also accountable for heightened student learning are teacher preparation programs, which need an overhauling; state departments of education, which need help in moving from compliance monitoring to supporting teaching and learning; and local and state school boards, whose performance should be linked with achievement trends. Finally, the whole system needs "good economic models to provide guidance about costs for educating all students to a proficient level." He recommended, somewhat facetiously, an experimental design in which one-third of the schools increased teacher salaries considerably, another one-third increased the number of teachers, and the remaining one-third served as a control group. With data from this design, "we could say to a legislative finance committee: It is going to cost x, but you've only appropriated this percentage of x…"

Considering factors outside of the school/teacher unit means that the focus includes school delivery standards, Marion said, and implementing these standards often "is contingent upon the resources provided by the state and federal

governments." Marion advised continuing to "squeeze" as much out of the current educational system as possible, "but if you believe as I do that the potential for the current system is limited, we need to continue to work to build comprehensive educational systems that can truly support meaningful teaching and learning."

Toward Multipurpose Test Design for Accountability and Improvement

CRESST has been working for 15 years on multi-purpose test designs, a lonely endeavor, when the work began, according to Eva Baker, co-director of CRESST at UCLA. The work has evolved into a model that uses cognitive demand "as the crosswalk between standards and assessment." Baker, however, talked about a middle path, "the place between modal and model."

Taking a page from systems theory, CRESST's work is a process that identifies high-quality objectives, builds appropriate measures, sets targets, and supposedly designs aligned instruction. It needs to build concomitant capacity of the users so that teachers, students, and administrators understand the inferences they are expected to draw; that is, they use the data for different purposes, including accountability, diagnosis and improvement. The theory of accountability in current reforms anticipates that people will be able to analyze the strengths and weaknesses of assessment results using systems or support environments such as CRESST's quality school portfolio.

This view of multipurpose assessments was developed early and is continuing in technology-rich environments, Baker said. The model focuses on different kinds of cognitive demands that are then embedded in different subject matters as appropriate. The idea is that the task structures can be used either for instruction or for assessment. An improvement on the original is the capability now to "get very explicit in expressing cognitive demands and content," Baker explained.

The cognitive demands in the assessments drive the teaching/learning situations, she added. However, the "granuality" provided by such assessments is helpful to teachers only if they have alternatives for action. "It doesn't help to understand more and more if you have no ways to act," Baker said. The transfer feature is critical because there needs to be assurance that the full domain of learning has been acquired, not simply practicing test items. We also need to draw inferences about what is causing achievement, which leads to the need for scalable measures of classroom practice. "We don't have much to go on when it comes to knowing what's going on when the classroom door is shut," she added.

To date, the model-based assessment system has shown that it can be produced to scale inexpensively and that teachers have changed their instruction with assessment helping to produce those changes. There is hope that it is possible to develop assessments using this model and the theory behind it, Baker said, "but the models have been battered by constraints and schedules." The modal practice in schools is a test-driven process, not a standards-driven process, so "the theory-based model is found practically nowhere." The lockstep curricula provides no space or time for alternatives based on classroom assessments.

"For poor kids, things are better than they used to be," Baker concluded, but there is uneven performance in high schools, and NAEP does not reflect much national improvement. In the worst case, she said, "we will have a two-tiered system in which smart kids get augmentation, not necessarily in school." A better case is that "we don't fold. We argue that there is evidence for test-based curriculum." Perhaps, "we may have to take a step back and come to a notion of what is the best we can do with the constraints that we have."

Robert Mislevy of CRESST and the University of Maryland approached the subject by discussing intuitive test theory as contrasted with "scientific" test theory. The former is everyday reasoning about tests, and focuses on items and scores. The latter sees assessment as evidentiary reasoning, and it uses probability models to connect "what you see with what you make inferences about," he said. It applies less familiar concepts such as measurement error and techniques such as item response theory.

Most reasoning is based on everyday experiences, and everyday test theory works fine for classroom quizzes, Mislevy said. But it fails for more complex assessments such as adaptive testing, national assessments, and linking results from different tests. "It is our responsibility as education experts to understand how others view testing, including parents and policymakers. We must help policy makers design systems that can actually do the job they are meant to." To do so, Mislevy tackled several popular beliefs about testing.

The first level of common thinking about testing declares that "a test measures what it says at the top of the page." Another belief says that any two tests that measure the same thing can be made interchangeable, with a little equating magic. "That would be nice if it were true," Mislevy said. It is also commonly assumed that "a score is a score is a score," or "don't give me any of this measurement error stuff" even though they do matter when tests are high stakes. Another misconception is that multiple

choice questions only measure recall while in fact, as Mislevy as showed through several examples, they can require a depth of content knowledge. Finally, there is the belief that technology will solve testing problems by making it possible to get voluminous amounts of data. It is not volume of data, but sound design which is needed. He illustrated this point with Cisco Learning Institute's NetPass prototype, which creates online performance assessments of networking skills.

Mislevy said these various beliefs show that assessment policy is often based on intuitive theory, which can lead to frustration and anomalous results. He noted three ways of dealing with intuitive theory: tell people what they've already planned is wrong and why ("the not-fun way"); be involved in project planning early on ("the good way"); or the "good and fun" way which is to be involved in existence-proof projects that demonstrate what can be done, outside familiar forms, using scientific test theory.

Using examples of open-ended items, Lorrie Shepard of the University of Colorado at Boulder and CRESST contended that good assessment tasks are interchangeable with good instructional tasks, and that large-scale and classroom assessments share many characteristics (as well as have some major differences). The "Knowing What Students Know" approach illustrated in the National Research Council report three years ago stresses the importance of coherence, if an assessment system is to support learning. Therefore, it is important to consider the shared characteristics of large-scale and classroom assessments. The shared model includes developmental progressions over time and the same conception of disciplinary knowledge and competence. The latter includes a focus on evaluating understanding, and reasoning; offers a clear vision of what constitutes mastery; targets both general and domain-specific forms of cognition; and selects complex and context-based tasks that are open to multiple approaches. The Delaware Comprehensive Science Assessment and The Queensland, Australia formative assessments were offered as examples of assessments linked to curriculum standards that could be used for both formative and summative purposes.

Large-scale assessments provide data for monitoring and accountability, but can serve other purposes, Shepard said, including program diagnosis and professional development targeted at topics students are not mastering according to the assessment data. She cautioned against relying on what test publishers consider alignment with curriculum standards, which is limited to what fits within test blueprints. Rather,

according to Shepard, the idea should be to look to "more complete and substantive alignment that occurs when the tasks, problems, and projects in which students are engaged represent the range and depth of what we say we want students to understand and be able to do."

Assessment reformers have used titles such as authentic, direct and performance assessments, all meant to convey the idea that assessments must capture real learning goals, Shepard concluded. Embodying worthy goals is still a core issue, she said. The Knowing What Students Know ideal system of coherent conceptual models behind both large-scale and classroom assessments "will require much more than token constructed-response items on accountability tests."

Commenting on the presentations, discussant Robert Glaser of the University of Pittsburgh said that the presenters in this session were in "high agreement on the need for an assessment system to support learning, though they couched their views with different contexts and problems." Glaser proposed considering the name of formative assessment in contrast to informative assessment. "We have to ask," he said, "what kind of information are we presenting to parents, students, and teachers on the basis of this assessment."

Appropriately designed assessment situations can have substantial impact on the quality of information provided to teachers and students for instructional decision-making and meaningful learning. Informative assessment can include teaching and learning by communicating learning goals, interpreting student performance, tracking progress over time, and suggesting appropriate directive actions. The necessity to develop tools is compelling in the context of current demands for standards and accountability. In this context, a significant influence is the renewed interest in uniting the fields of cognitive psychology and psychometrics to comprise assessments of performance achievement and competence.

We focus on the design of assessment situations in the course of classroom learning, and concentrate on informative assessment procedures. We use the term informative to refer to measures that can provide a view of the learning goals, information about the present state of the learner, and action to close the gap. The term "formative" has been used to refer to a placement of assessments during the course of an instruction unit, while use the label "informative" draws attention to the instructional purpose to improve student learning. Further, assessments can be informative of various aspects of achievement for various audiences. All assessments

can be informative in some way; the key issues are of what, for whom and how these measures inform. Glaser concentrates on the information given to teachers and students to facilitate teaching and learning. He also adheres to the general framework for assessment design put forth by the Committee on the Foundations on Assessment (NRC, 2001). This committee identifies three key elements of assessment: cognition—theories about learning and performance (and targets for assessment); observation, tasks used to illicit information about learning, and interpretation for methods for scoring and validating assessment results. Glaser's particular focus is on how these elements are presented to teachers and students in instructionally meaningful ways.

Glaser focused on several properties of assessment design that enable teachers and students to describe progress in terms of "cognitive" features of performance, and then act on that information to improve learning. Operationally, he intends to suggest design elements for situations that maximize the information provided about student performance and competence—particularly in terms of criterion activities and standards— and de-emphasize information provided by normative reference to group standing.

There is still more concern with normative standards than with performance and competence, Glaser said. If assessment is to inform instruction and teaching, "it ought to concentrate much more on self-regulation and meta-cognition than enable students to look at their own performance." Glaser also suggested the need to measure how a student represents a problem; the need to look at students' access to rules for performance; and the need to see how assessments capture the learning that takes place on the basis of existing knowledge.

Technology in Support of Learning and Assessment

States are pursuing online testing at multiple grade levels, in all key content areas, for a variety of populations, and with both low and high stakes, according to Randy Bennett of the Educational Testing Service. States are turning to online tests, he said, because of the speed of the scoring and reporting, the ability to customize tests to the skill levels of the individual student, the promise of being able to measure things that can't be measured on paper, and an eventual reduction in costs. There are some major issues, however, including the near-term costs; timelines for development; equipment, software, and network availability and dependability; security; and measurement and fairness. According to Bennett, a critical measurement and fairness

issue concerns comparability; that is: "Do the scores mean the same thing if we use the computer?"

Comparability relates to whether scores mean the same thing across different conditions, like the delivery mode (paper vs. computer), the characteristics of the computer platform on which the student takes the test (large screen vs. small screen), and the manner in which constructed responses are scored (on screen vs. on paper). With respect to delivery mode, for example, the student taking a paper test may encounter several items per page, while on computer only a single item per screen may be presented, and this difference in presentation may affect performance.

Comparability is important, Bennett said, when scores need to have common meaning with respect to one another, some reference group, or a content standard. If the scores are not comparable across conditions, decisions may be wrong with regard to, for example, promotion or graduation, diagnosis or learning progress, school effectiveness, or group proficiency.

For the foreseeable future, those states that introduce online tests will need to supplement them with paper versions. Paper versions are still needed because not all schools have enough computers, and some students don't yet have adequate computer skills. In such situations, comparability will be important.

According to meta-analytic research on adult test-takers, scores from computer tests are generally comparable to those from paper tests for multiple-choice examinations like the ones used in educational settings. A study using data from the GRE, GMAT, SAT 1, Praxis, and TOEFL found that the delivery mode "consistently changed the size of the differences between some groups, but only by small amounts," Bennett said. For performance tests, like those in essay writing, however, a study of adults found that they scored higher on paper than computer tests.

At the K-12 level, several studies have found that students scored higher on reading and math tests given on paper than in the online versions, but also only by small amounts. In writing, a NAEP study found no significant difference in mean scores for 8th-grade students between paper and online versions, but facility with computers predicted online test scores, suggesting that the scores were not comparable across delivery modes.

Bennett recommended increased research on identifying likely sources of irrelevant score variation, publishing the results in peer-review journals so that they

could be vetted by the field, and equating scores if the rank orders were very high but the distributions did not match. In conclusion, Bennett said that score comparability may be affected by variation in delivery mode, among other things, with potentially undesirable consequences. Education agencies must increase their efforts to study the impact of variation in delivery, he said, and take steps to manage sources of variation found to affect performance. Without serious attempts to do these things, "we may jeopardize the very benefits that we hope to realize in applying new technology to assessment."

Greg Chung of CRESST/UCLA described several studies under way by CRESST to learn more about distributed learning and technology, and how to scaffold more complex learning from assessments. Distributed learning ranges from traditional online courses to many users taking many courses, any place, anytime. Distributed learning is learner centered and requires more autonomous learners because of less instructor and peer support. This anytime, anywhere instruction also means anytime, anywhere assessment, Chung said. The challenges are to decide what to do with the information, what information to use, and how to develop automated reasoning support "to infer what is going on." Just counting clicks "is not where we want to be."

With tasks involving modifying a bicycle pump, determining the parentage of a person, or knowledge of rifle marksmanship, Chung illustrated how tasks involving low-value information can gradually be modified to be more complex. The cognitive demands underlying a task, and "what's going on in someone's head with respect to the task," can be understood by developing online measures, he said. More research is needed on this aspect and others related to distributed learning, but we know that these systems will increase in their development and use in education and training contexts. Already, 89% of 4-year public postsecondary institutions have at least one online course, 48% offer degrees incorporating online credits, and the military is spending billions on online-related training development.

John Bransford of the University of Washington suggested that the concept of "adaptive expertise" (a concept originally suggested by Hatano & Inagaki) provides a useful gold standard for education, and that adopting this standard requires the development of new kinds of assessments. Working with other researchers from around the country—especially Dan Schwartz of Stanford and Nancy Vye of University of Washington—he has developed a tentative analysis of adaptive expertise that

includes at least two orthogonal dimensions: efficiency and innovation. Adaptive experts are high on both dimensions; routine experts are high on efficiency.

Typical assessments are based on efficiency, Bransford said, and they test the direct application of previously acquired schemas and skills in environments that typically involve "sequestered problem solving." "If you are not efficient, you are overwhelmed," he said, "but efficiency is only one dimension. You need to add an innovation dimension, which encompasses invention and restructuring rather than direct application of existing schemas." A key point about innovation is that it often involves a movement away from efficiency--at least temporarily. And it often involves the need to admit that one is wrong and needs to do something differently. Innovation can be facilitated by the right kinds of interactivity among learners and by chances to "bump up against the world" in trying things out.

In order to illustrate why new approaches to assessment are needed to assess adaptive expertise, Bransford drew an example from an experiment with students at all levels—including middle school, college, and principals—in which they studied a report on the decline and renewal of bald eagles. Asked to describe what they could do to protect the species, even the more mature adults could only come up with low-level answers. There was also considerable negative transfer (e.g. the idea of bringing in some eagles just like people bring in wolves to repopulate forests), but when asked what they could do to learn more in order to save the species, their answers were much more sophisticated. They asked sophisticated questions and considered their initial (often erroneous) hypotheses to be tentative. Overall, it became clear that college students and adult, especially, were well prepared to learn. The idea of "preparation for future learning" (PFL) involves a different view of transfer than the idea of the direct application of existing skills and schemas. Bransford and colleagues argue that efficiency based assessments often make people "look dumb". Even more important, they can blind us to the value of educational experiences that look weak on the basis of "sequestered problem solving" efficiency measures but look strong when PFL (preparation for future learning) measures are used (e.g. see Schwartz, Bransford & Sears, in press).

"The more we explored the idea of PFL assessments, the more we saw the need to change typical curriculum structures, including problem based learning," he said. "We moved to environments that encourage students to learn to work smart by inventing tools and social procedures in order to become more efficient." Curricula that

support "working smart" require quasi-repetitive activity cycles, or QRACs, which include several cycles of action, reflection, and revision. This is very different from most approaches to curriculum which involve teach A and then test A, teach B and then test B. Rather than being cyclical, the latter approach is more like a conveyor belt.

Adaptive expertise and ideas of PFL transfer can supply what is missing in current assessments because they layer in what people need to get along in the world, Bransford said, and highlight what isn't being captured in current assessments. He also said that PFL assessments could not be implemented on large scales without technology because "we need to see how people select resources, how they grapple with them and use them to change their initial thinking, and the nature of their social networks—such as whom they contact and how they collaborate."

The major question during the discussion period was how the ideas from the presenters could find room in accountability systems under NCLB. The standards could include them, Shepard said, provided fairness is part of the picture. Students will need to know "what you want me to do."

Assessing Instructional and Assessment Practice

Studies of large-scale educational programs often need accurate descriptions of classroom practices. However, such descriptions are difficult to obtain in an efficient and timely manner. Common methods for measuring instructional practice have their limits, according to Brian Stecher of CRESST/RAND, that can affect their validity. Observations can be complex, time consuming, expensive, and subject to the biases of the observer. Similarly, survey responses can have a self-reporting bias, be distorted by faulty memory, and suffer from the lack of a shared understanding of reform terminology. Maintaining classroom logs has some of the same problems.

To address this problem, Stecher and his colleagues developed a short "vignette-based" measure of mathematics instructional practices that asks teachers to rate the degree to which various teaching practices correspond to what they do in their own classrooms. The Vignette-Based Study of Reform Teaching Practice, part of the Mosaic II Project funded by the National Science Foundation, "attempts to measure intention to engage in reform-oriented teaching," Stecher said. Teachers in the project respond to specific, hypothetical, but familiar situations with alternatives stated in clear behavioral terms. This approach presents realistic situations and choices, using common terminology, which standardizes the collection of teacher responses.

In the study, a panel of math experts used the National Council of Teachers of Mathematics standards and other documents to create vignettes for two common fourth-grade math topics. The options given to teachers represent a range of high- and low-reform actions and are parallel across the two topics. Each option is assigned a reform value, which allows the set of responses to be combined into an overall measure. An example of the response options is below.

After praising both groups for using effective strategies, how likely are you to do each of the following in response to these two explanations?

		(Circle One Response in Each Row)			
		Very unlikely	**Somewhat unlikely**	**Somewhat likely**	**Very likely**
a.	Ask the class if they can think of another way to solve the problem	1	2	3	4
b.	Suggest that the class check the results by using a calculator				
c.	Tell them the first group's method is faster	1	2	3	4
d.	Tell them they are both right and move on to the next problem				
e.	Have a classroom discussion about the differences between the two approaches	1	2	3	4

Analyses with 80 fourth-grade teachers who took the survey showed that teachers had fairly consistent responses across the two vignettes and their responses were moderately correlated with those obtained from classroom observations and more traditional teacher surveys and logs. The results provide guidance to inform the development of measures of instructional practice.

On the other hand, the researchers learned that the structured vignettes were difficult to develop, reading demands were high, and the evidence for validity was mixed. For example, the study found that teachers' responses were stable across parallel math contexts. In addition, the "reform inclination" scale derived from the vignettes correlated with several survey and log measures of reform practice (though not with

observations). On the other hand, the "Euclidean scale" of reform, which was derived in terms of the distance of each teacher from an ideal high-reform teacher, correlated with observational measures of reform-oriented practice, but not with surveys or logs.

"We won't say this strategy has been entirely successful," according to Stecher, but he considers it worthy of more study. The next step, he said, is to study the relationships of vignette-based measures with student outcomes, particularly achievement, which is the overall purpose of the Mosaic Project. Also, the quality of the vignettes needs to be improved through interviews with teachers, the effect of the length and level of detail in the vignettes on teacher responses needs to be evaluated, and the project needs to explore how the vignettes could be used in other contexts. Studies are needed to determine if the vignette approach can be brought to scale.

Another new experiment on documenting teacher practice is to collect artifact packages, or "Scoop Notebooks." As explained by Hilda Borko of CRESST/University of Colorado, the notebooks scoop up "a typical week's worth of instructional materials such as lesson plans, assignments, tests, student work, photographs of the classroom, and teacher reflections on class sessions and student work." The basic question that frames the collection of materials is, "What it is like to learn math in your classrooms?" The research project is based on the premise that an artifact collection, such as the Scoop Notebook, has the potential to overcome limitations of surveys and case studies as methods for measuring instructional practice by representing what teachers actually do in their classrooms—rather than what they report that they do—while requiring fewer resources than case studies.

The researchers used data from 30 middle school teachers in California and Colorado to analyze the reliability and validity of the Scoop Notebook. Data included notebooks completed by the teachers, researcher observations, and audiotapes of lessons (for a subset of 8 teachers). They developed a scoring guide on 11 dimensions of instructional practice, such as cognitive depth, problem solving, and assessment, which raters used to score the notebooks, observations, and transcripts of audio taped lessons.

To assess the reliability of Notebook ratings, the researchers examined agreement among raters along the 11 dimensions, for ratings based only on the Scoop Notebook. They found moderately-high to high levels of agreement among raters on all dimensions.

One set of validity analyses compared ratings based only on the Scoop Notebook with ratings based on the Notebook plus another source of data (observations or

transcripts). The researchers found moderately high levels of agreement for all dimensions, although agreement was lower for some dimensions such as Mathematical Discourse and Assessment. "Some dimensions and teaching practices present greater challenges than others for artifact-based tools such as the Scoop Notebook," Borko said. In addition, teachers vary their activities from one day to the next, and "raters don't always agree on the weight to be given to different activities." Disagreements among raters may be greater when there are inconsistencies in the data due to variations in a teacher's instructional practices.

In another set of validity analyses, the researchers found a substantial difference between ratings of Colorado teachers, who were using more reform strategies, and California teachers, who used more traditional strategies (the former had higher scores). Borko concluded that the artifact collection "is useful for describing classroom practice in broad terms, but it should not be used to make high-stakes decisions about individual teachers." She recommended more research on the differences in scoring among the raters, and the exploration of why some classrooms and some dimensions are more difficult to rate than others.

Noreen Webb of CRESST/UCLA reported on an inquiry to find out why students use such low-level discussions in collaborative groups, even when they have had extensive preparation and practice in collaborative problem solving. For this study, researchers examined teacher modeling of discussions. In four 7th-grade general math classes, students worked in heterogeneous groups for one semester. All teachers and groups were audiotaped for five class periods, and the students were given pre- and post-tests.

The students and teachers received ongoing training for group work. Students who understood the problem, for example, were told to refer to a "helper" classroom chart with such guidance as: "notice when other students need help." "Tell other students to ask you if they need help." "Be a good listener." "Give explanations instead of the answer." "Check for understanding." "Praise your teammates." Students who did not understand the problem could refer to a "helper" classroom chart that told them to: "Recognize that you need help." "Choose someone to help you." "Ask clear and precise questions." "Keep asking until you understand."

The researchers found that despite the training on higher-level group work, teachers did not noticeably change their style of instruction. In most cases, the teacher presented the steps in a problem, requested numerical answers only, did not explain

why student answers were correct or incorrect, did not probe student thinking, and focused exclusively on numerical procedures. This was not surprising, Webb said, "because most of the training was for students." While teachers practiced the same training activities, teachers focused on how students should behave during small group work rather than how teachers might use the principles of effective helping in the training activities to inform and change their own classroom instruction.

Despite the wall charts and training to support high-level helping behavior, the students usually followed the teachers' modeling and peer assistance was generally poor, i.e., helpers dictated the calculations. Helpers didn't try to determine a help-seeking student's level of understanding either before or after providing help. Students seeking help did not ask specific questions or reveal their level of understanding, nor did they use the help to test their understanding.

The researchers found few instances of "either teachers or students trying to find out the thinking of students needing help…and it was rare for these students to explain what they were having trouble with, or to use the help to try to solve problems on their own." Basically, Webb concluded, the students "did what the teachers did, not what the teachers said." A next step for the research is to identify teachers with useful helping styles and determine if students adapt that same helping style when assisting other students.

One of the stumbling blocks to studying instructional and assessment practice, according to Dylan Wiliam of the Educational Testing Service (ETS), is that "real lessons do not use high-quality instruction." He reported on research at ETS and formerly at Kings College in England, that defined and searched for formative lessons and assessments that actually shape learning. The role of teachers, he noted, "is not to teach, but to create environments in which students learn. Right now, teachers are working too hard and students, not enough." High quality formative lessons and assessment provide students with strong feedback. In the formative context, assessments should monitor, be diagnostic, and move from "what is wrong to what to do about it." The first priority of formative assessments (which can be external), he emphasized, is to serve learning, not accountability. Wiliam showed the following chart to help describe good formative assessment and the role of teachers and students.

	Where the learner is	Where the learner is going	How to get the learner there
Teacher	Evoking information	Curriculum philosophy	Feedback
Peer	Peer-assessment	Sharing criteria	Peer-tutoring
Learner	Self-assessment	Sharing criteria	Self-directed learning

In the above table, for example, a learner might use self-assessment to establish their current knowledge, and as a result of activities provided by the teacher, be clear about what it is they are trying to achieve. If they then know what to do, they would be able to engage in self-directed learning, although more typically they might get support from peers, or the teacher.

Formative assessment can be thought of as an aspect of the regulation of learning. Key components of this regulation are the creation and the use of "moments of contingency"—those times in a lesson where a teacher can go in different ways, depending on the evidence the teacher has about the extent of the students' learning. There is increasing evidence that the use of formative assessment increases student achievement, even when such achievement is measured via state-mandated standardized tests.

Politics, Policy, and Improving Assessment

Leading off a panel discussion on the politics of improving assessment with a focus on states, Michael Cohen of Achieve, Inc., said that content standards were getting better over time. They are more clear and specific, set by each grade rather than a grade span, and are more rigorous. There also will be about 600 assessments in place by 2005, "but I can't say the quality is getting any better."

According to Cohen, the 15-year effort to improve schools before NCLB was established had broad support and accomplished sustained improvement, "but everywhere you look we were coming up short on the capacity to pull this off." This is especially evident when considering two issues facing the states—high school reform and defining what "proficient" means.

Drawing from the Diploma Project, in which Achieve, Inc., is a partner, Cohen shared data about the large number of high school graduates who are unprepared for college (53% take remedial courses in college, and most students who take remedial courses fail to earn a degree). The Diploma Project, which also includes Education Trust and the Fordham Foundation, is working with five states to create end-of-high-school benchmarks of the knowledge and skills graduates will need to be successful both at college and in the workplace.

Analyzing high school exit exams, the Diploma Project also is developing definitions of "proficient" that are anchored to college and work-ready benchmarks. The Project's analysis, he said, found that more than half of the algebra covered on state exams actually covered the content of pre-algebra. Moreover, Cohen said, "states must establish robust and coherent assessments systems that are aligned with standards and that provide useful information to postsecondary institutions. " His advice to states: don't lower standards, don't delay high stakes, improve the tests over time, and build more comprehensive assessment systems.

From the viewpoint of students, the educational reform movement had not gained much traction, especially since 1988, according to Kati Haycock, executive director of the Education Trust. There had been much progress in improving the achievement among poor and minority students, but in the 1990s the gaps began to widen again. Then, many states failed to follow through on the 1994 reauthorization of the Elementary Secondary Education Act. They did not develop the assessments called for in the law, limited-English-proficient students were not included in the assessments, they used weak definitions of Adequate Yearly progress, and disaggregated data were unavailable in most states. The gaps kept growing, Haycock said, laying the premise for the passage of NCLB. While poor and minority students "were effectively hidden from public view, they were not hidden from Congress."

Even before the 2000 elections, the Democratic leadership was determined to act on the failures, especially targeting state accountability systems. The election brought in a president "with a Texas-style accountability" approach, who joined with key Democratic leaders and pressed Republicans to come along. The bottom line, she said, "is that very different politics produced NCLB."

It is easy to focus on the problems created by the law, Haycock noted. However, schools are a lot more focused on the "hidden kids" than ever before, the law "is creating a treasure trove of data validating how much schools matter," and the NCLB

experience is laying the foundation for next-generation approaches such as value-added measures.

Haycock also said that more states and districts are moving toward benchmark assessments to inform instruction, value-added analyses to measure growth, and other new ways of using assessments for improvement.

Focusing on recent developments in California's Academic Performance Index (API), Edward Haertel of CRESST/Stanford University discussed the changes in how it is calculated and the lessons learned. The API is a school-level summary of student test performance mandated by the state in 1999 and the pre-NCLB foundation for the state's public school accountability system. It used to be so simple, he said, with the Index based only on SAT-9 scores in core subjects, with weights given to the subject areas. Over time the measure changed to accommodate changes in the tests (easing out SAT-9 in favor of the California Standards Test); the heterogeneity of the state's schools (such as small schools and alternative schools); the evolving rules for special situations such as testing accommodations, mobility, and partial records; changes because of NCLB (massive because of changes in what sub groups are included, how participation rate is calculated, what tests are given at each grade, data release timelines and others); and refinements in data acquisition and quality control.

The state respects the maxim, "If you want to measure change, don't change the measure." To measure year-to-year growth while at the same time accommodating change, two APIs are calculated each year, one matching the previous year's formula and the other matching the following year's formula.

"We had a system that was evolving rationally," Haertel said, "but NCLB has made that all but impossible." The first lesson learned from the California experience is that long-term stability is probably unattainable, he said. Also, the complexity increases over time; technical and policy considerations often point to divergent courses of action; and fairness is a matter of degree because a uniform system must be imposed on a heterogeneous set of schools, and choices must be made that work to the advantage of some schools and not others. With hard work, however, Haertel said, a system can be created that is trusted and respected, despite its imperfections. That depends on an enormous behind-the-scenes effort to maintain a reliable and valid accountability system.

He advised policymakers to: create and use a stable technical advisory committee; plan ahead, but stay flexible; resist pressures to use tests for new and

different purposes; and remember that the school-level index is only one part of an accountability system and only one part of an education indicator system.

Geno Flores, deputy superintendent of the Assessment and Accountability Branch in the California Department of Education, provided some of the behind-the-scenes picture of California's assessment system. Moving from the dedicated heart of a "Rocky" to the highly specialized and destructive "Rambo," the California accountability system has struggled in the legislature. In the last state legislative session, Flores said, "legislators made testing out to be an evil," and a change in the party running the finance office, combined with legislature cuts, made his assessment office lose 12% of its staff and absorb cuts that eliminated item development. Still, the API system in the state has changed the way people talk about schools. "Now, they are not asking about school-to-school comparisons but about API and growth rates, disaggregated data, and school improvement."

Even though fewer schools made AYP improvement targets in 2004 than in 2002, more students are taking the tests and more are moving out of the lowest performing group, Flores said. Item types have been changed to be more instructionally sensitive. Charged by the media with dumbing down the test with the changing of the items, Flores said that it was important for tests to be able to detect improvements in instruction and asked "How can you be tested on what you haven't been taught?"

NCLB changes the accountability workbook constantly, and new amendments to the California accountability plan will mean more work for his department and schools. Current challenges include recent assessment changes, graduation rates for alternative schools, data availability on first-year limited-English proficient students, and decisions about sub-group sizes.

The discussion of panel presentations focused on the future. Asked if it will be possible to track individual children under a value-added system, Haertel predicted there would be a student identifier system in a few years despite large technical problems. Asked about changes in NCLB, Haycock predicted that it is unlikely for changes to be made before reauthorization in 2007, and even though there are real issues with special education and AYP, any modifications would be a year away. She welcomed collaboration of groups, such as assessment experts, on thinking through the problems.

Asked if the only goal of high school is to prepare for college, Cohen replied that the goal is to minimally prepare students to have a choice, "and to do that, we need to

give them college-going skills." Perhaps there is a problem in requiring all to prepare for college, he said, but so far "no state has designed a system to do that, so we are a long ways" from having the problem.

Next Steps for Accountability

The concluding panel of the 2004 CRESST conference looked ahead to the next steps in accountability with some fresh ideas and some cautions.

Zooming away from so much focus on NCLB, Daniel Koretz of Harvard University and CRESST called for much greater attention to how tests should be designed for use in accountability systems than we have seen in the past. Even though NCLB will eventually change or even be replaced, he said, the issues of test-based accountability will remain. Today's emphasis has been building for 30 years, beginning with state minimal competency testing. But it developed without an empirical basis and without research to say that one system of test-based accountability works better than another under certain conditions.

One unresolved problem is that of establishing sufficient breadth and avoiding excessive reliance on recurrent content or style of presentation. "We know how to establish a sufficient sample for low-stakes uses," he said, "but the issue is limiting predictable reoccurring presentations when teachers have a strong incentive to focus on them." Teachers and coaching businesses like Princeton Review recognize the recurring presentations "and teach teachers how to take shortcuts in response to them," he said. States have to reuse items to link tests from one year to another, but there is not enough research on how to do this well.

Some states are conducting pre-tests to ascertain the quality of tests, Koretz said, but these too are complicated by the current high stakes. "Often students are getting so fed up with tests that they are turned off, and when faced with harder open-response items of field tests, they don't try to answer them."

Another question Koretz wants answered is: How do we know when improvement is real? "We can't take gains at face value," he said, "and we ought to be doing more work analyzing which gains are meaningful." We also need more study about how to respond effectively to variations in performance. "We need to put pressure on low performance, but be realistic," he said, pointing out that variability in performance occurs in high-performing countries, and that race and ethnicity contribute very little to the total variability of performance. If the 2014 NCLB requirements were

met using NAEP proficiency standards, "mildly retarded students in the United States would outscore one-third of the students in Japan." Under the current accountability systems in the United States, students heading for MIT and mentally retarded students are held to the same standards, "which doesn't make sense."

A former special education teacher, Koretz said that the psychological effects of current testing have not been adequately addressed. Excessively high expectations can demoralize students and teachers as well as lead them to look for shortcuts. Research has not addressed where to set standards to prevent these responses. Nor has there been sufficient research on how to create incentives for teachers, such as what kind of sanctions yield the desired results.

The research to produce better test designs needs to be systematic and well funded, Koretz said. "We have embarked on a really powerful social experiment," he said, "that has a profound influence on children. The people who want these programs should be ponying up the money to allow us to find out what the programs are doing and how to make them better."

One area of testing policy that has been almost totally overlooked is the influence of and the information that can be derived from college-level placement tests. The issue is part of the larger failure of the K-12 and post-secondary systems to communicate with each other, according to Michael Kirst of Stanford University. Drawing from the BRIDGE project's study of K-12 and higher education articulation, he focused on the "dark continent of placement testing" as it affects students who do not go to selective colleges. While it may not always be wise policy, more than 88% of high school students want a college degree, and the college-going rate has increased 20% in the last 20 years. Graduation rates, however, are up only 3-4%. Students are enrolling at broad-access institutions, especially community colleges, but "they get a wake-up call when they take placement tests," Kirst said. Sixty percent wind up in remedial courses, and in community colleges, 60% drop out.

Part of the problem is that the K-12 and higher education systems establish standards in their own orbits, and in only a handful of states is there any conversation between the two levels. As a result, students face multiple and confusing tests. In high school, for example, language arts tests focus on literature, while college tests focus on rhetoric. State writing tests are more rigorous than college, while each level emphasizes different aspects of math. One problem is that many seniors do not take any math during their senior year, thus go 18 months without math instruction, and then take a

placement test. Moreover, Kirst said, many placement tests are locally developed and have not been validated.

Kirst recommended much more work on aligning a K-16 system and research on the validity and reliability of college placement tests. He also suggested that high school tests be used for college admissions, which would give them greater status among students. Furthermore, the 12th grade assessment under NAEP could be used as an evaluation of college readiness, and there needs to be a K-16 data and accountability system. Such policies would be difficult to enact, Kirst said, primarily because the higher education community is not interested in evaluation and relies on its "churning" students in and out without getting any reward for keeping students until graduation. "But this is a new century," Kirst said, "and gives us new hope and lots of work."

Jamal Abedi of CRESST/UCLA focused on results from 12 years of CRESST research on sub-group test performances, especially English-language learners. When one of these groups consistently fails to make AYP, he pointed out, an entire school would be considered eligible for reconstruction within four years of the benchmark year.

One of the issues research has delineated is the classification of students within a school. Students have different levels of English language proficiency, he noted. Another issue is the quality of measurements, especially because the language of academic tests is normed for native English speakers, and their validity for limited English proficient (LEP) students is questionable. These tests should not be used for high-stakes decisions about LEP students, he said

Abedi made several recommendations, including: identify new and confounding variables that might negatively affect measurement outcomes and control purposes for LEP students; specify the content being measured in high stakes tests and separate that from construct relevance; and identify accommodations for LEP students that actually help with their language needs.

Abedi noted that schools with high numbers of LEP students have low baseline scores and cannot be expected to improve at the same rate as other schools. They need a higher caliber of teaching, quality professional development, and after-school tutoring, but he was not sure if NCLB provides the right level of support.

A major issue in accountability is that when LEP students reach a certain level of proficiency, they exit from the programs, so that LEP classes always remain unstable

and low performing. Although current waivers from the U.S. Department of Education address this problem, he said the policy does not go far enough and that test scores of exited LEP students should be counted in the LEP programs as long as the students remain in the school. On the other hand, schools should not be allowed to ignore the LEP population when it represents only a small proportion of a school's enrollment.

Looming in the future is the science testing requirement under NCLB, and as a final topic for the conference, Richard Shavelson of CRESST/Stanford University turned to activity in this area. Because assessment people are exhausted and overwhelmed, this upcoming test is getting short shrift, he said, and people are saying: "Just give me something off the shelf."

Even though science achievement "has been flying under the radar," some experts and SEAL (Stanford Education Assessment Lab) researchers have been working on a project to develop a Comprehensive Assessment of Science Achievement, one that tests what is important and one that is coherent with links to formative (classroom) testing for improving teaching and student learning, Shavelson said. They are looking not only at both declarative knowledge (knowing facts and concepts in the domain) and procedural knowledge (knowing how to use routine procedures and aspects of problem solving), but also at schematic knowledge (knowing why and conceptual models of how the natural world works); strategic knowledge (transfer, or knowing when, where and how knowledge applies); epistemic knowledge (knowing how we know--how science knowledge is built and justified); communication skills (communicating ideas clearly and concisely in the genre of science); and motivation (commitment to doing something on the basis of knowledge).

Unfortunately, Shavelson predicted, "most of this work will be ignored." The fallback may be on multiple-choice tests "that are unlikely to measure the range of outcomes envisioned for science achievement." However, the future would look better if some of the innovative assessment were incorporated in the science achievement assessment, relying, in part, on computer and web-based assessments to be developed more fully, he said.

The CRESST conference small group sessions featured ongoing work on the assessment and accountability issues raised in plenary sessions, but also focused on groundbreaking work to embed quality assessments in classroom instruction. Towards this end, CRESST research teams have been developing a number of tools for teachers and schools.

One is the CRESST performance assessment design model that focuses on the cognitive demands of assessments and is intended to improve the quality of assessments used for multiple purposes. CRESST research has moved from small-scale research settings to large-scale trials in Hawaii and the Chicago Public Schools, and most recently was used in the Los Angeles Unified School District (LAUSD) as part of their standards-based accountability system. Writing was a particular LAUSD focus with the goal of creating assessments that could help teachers identify the areas of greatest student need and, at the same time, provide information to the district on the quality of student writing.

In the spring of 2002, the CRESST/LAUSD writing assessments were administered to about 300,000 students in Grades 2-9, and then scored by about 29,000 LAUSD teachers. The process showed that (a) research-based models can lead to high-quality, "learning-like" tasks; and (b) performance assessment scoring can be successfully conducted on a very large scale. The CRESST assessments won broad teacher support, which continued well after the CRESST effort finished, according to Joan Evans, director of standards-based education for LAUSD. More importantly, writing performance increased substantially. In 1998, 22 percent of the LAUSD students were scoring at proficient or above levels in writing. By 2003, after using the CRESST performance assessments for several years, 49% of the same cohort of students scored at proficient or above levels.

CRESST researchers also found the following:

- How often students read and wrote about literature in class was an important predictor of student achievement.
- The performance tests accurately predicted how students would perform on high school exams.
- Although there was a high number of English-language learners among the students tested, only 11 percent of the variation in scores could be explained by student background information.
- The effects of opportunity to learn (OTL) varied significantly across classrooms.

"The substantial differences that we observed within the same classrooms in terms of the performance of students and the levels of OTL they report being exposed to," concluded Felipe Martinez of CRESST/UCLA, "suggests that reform efforts need to

pay as much attention (or more) to factors at the classroom level as to those at the school level." The great challenge for researchers in this area, added Martinez, is the development of rich and reliable, and at the same time scalable and economic, indicators of OTL to be used in large scale studies and assessment systems.

In another part of the study, CRESST researchers looked at the validity and reliability of the essay tests. Ninth-grade student essays were studied to determine how well student performance could be generalized from one writing task to other similar ones. Students performed somewhat differently on different tasks, leading the researchers to conclude that 3-5 essays are needed to make reliable judgments about students' writing abilities.

The CRESST performance assessment model also had the effect of improving the consistency of teachers' evaluation of student writing across the district. The research team trained a cadre of scoring trainers, said David Niemi of CRESST/UCLA, using rubrics based on trial models. The trained teachers, in turn, trained teachers in their schools to score student papers in the spring, and a sample of papers from all schools was collected and rescored in the summer. Feedback from this rescoring, followed by district efforts to insure that local training occurred as intended, helped to improve the consistency of teacher scoring from one year to the next.

By using the CRESST model, teachers "became true believers," Evans said. Today, they are adapting what they learned from CRESST to analyze any assessment, including those from test publishers, "wanting to know where are the rubrics, anchor papers, training papers, and trainers," she added. In schools where Open Court has been adopted, for example, teachers use the CRESST assessments each quarter in order to maintain the "rigor" that those assessments give to them.

CRESST technology tools also are helpful because they contain scoring examples and samples of student work, explained Evans. The interactive nature of the website developed during the project allows teachers to score student work and get immediate feedback on their score with appropriate commentary indicating why the work merits the score that it receives.

Niemi's overall conclusion: "It is possible for a district to develop, implement, and score assessments that have an impact on learning, and this model can work in a district the size of Los Angeles."

Comprehending Reading Comprehension

Drawing on earlier work funded by the National Science Foundation that focused on early reading intervention, Alison Bailey and Margaret Heritage of CRESST/UCLA described a similar model of assessment to address difficulties in reading comprehension. In the earlier research, a literacy development checklist (LDC) had been used to help teachers "see struggling readers through a research-based lens, "explained Bailey during a small group session. Bailey and Heritage expanded the checklist to focus on the black box of teachers' knowledge of reading comprehension and formative assessments. The checklist covers different aspects of reading comprehension, including concepts and specific skills. Other checklist components include word recognition, fluency, text presentation and structure, vocabulary, syntax, discourse in oral language, and appropriate use of language. In addition, the checklist helps teachers to collect information useful to unlocking reading comprehension problems.

Using the checklist, teachers can better evaluate a student's background knowledge, understand how students integrate their knowledge with what they are reading, and identify the strategies students use to comprehend what they are reading, explained Heritage and Bailey. Additional checklist data can help teachers evaluate a student's opportunity to learn outside of the classroom including their access to a local library and literacy practices in the home.

Tools to Improve Technology Use and Public Understanding

CRESST researchers provided annual updates on the Center's long-term involvement in developing technology tools for assessment. CRESST has been active in using technology, according to Bill Bewley of CRESST/UCLA because "technology allows us to evaluate student learning in new ways and with greater efficiency." When used successfully, he added, technology can lead to improved data analyses.

CRESST has developed major technology tools for a variety of purposes including assessment, data collection, analysis and reporting, and automated scoring. Bewley described four types of CRESST tools including: eye tracking to measure focus of attention; knowledge maps to assess content understanding and task analysis; click-stream analysis to collect data on users' information selection in problem solving; and a courseware rating tool using research-based assessment and instructional strategy

guidelines. The latter evolved from guidelines in a binder accompanying a comprehensive software evaluation tool.

Part of a test bed of assessment tools, the research was supported by the Office of Naval Research, the U.S. Department of Education, and the Interagency Education Research Initiative. With the technology, researchers have been able to measure learning rates of novice and experienced marksmen for the Navy and found high levels of agreement between automated and expert scoring. The CRESST research team looked at more complex knowledge mapping structures in order to locate clusters of knowledge and relationships, and identify conceptual weaknesses. These learning processes, Bewley said, "require assessments on the fly, where technology has great efficiency."

Terry Vendlinski of CRESST/UCLA described two projects where researchers studied student problem solving strategies and the relationship of these strategies to student cognition and learning. The research team used Artificial Neural Networks and Lag Sequential Analysis to isolate problem solving strategies from click stream data generated by students solving computer tasks. In the first project, researchers studied students taking a high school chemistry course, while in the second, they studied Navy officers taking a risk management course. The researchers found that students, both high school students and Navy officers, used one of three approaches. The first approach was to try and solve the problem with minimal information. High school students for example, would often use a single chemical test to identify an unknown compound, almost always producing an incorrect answer. In the second approach, students conducted virtually every possible test on the unknown compound, again producing an incorrect answer. Vendlinski believes that the students using the second approach either got lost or didn't understand what the tests told them. The third group, who usually correctly identified the compound, used just enough tests to solve the problem, suggesting that these students had strong prior knowledge and could apply it effectively to the problem.

"We learned that students who used limited *or* copious amounts of information kept using these same strategies even when they consistently produced incorrect answers." The challenge therefore, added Vendlinkski is to get students into the more effective problem solving group, most likely by increasing their prior knowledge together with their ability to apply that knowledge to a problem.

In his conference presentation, John Lee described research showing that at least half of all administrators and teachers do not do any data analysis, and even when they do, most analyses are limited to rather routine curriculum decisions and some longitudinal data collection. To encourage robust use of data at the school level, explained Lee, CRESST developed the Quality School Portfolio (QSP), a free data management and analysis software program in use by 120 school districts across the country.

Using QSP, schools can generate 23 different reports based on a school's goals, demographics, achievement, and other factors. CRESST's web-based QSP training is done school wide, added Lee, helping educators to design effective questions, then collect and analyze data. A corollary tool is the new Assessment Design and Delivery System (ADDS) that helps teachers look at the cognitive demands of a particular task, then use quality assessments to evaluate higher order thinking skills. The data collected by ADDS can be exported into QSP, Lee said, then effectively analyzed to help refine instructional strategies.

A presentation by Jacquey Barber focused on the role and products of the Center on Assessment and Evaluation of Student Learning (CAESL). CAESL is one of a dozen national education centers funded by the National Science Foundation, with partners from the Lawrence Hall of Science at the University of California, Berkeley, WestED, and CRESST. According to Barber, associate director of the Lawrence Hall of Science, CAESL is the only NSF center that focuses on public understanding of assessment.

"We found very little information for parents that was interactive and that didn't give a particular point of view on assessment," Barber said at the concurrent session with Ron Dietel on tools for public understanding of assessment. The tool kit contains a variety of resources to adapt for the situation where they will be used, from a 10-minute presentation at a PTA meeting to an all-day workshop. It includes a framework and materials for three two-hour sessions, seven short PowerPoints, assessment briefs developed with CRESST, and 22 stand-alone handouts. The toolkit materials were pilot tested with different kinds of parents and reviewed by teachers as well. "The toolkit provides objective information about assessment, raises important accountability issues, and promotes discussion," said Barber, "yet allows parents and the public to make up their own minds on key accountability issues."

CRESST has developed a different sort of tool to create better public understanding of assessment, directed at audiences from parents to policymakers.

Dietel, CRESST assistant director for research use and communications, listed a number of barriers to public understanding of assessment and accountability. Among these are the rapid numbers of changes in state tests, state accountability ranking systems, school report cards, student report cards, and the sheer volume of data that is now publicly available. "Many parents and the public rely on the media to explain this new assessment environment," said Dietel, "but, there are a fair number of misunderstandings on the part of reporters, as well as frequent turnover among reporters covering education."

CRESST's media support includes personal contact with as many as 200 reporters each year who need knowledgeable research experts to comment on specific accountability programs or help explain test scores. CRESST created a web page just for parents with easy-to-understand articles on assessment. Their policy briefs and newsletters can also be helpful. Further, CRESST uses other organizations and publications to inform more people about assessment issues and research-based best practices. CRESST partners are a key connection to audiences ranging from other researchers to policymakers, practitioners, and parents.

Assessment Portfolios – for Teachers

At a concurrent session on indicators of teacher assessment practices, researchers from CRESST and CAESL shared their work on assessment portfolios that science teachers use to embed assessment in their instruction. "Existing assessments embedded in curriculum units are often of marginal quality," explained Maryl Gearhart of UC/Berkeley. "Teachers either need to refine them or find others."

CAESL, the organization supporting this project, represents new forms of collaboration among curriculum developers, teacher developers, teachers, and researchers. One premise in CAESL's mission is that classroom assessments are the foundation for a quality multi-level assessment system that incorporates high-quality goals for student learning, quality assessments, and quality use of assessments.

The assessment portfolio is a professional development tool that serves several functions: individual teacher reflection, team collaboration, coaching, and leadership. The first section of a portfolio contains a plan for a curriculum unit which as Gearhart said, "has been deconstructed by teachers, then reconstructed so that the learning goals are conceptually organized and guide teachers in the design of an assessment plan to track student progress over time. She also noted that: "The process is very messy with a

lot of things on stickies, but guiding principles evolve as teachers create a curriculum flow." Other parts of the portfolio include the assessment plan, the assessments, interpretations of student work, and assessment revisions.

The study of this strategy included 30 teachers and administrators over 18 months, during which teachers constructed a series of three portfolios. Teachers and administrators involved in the study are from districts committed to inquiry-based learning, and almost all are veteran teachers. Data include surveys and portfolios for all participants, and observations and interviews with selected cases.

One of the guiding questions for the study is the relevance and utility of psychometric constructs for classroom assessment. The components of a quality assessment system are familiar (e.g., valid for the purpose used, developmentally sound content, reliable and accurate), "but the challenge is to apply them to classroom assessments," said Cheryl Schwab of UC/Berkeley.

The portfolios capture the quality of the assessments teachers are using in their classrooms, such as use for instructional improvement, student involvement with assessment, and timeliness and usefulness for all students. Teachers tended at first to depend on scripted assessments, then gradually learned to refine assessments to improve their quality. The assessment plans often contained too many assessments to implement, but teachers learned to select fewer and to integrate them more strategically with the curriculum. "Teachers continued to need support for the work," Schwab said, "but if you provide too much support, it begins to look like a prescription."

According to Sam Nagashima of CRESST/UCLA, survey findings provide evidence of growth in teacher skills with assessment and improved alignment of assessments with instructional goals. In a recent survey, added Nagashima, teachers expressed less satisfaction with their assessment findings "suggesting that they were becoming more knowledgeable about assessment quality and thus more critical in their ratings."

A goal of the project is to involve the teachers in leadership activities that will create a coherent science assessment system in their districts from Grades K-12. "But science is not a priority in their districts," according to Craig Strang of UC Berkeley/Lawrence Hall of Science, "and they are trying to achieve some changes where there is not a lot of attention right now."

The participants in the project have developed a matrix of what needs to be addressed in their districts to bring about systemic change in science assessments in the next five years. The planning matrix includes strategies for professional development, parent and community involvement, development of a quality assessment system, and cultivating their own replacements in leadership roles. In the 2004-05 school year the project teachers were selecting a completed unit to teach others in their school or department, serving on committees, and giving presentations.

"They probably all will be retired by the time they have a thorough level of expertise," Strang said, "so you cannot focus on making sure they are experts before they can be leaders. Help them become critical consumers and know where to find help, but don't wait until they know everything."

Help Is Here: A Template for School Report Cards

Reporting mandates have made school information transparent to parents and the public, while at the same time flooding the same audiences with reams of data. In a special presentation of CRESST's work on school report cards, Kyung-Sung Kim outlined what now must be reported publicly under many state laws or under the NCLB act. Data includes student demographics, student outcomes (attendance, graduation, dropouts); AYP data (student performance and participation rates); school/classroom safety, facilities and other characteristics; teacher characteristics; and any special indicators such as survey results.

Trying to format all of this information has evolved in the past six years at CRESST, beginning with simple data collection and basic software use, then continuing to summary information with engaging graphics. Responding to schools' diverse format needs that frequently changed each year, CRESST developed a school report card template that incorporates data from many sources, is simple to use, and provides a selection of graphics and other format options.

According to Kim, CRESST is now able to help schools and districts with the entire process of reporting school progress to the public. This includes data cleaning from many sources; report card design; summarizing aggregated and disaggregated data; and posting the report cards, either web-based or printed.

Measurement and Evaluation Issues

Researchers from a session on measurement and evaluation issues, examined the validity of inferences and consequences of adequate yearly progress accountability rules. This is perhaps one of the most important issues in school accountability and evaluation today. To what extent are test scores a product of good schools or a product of good students who happen to be enrolled in those schools? CRESST researcher Pete Goldschmidt delved into this issue by conducting a longitudinal analysis of elementary schools located in a large urban school district in California. He used a 10-year panel data set of state reading and mathematic assessments examining within-school changes and their effect on school performance. Through a two-stage hierarchical growth model within the general value-added framework, Goldschmidt measured the effect of changes in school demographic composition over time.

"Simply looking at school achievement over 10 years doesn't tell us much about the school," Goldschmidt said. "Correlations over the years will be high," he added, "but we still don't know if we are looking at the quality of the school or the characteristics of students enrolled at the school."

One answer is to move to more complex models, according to Goldschmidt. In terms of school context, the focus of growth and value added models is generally the effect of context on between-school differences in achievement growth, but not on how changes in context within a school might affect differences in achievement growth within schools.

A preliminary analysis of the data in the study suggests that changing demographics do not necessarily preclude schools from meeting adequate yearly progress. What is not measured in this model, said Goldschmidt, are changes in teachers or in school policies and practices, which need to be considered as well. The research team is further investigating school characteristics that account for a pattern of differential year-to-year performances.

Other papers at the session compared models using student characteristics compared to models using policy and practice effects. KC Choi and colleagues conducted a study in a diverse, large urban district, using two time points on an ITBS reading assessment. The researchers learned that school effects were totally different, depending on which model was used (adjusting or not adjusting for socio-economic factors). In terms of adjustment, the researchers found that once the initial status of students is included, adding new socioeconomic status (SES) factors has very little

impact. They also found that adding in school SES factors may over-adjust the school effect. They concluded that SES needs to be considered when measuring adequate yearly progress, but care should be taken to avoid over adjusting; and policy and practice effects need to be part of the model.

CRESST partner Yeow Meng Thum in his presentation said that he examined the valid use of performance levels to rank and sanction schools. This is a critical issue because cut scores, whereby scale scores have been transformed into performance levels, change as a function of the procedures and protocols employed to produce them. The problem is exacerbated when we attempt to monitor progress over time in terms of performance levels. Using data simulated for a design that varied several key components of standard setting procedures, Thum and colleagues analyzed the potential impact of shifts in cut scores on actual SAT-9 performance data from four urban schools. Their findings concluded Thum suggest that large-scale test publishers need to study and continuously monitor the cut-scores for drifts or shifts in order to maintain confidence in performance levels used for high-stakes decisions.

Derek Briggs of CRESST/University of Colorado at Boulder reviewed two popular methods for measuring adequate yearly progress: first, the cross-sectional model used under NCLB and second, the value-added growth model being urged as an NCLB alternative. His detailed analysis, as described in his presentation, indicated that it is risky to draw valid inferences about a classroom teachers' effectiveness (cause) on student achievement (effect) from either model. Each may be causally interpreted, concluded Briggs, but there are a number of problems: interpretation of the data under a value-added model can be biased; neither model produces meaningful interpretations of why students score as they do; and the NCLB model does not describe student learning. Consequently, the validity of either model is in question. An approach that avoids the sorts of causal inferences that are so problematic in both models would be to evaluate the effectiveness of the value-added accountability *systems* as the treatments intended to effect student learning.

Future research, Briggs said, should examine (a) the extent to which cross-sectional and longitudinal data provide different pictures about student learning, (b) whether estimates of teacher/school effectiveness are sensitive to innovative professional development programs, (c) whether estimates of teacher/school effectiveness can be externally validated by measures of teacher/school quality that are not a function of student gains on standardized tests, and (d) how the interpretation of

teacher/school effects may change as a function of the frame of reference for the counterfactual comparison. Briggs is currently exploring this latter issue with empirical data.

Several CRESST initiatives currently focus on evaluating specific programs; from the adoption of English language programs to after-school and arts programs. The researchers discussed findings from two recent evaluation projects with very different foci—one focused on the evaluation of a whole-school arts infused instructional model (The Artful Learning Model), and the second on a state-wide professional development initiative (The California Professional Development Institutes). To examine the effect of the Artful Learning Program, funded by the Grammy Foundation, Noelle Griffin of CRESST/UICLA used interviews, pre- and post-implementation surveys of teachers, and analyses of school-level achievement and instructional materials.

The Artful Learning Program's theory of action was that the "joy, discipline, and commitment required by the arts will provide a useful framework for the overall improvement of education and for individual growth." Griffin's study found several barriers to fulfilling this goal. These included turnover of leadership and inadequate school administrative and district support, leading to a loss of program momentum. Griffin also found that elementary schools had more successful implementation than middle schools, the latter having more teachers who questioned the usefulness of the initiative.

The research team found that creativity in the program was higher when schools relied on outside arts specialists. "When the implementation was self-contained, many schools used a curriculum that didn't apply to the subject," Griffin said. While teachers rated the professional development offered by the program as very high, they struggled with assessment. Eighty percent said the professional development prepared them well for instruction, but only 42 percent said it helped them with assessment. The dilemma, according to Griffin, "is that teachers tended not to use assessments in artful learning, then would ask how do we know if it [the program] is working?" The programs are settling in at the schools, but assessment and parent involvement have remained weaker elements.

As part of an evaluation of a state-wide professional development initiative in reading, Griffin and LeeAnn Trusela of CRESST/UCLA studied teachers' experiences with the adoption of published reading programs in primary grades in California. The researchers studied the professional development offered for the two state-adopted

reading programs in California, Open Court and Houghton Mifflin. Both are 40-hour institutes during initiate program adoption with a 40-hour follow-up option. While the researchers primarily studied the quality of the professional development, they found that teacher's attitudes toward the adoption had a large impact on implementation.

A majority of teachers felt the adoption had been top-down and that they were not given enough time to implement the programs properly. One-third of them were critical of the program's writing components, and many were concerned that it did not allow for varying student proficiency levels. There was considerable overlap between areas teachers deemed lacking in training and areas they had difficulty implementing. Teachers wanted more training and opportunities for practice and preferred trainers who had had experience using the materials in their own classrooms. The researchers concluded that further research on large-scale program adoptions needs to investigate more thoroughly how teacher beliefs and school and district-level decision-making processes affected program implementation.

Denise Huang of CRESST/UCLA and her colleagues conducted an evaluation of the L.A.'s BEST after-school program in the Los Angeles Unified School District, collecting data from more than 100 sites. L.A.'s BEST provides academic support, recreation, community and cultural programs, parent involvement and performing arts experiences for at-risk students.

Data from the studies indicate that students either improved or maintained their SAT-9 scores and fewer L.A.'s Best students were in the below basic and basic categories in both reading and math. The L.A.'s BEST program tended to have the largest effect on English-language learning students (most of the attendees were bilingual), who reported lower levels of concerns about school safety, higher cooperative and study skills, and better conflict resolution skills. These students also had higher confidence about doing well in school. Additionally, L.A.'s BEST parents tended to become more involved in their children's schools once their children reached high-school age.

Exploring the issue of alignment and opportunity to learn, Zenaida Aguirre-Munoz of CRESST/Texas Tech examined the use of academic language from a functional linguistics perspective with English learners. "Academic language is language used in the classroom for the purpose of acquiring knowledge," said Aguirre-Munoz, "not language for informal discourse." She and other researchers developed professional training for teachers on how to dissect linguistic features of academic text

that focused on key elements of written academic language including noun and verb phrases, in addition to text organization at the clausal level.

The purpose of the training was to deepen teachers' knowledge of language structures and provide them with instructional resources in academic language of English language arts. Thus, strategies highlighted writing instruction aimed at developing students' academic English proficiency. For example, moving a phrase in a sentence can provide more clarity and make it easier for English learners to understand the text's message. Another strategy is to expand on a noun phrase with prepositional phrases and embedded clauses.

After training, teachers were able to examine student writing such that they were (a) more informed about students' progress in academic English writing; (b) able to plan subsequent instruction that builds on students' current developmental level; and (c) provide specific feedback to students that went beyond superficial aspects of writing, such as mechanics.

Another opportunity to learn study was conducted by Maria Ruiz Primo and colleagues at Stanford University. They used students' science notebooks to examine the nature of instructional activities in science classrooms, the nature of teachers' feedback, and how these two aspects of teaching related to the students' performance. Each entry of each student's science notebook was analyzed according to the characteristics of the activity, quality of student performance, and teacher feedback.

Results across 10 fifth-grade classrooms indicated that the intellectual demands of the tasks found in the notebooks were, in general, low. Teachers typically asked students to record the results of an experiment or to copy definitions. These types of tasks by themselves are not challenging enough to either engage students in scientific inquiry or help students improve their understanding in science. Results also showed that only 40% of teachers provided feedback to students, even though the notebooks revealed errors or misconceptions on the part of students. Furthermore, when there was feedback, it was usually just a grade, checkmark or a code phrase rather than any useful comment.

Low student performance scores revealed that students' communication skills and understanding were far from the maximum score and did not improve over the course of instruction during the school year. However, inferences about student performance using notebooks were justified. High and positive correlations with scores

obtained from performance assessments indicated that students' notebook scores can be considered an accurate indicator for students' science achievement.

The researchers concluded that notebooks can be used as a source of information to obtain a partial picture about opportunities students have to learn science. Unfortunately, neither students nor teachers were using notebooks as a valuable source of information for teaching or learning.

"Measures of Alignment and Opportunity to Learn"

Another session provided a specific strategy to improve opportunities to learn in the classroom. Called the Instructional Quality Assessment (IQA) Toolkit, the strategy is being developed by the Learning Research and Development Center at the University of Pittsburgh, explained Lindsay Clare Matsumura at the conference.

Intended for use in high-quality professional development programs, the IQA Toolkit is linked to a wide range of research on instructional quality with a conceptual base drawn from the National Research Council's study on *How People Learn.* The Toolkit provides specific information about managing teacher instructional behaviors, explained Clare Matsumura, such as choosing proper texts and developing rigorous activities around them. This means that the model for the assessment tool is:

- Learner centered, showing teachers how to draw out instructional direction from student misconceptions;
- knowledge centered, in that the model goes into depth on core concepts;
- assessment centered, so that teachers can help students develop a clear understanding of what they need to know and do, and develop assessments that allow students to demonstrate high-level thinking; and
- community centered, building classroom organizations that create an academic learning community in the classroom.

Discussing only the observational piece of the toolkit, Clare Matsumura used a classroom video vignette to illustrate how the IQA encourages assessment of "teacher press," or the evidence of how a teacher is involving all students in rigorous thinking. The video, covering reading comprehension, also yielded clues as to the teacher's choice of text, and the rigor of lesson activity and discussion. A model for math instruction, she said, uses different rubrics, such as looking at patterns and applications of new knowledge.

The IQA toolkit, Clare Matsumura said, is a way to look for teacher expectations of high-level work, which receives very little attention in American classrooms, she added.

References

Pellegrino, J.W., Chudowsky, N., and Glaser, R., (Eds.) (2001). Knowing what students know: The science and design of educational assessment. Committee on the Foundations of Assessment. National Research Council.

Schwartz, D. L., Bransford, J. D., Sears, D. L. (in press). Efficiency and innovation in transfer. (2005). In J. Mestre (Ed.), Transfer of learning from a modern multidisciplinary perspective (pp. 1 - 51). CT: Information Age Publishing.