Celebrating 20 Years of Research on Educational Assessment: Proceedings of the 2005 CRESST Conference

CSE Technical Report 698

Anne Lewis

August 2006

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.

The findings and opinions expressed do not reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences or the U.S. Department of Education.

Table of Contents

CELEBRATING 20 YEARS OF RESEARCH ON EDUCATIONAL ASSESSMENT: PROCEEDINGS OF THE 2005 CRESST CONFERENCE

Anne Lewis

Introduction

The 2005 CRESST conference marked the 20th year of work on critically important accountability topics by the UCLA institution, "a tremendous accomplishment for a research center," according to Aimee Dorr, dean of the Graduate School of Education and Information Studies. In her welcoming remarks, Dean Dorr described why CRESST has achieved such longevity. The center is "independent, very lively, grounded in practice, and very forward looking, with many top accountability experts from around the nation," said Dorr, "interested in new technologies and helping to shape the future of education." She also noted that although it was "good fortune" for the center's senior partner to be located at UCLA, "it is a partnership throughout the country, and one that enriches us here as the partners do on the national scene."

The anniversary for CRESST was an opportunity for the conference program to focus on the achievements in the use of assessment to improve student learning. The two-day gathering described many of the lessons learned from a century of testing. The discussions also featured the newest CRESST initiative, known as POWERSOURCE, a $10 million grant from the Institute of Education Sciences at the U.S. Department of Education to develop formative mathematics assessments in the middle grades to improve student performance and learning.

THE HISTORY OF EDUCATIONAL ASSESSMENT

Four long-term CRESST colleagues set the conference agenda with an opening panel discussion on the history of assessment, a presentation that ranged from the centuries-old Chinese civil service exams to today's assessment policies and their future.

Learning from History: Assessment and the Improvement of Learning

Speaking just days after Hurricane Katrina revealed the "incompetence and immorality" of the neglect of the poor in the Gulf Coast region, Edmund Gordon drew parallels with issues in the use of assessment. The crisis in the Gulf Coast, said

the John M. Musser Professor of Psychology Emeritus at Yale University and director of the Institute for Urban and Minority Education at Teachers College, showed that "we have known about these conditions, we've continued to create these conditions, and we've done nothing to mitigate their destructive power." Similarly, "we have known about and neglected the problems of the use of assessment to improve learning for a long, long time."

Gordon traced the efforts of early testing pioneers including Alfred Binet, who encouraged the use of tests to improve learning rather than just document it. The dilemma of assessment policy today is not new, said Gordon. "Because of the profit that can be taken from making assessments," he said, "we have found greater advantage in pursuing, not the formative or the developmental potential of these tests for the improvement of education, but the equally valuable use of these tests to document status and, more recently, to pursue accountability." At a time when the nation is making the biggest change in education policies since the 1960s, "we're embracing and expanding the use of tests to measure and document status to the neglect of the use of such technology to inform the improvement of learning."

Contemporary researchers have provided good ideas to use, Gordon said. These include criterion reference testing strategies to help educators, deconstructing test items to reveal the intent behind the tasks, curriculum-embedded assessments, and portfolios. The reasons such ideas fail to influence assessment policies, according to Gordon, include politics and profit making, but most importantly because they represent contradictory alternatives.

For example, there is a contradiction between the pursuit of validity and reliability. "Both are important," he said, "but as we try to generate reliable data, we have to manipulate situations and phenomena in such ways as to threaten the validity of the thing we're measuring." Another contradiction is between modification and prediction. Testing has created a capacity to make relatively sound predictions, "but increasingly we are called upon to use these data to modify the thing that we are trying to predict. A third contradiction described by Gordon is in the tension between the description of learning processes and the determination of status. In other words, educators must be concerned with "situative" criteria or how a person functions in a particular situation, but also must be equally concerned with normative criteria, or that which can be generalized.

Contradictions need not be constraining, however, Gordon concluded. As others have written, it is important to understand the contradictions and use them to craft appropriate action. I believe that the preferred role of measurement in education is not to document status or even to measure progress, he said, but to inform teaching and learning, and, ultimately, to cultivate intellect.

Learning from Past Mistakes

Lorrie Shepard, CRESST partner at the University of Colorado at Boulder, agreed with Ed Gordon, saying that psychometricians have often believed that assessment should be used to improve student learning. In fact, in the first volume of Educational Measurement on 1951, both Ralph Tyler and Walter Cook discuss how educational measurement should facilitate learning. If this is what they believed, "how did we end up with the current situation?" she asked.

In Shepard's view, there were three critical errors in the thinking of early psychometricians. First, they trusted the objective formats to represent important learning goals and to replace what Edward L. Thorndike had called "the scandalous unreliability of teacher-made essay tests." She acknowledged she was not critiquing multiple-choice items, but was criticizing how they were typically used. The first assessments were mostly historical fact items, Shepard explained, and were believed to be sufficient. The second mistake was that the pioneers of assessment did not foresee the important differences between day-to-day formative assessment in classrooms and annual formative program evaluation. Tyler, who is a hero in Shepard's opinion, especially did not make this distinction, believing that teachers would convene with measurement experts to make a standardized testing system primarily for the district level.

Tyler's rhetoric was about learning and classrooms, she said, "but he created and then launched what became the large-scale assessment apparatus that we have, which is not consistent with his own rhetoric." The final mistake of Tyler and his colleagues was their belief that knowing what students didn't know would be enough to know what to do about it. If you could point out to teachers what items students missed, that would be enough to help them develop interventions.

"These are the big ideas that we have been working against" in trying to reform assessments over the past two decades, Shepard said. Researchers recognized the negative effects of what assessments had become such as test score inflation (students did not really know what the tests said they knew) and curriculum

distortion (the major cause of why students were not really learning). The remedies took very different forms in the United States compared to other countries. The U.S. developed performance assessments, creating test items that looked like the real learnings that were intended. Australia, the United Kingdom, and New Zealand focused more attention to the process, or ways assessment should be used to inform instruction, and on formative assessments.

Shepard gave examples of exemplary performance assessments developed in this country, and gave special attention to the research on formative assessments. She described research on the motivational and cognitive effects of classroom assessment and a paper by Royce Sadler, who developed a model of formative assessment that includes feedback and self-monitoring. The Assessment Reform Group, started in 1999, promoted the idea of "assessment for learning instead of assessment of learning," or a basic distinction between formative uses of assessment and summative uses of assessment, explained Shepard. This group commissioned a study of formative assessments by Paul Black and Dylan William, who found that well-designed formative assessments produce large positive effects, and when they are used, they help low-achievers the most. She attributed this effect of formative assessments to the fact that they help scaffold learning for low achievers so they gain some of the same metacognitive insights typical of proficient learners.

The 2001 report of the National Research Council, Knowing What Students Know, presented cognitive science findings on key aspects of learning processes, and these can be translated into targeted features of formative assessment, Shepard added. She noted that many people at the CRESST conference helped fuse learning theory with measurement theory in the NRC report. The aspects that translate well into formative assessments include accessing prior knowledge, strategic use of feedback, teaching and assessing for transfer of knowledge, and meta-cognitive benefits of self-assessment.

Shepard said she has tried to contribute to the assessment reform conversation by focusing on the socio-cultural aspects of learning theory, not just the cognitive. Contemporary psychologists also have gone back to Lev Vygotsky's concepts of dynamic assessment and instructional scaffolding and others who contend that learners develop an identity of mastery as they participate in a community of practice. "The social practices of your community include how you perform intellectually," she said, "and you can make someone smart or not by whether that's part of the discourse and part of the expectations in that community."

Also, "while psychometricians were asleep," subject matter experts in this country were blending a better representation of content with the processes of learning. Shepard provided examples from early literacy research, including using analyses of student work, and mathematics, where classroom discourse is used to contribute to helping students solve problems.

Gordon's references to the influence of money and politics on assessment policy were underscored by Shepard's description of the choices to be made—either create curriculum-embedded assessments and invest heavily in helping teachers to learn to do them, or purchase external mechanistic assessment systems that document that something is happening for policy leaders "but don't necessarily do all of the rich things that the formative assessment literature has promised." The latter strategy essentially is "hijacking" reforms, she said, and repeating mistakes one and two.

As for the third mistake, thinking that knowing what students don't know is enough, curriculum-embedded assessments can address the problem, provided teachers know how to make qualitative judgments and build on existing knowledge. Assessments can help, Shepard said, if they are built to uncover specific misconceptions. More helpful, however, are learning progressions, or a developmental scaffold that teachers can use to determine where students are and what their next step will most likely be. "Knowing how learning typically unfolds helps teachers know 'what next' and how to 'back up,'" she said. CRESST conference participants from Australia have been working on this reform for two decades, Shepard noted. In this country, the most productive work has been with emergent literacy and early writing. In the Netherlands, "learning-teaching trajectories" are being developed "that weave together learning progressions, curricular understandings of those progressions, and the pedagogy that goes with moving forward."

Shepard concluded by noting that she worries, as does Gordon, that many of these ideas have been around for a long time but have been prevented from being used by political and monetary impediments. "Are we recreating the mistakes?" she asked.

The Many Ways of Using Tests for Educational Improvement

While Shepard focused on the evolving rationalization for formative assessment, Edward Haertel of CRESST and Stanford University looked back at

district and state testing policies to find out "what history reveals about what was sensible at one time that, in retrospect, seems to have been mistaken." He discussed seven distinct models, or ways tests have been used to make education better, that in some instances turned out to be "bad ideas:"

- Group students into fast versus slow tracks, according to their IQ. The pressure to educate large numbers of immigrant students encountered the business efficiency movement and the development of school administration as a profession. Also, prevailing beliefs about the origins of group differences favored this kind of IQ testing. It was viewed as "a powerful new scientific tool." Today, IQ-based tracking is seen as having been based on unexamined assumptions about "nature versus nurture," as not recognizing the influences of language and culture on test performance, and as creating a negative effect on opportunity to learn.

- Divide the curriculum into tiny testable slices; test for mastery before moving on. The first example of this was the Winnetka Plan used in Winnetka, Ill., where the math curriculum was taught using mimeographed worksheets focused on very narrow skills. Students could take self-tests, then a secure test given by the teacher when they felt ready. Other examples of such tests, which are grounded in behaviorist psychology, include programmed instruction, mastery learning and criterion-reference testing. The belief that complex learning could be broken up into small bits, with the bits coming together again for effective practice in some other context, was mistaken, Haertel said. We now know that learning is best when it serves a purpose embraced by the learner. Mastery learning also failed because "it just gets boring."

- Diagnose learner's individual aptitudes; prescribe instruction accordingly. This differs from IQ testing in that the IQ model kept children on different tracks to reach their own individual capacities, while the diagnostic approach said all children would be brought to the same point, but get there by different routes. Today, it is recognized that different instructional approaches may be needed, but that the "local knowledge" of experienced teachers is usually a better guide "than any battery of tests we have been able to devise." Shepard's emphasis on formative assessments, Haertel said, points to resources that could be helpful.

- Use tests to determine the best curriculum or best instructional practices. Examples of this approach were the Head Start mandated evaluations, the Follow Through Planned Variation Study, evaluations of post-Sputnik math and science curricula sponsored by the National Science Foundation, and the current What Works Clearinghouse sponsored by the Institute of Education Sciences. It was thought that scientific methods can reveal

answers about the effectiveness of curricula and instructional methods. Today, Haertel said, "we understand that progress in education research requires a plurality of approaches," including randomized trials, case studies, basic research, and the investigation of new models. "And we know," he said, "that the process of dissemination is complex and that we can't simply look for the package that works the best, plunk it down, and assume it's going to work the same way in all classrooms."

- Require passing a minimum competency test or exit exam for high school graduation. These introduced high stakes for individual learners and grew from the perception that schools were maintaining high standards and that the high school diploma did not stand for much of anything. Minimum Competency Testing rhetoric was about basic skills and current high school exit exam rhetoric is about academic skills for college and the workplace, but it may well be that the actual tests are about the same, Haertel said.

- Use authentic hands-on performance assessments and portfolios to promote higher order thinking and to improve motivation. Examples included the New Standards Project, Vermont's portfolio-based assessments, and the California Learning Assessment System. These approaches considered multiple-choice testing to be pushing instruction toward superficial, isolated bits of knowledge. It was argued that the right kinds of tests, aligned with clear standards, could push instruction in the right direction. They were "ambitious and worthwhile" efforts, Haertel said, but they failed because of their "lower reliability, high cost, and rushed implementation."

- Evaluate schools according to the percent of students scoring at- or above-judgmentally determined cut scores representing "proficient" on tests that, by mandate, are aligned with academic content standards. This strategy was introduced in the 1994 reauthorization of elementary/secondary education and mandated in the No Child Left Behind Act reauthorization of 2001. The latter contains many worthwhile provisions other than testing, Haertel said, but "the balance seems to be a little bit too heavily on the side of testing and not enough on the side of capacity building."

From this historical perspective on the uses of testing to improve student learning, Haertel pulled four lessons. The first is that the educational system must acknowledge individual differences in student aptitude. He acknowledged that confusion about sources of within- versus between-group differences in test scores has made this topic difficult to discuss. "We recognize, as every teacher knows, that children develop at different rates, they have different interests, they have strengths in different areas," he said. "Human beings are complex." All children must be expected to master the basics, but "a sensible, humane educational system has to

provide alternatives to the one-size-fits-all assumption that everyone is going to go to college. There has to be room for a variety of different kinds of outcomes for different learners." To continue to insist on uniformly high standards for everyone, he said, will result in reducing the uniform expectation because, "as a practical matter, we can't get everybody to the level that we would hope the best students are going to attain."

A second lesson is that some crucial learning outcomes cannot be assessed with multiple-choice tests. "This is scarcely debatable," he added, and noted that Shepard had covered this point already. While performance assessment has a place in the classroom, "we do not know yet how to make it work well," he said. In large-scale assessments, performance assessment has worked best in the area of writing.

The third lesson is that tests, rewards, and sanctions alone will not bring about meaningful educational reform. "If educators knew how to improve test performance and eliminate achievement gaps," he said, "they would already have done so. We need to worry more about capacity building and looking hard at where the problems are."

Finally, educational research that will inform best practice requires multiple methods and perspectives. While the "What Works Clearinghouse" was based on the model of clinical trials in medicine, randomized controlled trials are only part of the research process, the end point of a long sequence of steps that leads to improvement in medical science. And, Haertel said, "as with medicine, education requires a plurality of research approaches."

As a bonus, Haertel added one more lesson, referring to the current model of standards-based reform: "This too shall pass."

From 2300 BC to the 21st Century

No one can talk about the history of achievement testing without discussing the ancient Chinese, H.D. Hoover, recently retired from the University of Iowa, asserted in the final presentation of the panel that focused on looking back on assessment. Hoover proceeded to do just that, describing the system of civil exams that began in China in 2357 BC and continued until the 1900s. The Chinese system included measurement of the six arts: music, archery, horsemanship, writing, arithmetic, and arts and ceremonies. Later, five studies were added to the exam

system: civil law, military affairs, agriculture, revenue, and geography of the empire. Every three years, civil servants had to take written exams in these topics.

In 1015 AD, the Chinese recognized the problems of validity and reliability, so they introduced multiple readers of the exams. "These were things that many people now think they have discovered," Hoover said, "then fast forwarded to the introduction of such exams in France in 1791, in Great Britain in 1833, and in the United States in 1883, modeled after the Chinese. Many people are credited with being the "father of educational testing" in the United States, Hoover said, but his choice is Horace Mann, who supported written exams (the university system had used oral exams up until then). New exams were introduced to the Boston schools in 1846 and were considered superior to previous exams because they were "impartial, more thorough, prevented teacher interference, determined how well students were taught, removed the possibility of favoritism, and enabled all to determine the ease or difficulty of the questions," as described by Mann said Hoover.

It was Fisher's Scale Books for handwriting, spelling, math, knowledge of scripture, and other subjects, published in 1864, that began the standardized testing movement, Hoover said, and they were followed quickly by New York Regents' Exams in 1865. The Regents' Exams began as a high school admissions test, evolved into a college admissions test, and became more like regular exams of today in the early 1900s. It was Joseph Rice who wanted to use such exams to evaluate schools and teachers, which Hoover says may have been the beginning of school accountability. Rice had a major impact on E.L. Thorndike, who probably would be picked by many as the father of achievement testing. "It is sort of frightening," Hoover said, "when you are a psychometrician like I am and read what Thorndike wrote in his 1904 book on testing and realize that I don't know any more than he did."

Thorndike not only built tests, he produced students who became a "who's who in educational measurement in the first half of the 20th century," according to Hoover. The other leaders in the field of assessment were Lewis Terman, who developed the Stanford-Binet test in 1916; and the Stanford achievement in 1923, and E.F. Lindquist, who launched the Iowa Testing Programs in 1929.

Hoover diverged from talking about testing pioneers to talking about test publishers, who entered, traded, and consolidated the testing industry throughout the 20th century. Houghton Mifflin started with the Stanford-Binet tests in 1916,

adding the Iowa Tests of Basic Skills in 1940. World Book, which became Harcourt, began its assessment involvement in 1918 with the Otis Intelligence test, added the Stanford Achievement Test in 1923, and others during the 1930s. The California Test Bureau, now owned by McGraw-Hill, began with tests in 1933; McGraw-Hill later acquired Science Research Associates, which had started publishing the Iowa Tests of Educational Development (high school level) in 1943.

"We hear so much about how kids are being over-tested," Hoover said, but "by 1928 between 30 and 40 million copies of tests were being used annually." By 1930, 1,300 tests and scales had become standardized. He ended his presentation with a brief history of the Iowa Testing Programs, started because Lindquist wanted academic tests that would allow students to compete on knowledge of subject areas just as they did in sports. "This was the first every-people test," Hoover said, because schools that participated in the academic meet had to test all students. Lindquist saw the benefits but also quickly saw the downsides, such as teaching to the test. He introduced the ITBS for elementary grades in 1930, then followed with the high school version a decade later. Lindquist's whole focus, Hoover noted, was on using testing to facilitate instruction, a statement that took the panel back to the premises of all the other presenters—that the struggle over assessment policy is to use them to improve learning, not just determine status.

TRENDS AND POSSIBILITIES FOR ASSESSMENTS TO IMPROVE STUDENT LEARNING

For the rest of the two-day CRESST conference, presenters and participants focused on the current progress and future possibilities of assessment systems. The panel discussions and presentations underscored the fact that assessment policies in the United States are still vulnerable to mistakes of the past but also still capable of making important strides in supporting student learning.

AYP for Schools: Consequences of State Accountability Design Decisions

State variations in the number of schools making adequate yearly progress are extremely large and make little sense when compared to state results on the National Assessment of Educational Progress, said CRESST co-director Robert Linn of the University of Colorado/Boulder. He illustrated his comments with charts of state performance from their own assessments and then compared those to NAEP results. Generally, states on their tests had a large percentage of schools meeting

AYP in 2004, said Linn, with 16 states reporting that more than 80% of schools met AYP. By contrast, NAEP scores in 4th grade reading for the same year showed that only seven states had more than 40% of the students at proficient or above levels. These differences threaten the credibility of current systems.

States have selected one of three AYP designs, Linn said, and the number of schools meeting AYP depends largely on which design they are using. One method uses a straight-line trajectory toward the goal of 100% proficiency (Arkansas has different straight lines depending on the subject and grade level). Another design uses a straight-line trajectory with plateaus for certain years in between the climb (Missouri and Colorado are examples). The third design, such as in Texas and Louisiana, backloads progress, with minimal progress at first and a steep climb in later years.

Two other crucial AYP design factors are the minimum number of children required to form a subgroup and the use of confidence intervals. Linn analyzed state plans with regards to these three design factors and found extreme variation between designs. With regard to trajectories: 4 states used a straight line, 19 included plateaus, and 24 backloaded AYP goals (3 were unspecified). With regard to minimum number of students: 3 states set the minimum above 51, 6 set it at 41-50, 14 at 40, 13 at 30-49, and 11 at 5-20. Eleven states are not using confidence levels, 6 left this unspecified, 14 set it at 95%, and 16 set it at 99%.

Linn discussed several reasons for the wide variation in state AYP designs. They include different definitions of the proficiency levels, state demographics that cause different numbers of sub-group students to be used for separate reporting, the trajectory selected, the minimum number of students for reporting subgroup performance, and the use of confidence intervals. Using Kentucky's assessment results as an example, Linn showed how changing one or more of these factors would make very large differences in the number of schools making AYP.

"State designs have a major influence on the likelihood that a school will meet AYP standards," Linn concluded. "With such variations among the states, we are not making good judgments about schools across the country." If the purpose of the AYP system is to improve learning, then the variations, according to Linn, provide a moral lesson "that it is hard to have a foolproof system from afar. Even with the best of intentions, we have wound up not having valid judgments and not accomplishing the goal."

Form Effects on the Estimation of Students' Progress in Oral Reading Fluency Using Curriculum-based Measurement

The Reading First program under the No Child Left Behind act is based on the principles of reading instruction and assessment defined by the National Reading Panel in 2000. It includes instruction and assessment in five domains (phonemic awareness, phonics, fluency, vocabulary and comprehension); three tiers of instruction (primary instruction, prevention, and intervention); and four purposes of assessment (screening, diagnosis, progress monitoring, and outcomes).

David Francis of the University of Houston reported on a study of progress monitoring, especially of the widely used DIBLS (Dynamic Indicators of Basic Literacy Skills). In progress monitoring, he said, there is monitoring of student progress toward year-end goals with regular feedback for teachers. In order to identify students needing modification of their instruction, the progress monitoring needs to be administered on a regular basis, be brief and easy to administer in the classroom, provide scores on a constant metric, be predictive of end-of-year outcomes, and be free of measurement artifacts such as practice and form effects.

The study used curriculum-based measurement, as employed in DIBLS, which Francis described as "ubiquitous in Reading First," but the characteristics are not unique to DIBLS. DIBLS has many strengths with the properties of good curriculum-based measurement, he said, including that it is a quick, one-minute probe of reading given once a week; has a teacher-friendly format with easy-to-follow directions; provides instructionally relevant information; and focuses on within-grade evaluation of student growth. DIBLS also was included in the study because it has a large number of content stories in place.

The study sought to determine the form (story) effects. Francis noted that readability formulas are not perfect, and it is difficult to precisely control text features, which will affect fluency. Also, there has been a lack of attention to scaling, he said. The researchers gave DIBLS oral reading fluency passages to 134 students in two Houston schools with diverse populations. The second graders were randomly assigned to six groups, and if the stories were all equivalent, "we shouldn't have been able to find any group effects." However, according to Francis, group differences in fluency were apparent at the baseline on the Texas Primary Reading Inventory, a common measure administered to all students, thus indicating a problem with the randomization of students to groups. However, after controlling

for the TPRI findings, the six groups did not differ on any particular DIBLS story, indicating that form effects could be tested provided that group differences were controlled by using the TPRI fluency measure as a covariate.

The form effects in assessment must be addressed if teachers are to form valid inferences about student progress and are to target the right students for intervention, Francis said. The problem, however, is not one of reliability in terms of a low correlation between alternative forms but "is one of inconsistency in scaling across forms." These form effects adversely affect validity and reliability of the slope estimates for individual students. The effects are not unique to DIBLS or curriculum-based measurement, he added, "but they have not been talked about in this literature and have been ignored in Reading First for all practical purposes." The literature on curriculum-based measurement implies that fluency inherently provides a constant scale, he said, "but if it is to be constant, forms must be parallel."

What is needed, according to Francis, is the development of a scale score that takes form difficulty into account. One potential solution, he said, is equipercentile equating. In conclusion, assessment for progress monitoring needs to have empirically equated forms with a scale score metric that factors out differences. Because of the large number of forms in use, he proposed a "FedEx model" that equates all forms to a single standard form based on percentiles.

Adapting Measurement to Reflect Accountability

There have been three stages in testing research and development, according to Daniel Koretz of CRESST and Harvard University. Until the 1980s,traditional psychometrics assumed, even required, a low-stakes environment, even though testing was in some instances high-stakes. Consequences were not a major focus of the research, and behaviors, especially those of teachers," mattered only on the margins," he said. The methods for research and development focused on initial representativeness of the test content (ignoring changes in representativeness arising from test use), the consequences of the testing were largely ignored in the technical analysis, and the level of scores was considered either arbitrary or normative (the focus was on the relative level, not the absolute level of scores).

The second stage of testing research and development was what Koretz calls "traditional psychometrics plus," covering from 1980 to the present. In this stage, theory called for attention to consequences, although research on consequences remained limited. This stage also included only very limited research on whether

gains in scores were meaningful (score inflation), in part because states and districts were not interested in it. Most psychometricians and the policy community "ignored all of this and continued as before," he said. An Internet search of test studies would confirm this response, he added.

The challenge now is to move to the third stage of research and development—accountability-oriented psychometrics. "The investigation of consequences must become routine," Koretz said. "Every study of large-scale assessments has found problems. Studies should be routinely looking at what happens in schools as a result of testing." Also, validation must include evaluation of inferences about gains, which would require the improvement of audit mechanisms and attention to developing representativeness of content. And behavioral responses to testing, especially of teachers, must feed back into the technical evaluation.

The first "next step" in this third stage of testing research and development would be to survey the whole field of technical analysis to map the influence of high stakes tests (accountability), according to Koretz. Other next steps include developing better ways of evaluating gains, developing ways of linking scores that are better suited to high-stakes contexts, developing and evaluating methods for pinpointing corruption at the item level, and examining tests to identify opportunities for inappropriate test preparation. On this last point, he sees no reason "why developers can't learn this Ñ test-prep firms do Ñ but it will take money and time." Finally, test development needs to be refined to lessen score inflation, a result of teaching to a finite number of test items. Koretz's ending comment, however, emphasized the first step: "We have a desperate need to study the effects of high stakes testing."

Struggling for Meaning in Standards-Based Assessment

People may not be getting what they think they are getting from standards-based assessments, in the opinion of Mark Wilson of the University of California, Berkeley. They often believe that state tests are providing a useful result for each standard, or ideal, he said, "but this is an illusion. Each standard cannot be assessed, Wilson said, "because of money limits to develop a test. The more test items, the less money for development of each one, which impacts the quality of the item. Alternatively, what we sometimes get are broader test items that are somehow related to all or a sub-set of standards, but not to each standard." Using California high school math tests as an example, Wilson showed that some standards get no

coverage in the assessment, some get only one item. Consequently, Wilson added, "standards-based assessments do not have high fidelity to standards but are what can be afforded." Even though there is a low density of items per standard, the assessments still maintain the "threat" effect that they might be tested. Consequently, schools and teachers try to teach all the standards.

A few common suggestions to address the above issues are to select among the standards, i.e., identify the "gold" standards, or sample standards over time and rotate their assessment items in the assessment. Another approach is to have a much smaller set of standards, leading to James Popham's recommendation for "instructionally sensitive assessments." It is time to think about more alternatives, Wilson said, noting that the current standards in math and science have been criticized for their lack of depth. Standards need to be interpretable by educators and policymakers—that means short and clear. They also need to enable a long-term view of student growth and allow a more efficient way to use the standards-based test items.

Wilson suggested that learning performances be organized into learning progressions and be measured by progress variables, as one way to improve standards-based assessment. Learning progressions are descriptions of the successively more sophisticated ways of thinking about an idea that students could use, and they allow for more than one way to show competence. Progress variables are an assessment realization of a learning progression. The aim, he said, "is to use what we know about meaningful differences in item difficulty in order to make the interpretation of the results more efficient."

The Berkeley Evaluation and Assessment Research Center (BEAR) has developed an assessment system using the above process, said Wilson, based on principles that are helpful for improved standards-based assessments. The principles include (a) organizing assessment in terms of developmental progress, (b) a strong match between the curriculum and the assessment, (c) emphasis on evidence of quality, and (d) the capability of being managed and interpreted by teachers. The quality principle means that the assessment has reliability and validity, evidence of fairness, and techniques used to achieve this include multidimensional item response models to provide links over time both longitudinally within cohorts and across cohorts. Achievement of meaningful measures is tough under any circumstances, Wilson said, "but it is especially so in an accountability system, so these principles can be helpful."

INTEGRATING ASSESSMENT AND LEARNING

Instructionally Linked Assessments in an Age of Accountability

The essential components of a standards-based education system are having a rough time in current policy and practice, according to Lauren Resnick of the University of Pittsburgh and a CRESST partner. Even though all states now have standards for learning, meeting the first component, "the standards are of varying quality in terms of content and specificity, and were created with differing judgments of quality." As a result, states are beginning to revise some of their state standards.

A second standards-based education component (tests aligned to the standards) has in general been carried out poorly with systematic inclusions or omissions that affect the quality of the process. "Basic knowledge and skills are over-represented," said Resnick, "while thinking, reasoning, applications, and higher order standards are underrepresented. She attributed the poor alignment partially to yearly testing demands and cost considerations, and partially to the public's perceptions that: "A test is a test," without attention to quality or alignment.

The third component, accountability based on standards, has been tied up in formulas because of the requirements of the No Child Left Behind Act. The effects of accountability requirements have been determined by the choices states make, such as the different ways they plan to reach 100% proficiency levels in math and language arts and the specific choices about disaggregation that they make. Resnick also wondered if the weak alignment between tests and standards has created a situation where testing "is dangerously ahead of the intent of a standards-based education system. Have the tests hijacked the standards?" she asked.

The fourth and final component of a standards-based education system is curriculum and professional learning aligned to standards. This activity has been left up to districts (a "local control" compromise) and "attention to it has been relatively late and spotty," said Resnick, "and dependent on the size and capacity of the district." Compromising this component is the view, now controversial, that individual schools should be the locus of instructional and professional decision-making.

The district response to the task of aligning curriculum and professional learning to standards contains three dimensions, Resnick said. These include the

teaching dimension, in which "teaching on the diagonal" should continuously link topical content and modes of thinking; the professional learning dimension, in which professional development of teachers and administrators should be focused simultaneously on the content and pedagogy of the district and supported by coaches, lead teachers, specialists and professional learning communities; and the monitoring dimension, which assures an ongoing assessment of teaching effectiveness.

Interim assessment of student achievement, also called benchmark assessment, is a tool for the monitoring dimension, according to Resnick, and is almost always described as diagnostic and formative. A typical version of such assessments consists of quarterly tests covering the particular standards that the district expects teachers to be focusing on during the quarter. In well-implemented systems, scores are returned to teachers quickly and efficiently, apparently meeting the diagnostic and formative intent of the system.

Resnick described this kind of monitoring as a dilemma, however, drawing somewhat from the experience in the Los Angeles Unified School District, which, she said, "has been very open in sharing information and concerns about the functioning of their interim assessments." One purpose of benchmark assessments is to predict how well students will do on state accountability tests. Another purpose is to see how well students are learning the curriculum that is being taught. Both purposes may be undermined if the interim assessment is not well aligned to the state standards. A third purpose of interim assessment is to diagnose student needs for additional instruction. But if used for that purpose, an important question is how well the benchmark tests will guide teachers and administrators in what they have to do to meet state testing expectations.

The dilemmas raised by interim assessments used for diagnostic purposes go even deeper, Resnick said. Questions need to be answered as to whether they should match the instruction or the state tests, when they should be given, how secure they should be, how much reliability they need, and who is to see the data from the assessments first. In short, she said, we must always ask, "are diagnostic assessments part of the accountability system or part of the teaching and learning system?"

Implications of Natural Information Processing Systems for Instructional Design

There are four principles at work in natural information processing systems, which have a common base with human cognition, according to John Sweller of the University of New South Wales. These include the information store principle, the borrowing principle, the randomness as genesis principal, and the narrow limits of change principle. First is the information store principle, which asserts that cognitive activity is governed by a store of long-term memory. "We are our long-term memories," Sweller said, "and they provide schemas for almost all cognitive activity such as remembering words or solving problems." The amount of information in the information store "is massive, immeasurable, and influences everything we do," he added. Evolution is a natural information processing system analogous to cognitive activity.

Second is the borrowing principle, which contends that virtually all of information in long-term memory is borrowed from the long-term memories of other people. "We imitate other people, listen to what they say and read what they write, which has implications on how you should present material to students and how you organize it," Sweller said. The evolutionary equivalent of the borrowing principle is sexual reproduction.

Third is the randomness as genesis principle, which addresses "what you do when you can't obtain information from someone else's long-term memory." Somewhere along the line knowledge had to be generated, Sweller noted, describing this principle's biological equivalent to be random mutation.

Fourth is the narrow limit of change principle, which contends that change comes in small bits. "Large, random changes in long-term memory would be disastrous," Sweller commented, "because they are most unlikely to be useful and instead would eliminate, previously learned, useful information." Instead, changes come about through our limited capacity "working memory" that only deals with small changes that can be tested for effectiveness before being incorporated into long-term memory. As a consequence, when dealing with new information that is not yet organized, working memory must have a very limited capacity so that new information, when transferred to long-term memory does not destroy the knowledge held in long-term memory. These limitations of working memory when dealing with novel information do not apply to information transferred from long-

term memory. There are no limits to the amount of information from long-term memory that working memory can deal with. That organized information is of most concern to educators.

These principles have large effects on instruction, Sweller contended. The borrowing principle, for example, goes against what many propose, "but if you do an experiment that either shows students information via worked examples or lets them discover it via problem solving, the work example group will do better than the problem solving group." Sweller also described the expertise reversal effect that occurs when, for example, solving problems becomes superior to studying worked examples with increasing expertise. "Knowing that the instructional procedure should change from an emphasis on worked examples to an emphasis on solving problems is useless," he said, "unless the learner's level of expertise can be measured so that instructors can know when to change the type of instruction." He noted that conventional tests do not allow assessments to be changed as expertise improves, which calls for rapid assessment techniques to be developed.

Sweller concluded, "If you are going to design assessments, you really need to make sure you have some conception of human cognition." And, he said, at the end of the day, "you need to place a lot of emphasis on randomized controlled experiments."

Problem Solving Assessments in Games and Simulation Environments

"For at least 30 years, researchers have been trying to understand how to promote meaningful learning and to develop assessments that can measure if people have learned in a deep way," said Richard Mayer of the University of California, Santa Barbara. In his opening remarks about assessing learning in games and simulation environments, Mayer noted that Bloom's taxonomy was a classic attempt to integrate assessment and instruction, specifically to assess meaningful learning at different cognitive levels.

Mayer presented research on the different forms of problem solving presented by games and simulations, and projects that have tried to assess the learning students derive from them. Games and simulations, he said, have the advantage of using visual modes of presentation to help people understand concepts. They also are interactive, utilizing new technologies to make it easier to create environments in which students can learn. He defined instructional games as involving one or more players competing to achieve some goal, which is accountable. An instructional

simulation is a multimedia environment that models a yet-to-be-learned system and allows for the user to respond interactively.

Games and simulations can be especially helpful in assessing students' problem solving, Mayer said. "We do a pretty good job of assessing factual and procedural knowledge," he said, "but it is more difficult to measure concepts and strategies." He described seven projects that involved assessing learners' knowledge gained by using games and simulations including: programming in Logo, designing a plant, and an aircraft simulation. The games and simulations were used in an after-school center. One of the research observations was that the games and simulations provoked the players to ask questions well after the game was over. Another key finding was that the games and simulations helped children improve their problem-solving skills.

Mayer said that problem-solving performance in any domain depends on the knowledge of the learner, and that conceptual and strategic knowledge can be assessed through games and simulations. It is possible, he said, to measure a learner's conceptual knowledge by asking them to specify what happens when a command is executed, such as translating a word problem into an equation or telling what numbers are needed to solve a word problem. Similarly, to assess strategic knowledge, he said, games and simulations can ask a learner to generate or follow directions, describe the output of a procedure, or to make judgments.

Assessment Insights from the Use of Cognitive Task Analysis to Study Expertise Development

Cognitive task analysis is a "protocol for interviewing many experts, one at a time, to capture sequence, alternatives and criteria for decisions and actions needed to perform complex tasks," explained Richard Clark of the Center for Cognitive Technology at the University of Southern California. The protocol includes: retrospection by an expert being interviewed while solving an authentic problem; interviewing three to four experts and distilling their individual solution strategies into one approach based on maximum efficiency and accuracy; and collecting a range of problem examples from experts for use during practice and testing. The goal, he said, is to develop an accurate and efficient worked example of best-solution strategies for a specific task.

Task analysis began when social Darwinism morphed into "equal opportunity to take a test" to qualify for jobs and educational opportunities. Reactions to the

movement resulted in the "scientific management of training and education" said Clark. Frank and Lillian Gilbreth followed with work lasting into the 1930s on the development of behavioral task analysis as the basis for individual/team instruction and testing for the best approach to work tasks ranging from surgical team strategies to bricklaying. "The tradition that came out of the work," he explained, "was not that everyone took a test and, if their scores indicated they could learn to perform a task was expected to achieve without much support, but that the way to support people's learning was to lead them toward expertise." The Gilbreth's task analysis research resulted in the QWERTY keyboard, advanced surgical team procedures and more productive bricklaying strategies.

Cognitive task analysis grew out of the Gilbreths' work, aided through research by cognitive psychologists such as Richard Shiffrin and William Schneider, ATI research and John Anderson's studies on the learning of automated knowledge. In education, "instruction and assessment," Clark said, "are largely focused on the learning of conscious, declarative knowledge, such as learning concepts, processes and principles." In education and testing, little attention is given to automated knowledge -- that is knowledge which has become automated with previous learning and frequent use, like knowing how to read without thinking about the reading process. But "learner decisions, like those made by top surgical experts and bricklayers, are largely automated," said Clark. "They are unconscious and difficult for teachers to describe during instruction." Researchers have found that even experts make wrong decisions a good part of the time, and they underestimate the difficulties of decision-making by novices.

Cognitive task analysis has been proven to be significantly more effective than behavioral task analysis, "think aloud protocols" or "conceptual analysis," Clark said. Researchers have found that cognitive task analysis improves the decisions made by medical students training to do surgery by as much as 50%; that worked examples are significantly better for decision-making among novices than discovery learning.

Clark concluded with five promising suggestions for improving assessment and learning through cognitive task analysis:

1. Focus learning assessment on knowledge automation including speed (secondary, timed response), number, direction (forward or backward reasoning), and the generalizability of hypotheses students generate

2. Understand that during the early stages in the learning of problem solving, decision steps automate first and the links between chunks of steps develop later;
3. Recognize the need for testing of applications of solution strategies to increasingly novel and challenging versions of problems
4. Note counter-intuitive, expert-novice development issues such as the findings that novices have significantly greater memory for details than experts but tend not to diagnose problems effectively;
5. Recognize possible connection between misconceptions and previously automated routines that are difficult to modify in order to learn new routines.

2005 CRESST DISTINGUISHED SERVICE AWARDS

Two long-term CRESST National Advisory Board members and distinguished educators received the 2005 CRESST Distinguished Service Awards. Tributes to Bella Rosenberg of the American Federation of Teachers and H.D. Hoover of the University of Iowa focused on their mentoring, their honesty in advising CRESST policy, and their prodigious contributions to education. Rosenberg, in turn, praised CRESST for educating a new generation of researchers. Hoover considered himself fortunate to have worked with the CRESST people but, most of all, his service on the CRESST board sine it began 20 years ago "has been fun."

Accelerating Future Possibilities for Assessment and Learning

Building on the research described in the previous day's panels and looking ahead to CRESST's new POWERSOURCE grant, Eva Baker, co-director of CRESST, described future possibilities for assessment research in the context of political realities. Leading off the panel discussion on the future, she warned that the research and development community has not recognized the trade-offs that are occurring in the rest of society and applied them to our research. It is a policy environment, she said, with these R&D principles: now rather than later, partial rather than full solutions, and lower cost rather than higher quality. These realities for the research and development community must be addressed, Baker said, else we risk "the abandonment of key tenets such as research validity and/or the loss of resources in the future." Research will continue under the pressure of policy requirements such as "short timelines, a tendency to buy on the basis of availability, resistance to understanding processes or consequences, and a lack of acknowledging that high quality research requires substantial investment."

Baker outlined new goals for assessment R&D. Researchers must show benefit over current options; simultaneously maintain or reduce costs; increase the velocity of outcomes (research, development, and evaluation, e.g., a process that in business is called the stretch); and increase feedback, devising "appropriate interactivity with the market and with individual consumers."

One possible method for meeting these new goals is to use learning in our assessment designs, Baker said. For example, focus on big ideas, the principles, themes, and schema that do not start with standards per se but with the main ideas underlying them. The assessment benefit is that the items are reusable, recurrent, and spiral.

Second, use knowledge representation to show relationships within and among key content and tasks. This strategy provides maps to monitor progress and develops teacher knowledge. Templates for administration and automatic scoring are also part of the product. Knowledge maps developed by CRESST are one example.

Third, integrate learning in assessment design through explanation, using the substantial evidence available for instructional and self-assessment. This means "telling what's going on," Baker said. "If a kid's explanation is incomplete, this is evidence for the teacher to find out what's going on. Kids also see how things go together when they generate an explanation." The assessment benefit is the diagnostic and feedback value. The procedure would be to use existing scoring rubrics, peer assessment, or automated scoring, or a selection of best explanations.

A fourth way to use learning in assessment design is through integrated worked examples, as described by previous presenters John Sweller and Richard Mayer. These provide a total picture of the goal in procedural or problem solving, Baker said. The assessment benefit is that they are completely scaled a priori and empirically integrated with instruction.

All four of these strategies, Baker said, support learning, teaching and transfer, and are being applied in the new CRESST POWERSOURCE grant. Using an experimental design, CRESST researchers are developing formative assessments in middle school pre-algebra and algebra and will measure their impact on learning and knowledge acquisition. The project, she said, will be an example of assessment R&D that is faster, cheaper, and of high quality; thereby meeting policy and practitioner needs.

Some Implications of Expertise Research for Educational Assessment

The way assessments have been built "often runs counter to how people use knowledge," said Robert Mislevy of CRESST/University of Maryland in introducing his work on assessment task designs. He shared with the audience his latest work focusing on trying to determine "in detail the knowledge, representations, development and contexts of expertise." The results, he said, can ground assessment arguments and task design, recasting them in terms of purpose, perspective, principles, and structures. The results also will be especially valuable, he said, to increase our understanding and assessment of complex proficiencies, situations, and performance. The assessment questions arising in expertise research were suggested by Sam Messick in 1994: First, what complex of knowledge, skills, or other attributes should be assessed? Second: what behaviors or performances should reveal those constructs? Third, what tasks or situations should elicit those behaviors?

Mislevy shared with the audience his colleague Irving Katz's work on the Architectural Registration Examination. The new computer-based architectural assessment puts a premium on thinking, not drawing, replacing a more traditional 10-hour, hand-drawn design problem. This work has already yielded a number of interesting and useful findings for task design. First, experts understand fundamental principles and act based on those principles rather than surface features of a problem. In architecture, Mislevy noted, constraints are a big factor such as the number of constraints, variation in importance and difficulty, and degree of conflict between the constraints. Architecture experts studied a problem longer, thought about which constraints would be hardest to solve, and solved those issues first. Thus expert revisions only required reworking, not a fresh start.

A second lesson for task design is the importance of interaction with a situation. "You don't write a paper without revisions," Mislevy said, and intelligence is built into situations, tools and processes. Experts are attuned to situations and go through cycles of hypothesizing, testing, and revising; or asking "what can I do to make things better?" A third lesson for task design is the value of external knowledge representations such as maps, forms, symbol systems, diagrams, charts and blueprints. They have roles in practice, such as gathering, sharing and transforming knowledge; and roles in assessment.

Cognitive task analysis can help ground assessment in terms of both validity argument and principled task construction, Mislevy said. Design patterns emerge

when looking across domains to find recurring difficulties, blockages, or overloads, particularly when using technology. Classes of expertise arise in many domains, he added, such as designing under constraints in engineering, problem solving (troubleshooting), or model-based reasoning in science.

In conclusion, Mislevy emphasized that insights from expertise research can improve the practice of assessment. Suitable conceptual frameworks, tools and exemplars, and design and delivery frameworks—all based on expertise research—are key to making technology-based and complex assessments possible.

Assessing Learning in Video Games

The use of video games as an assessment tool of adult learning is a relatively unexplored research area, according to Harry O'Neil of the University of Southern California/CRESST. Video games have great potential due to their widespread use and intrinsic motivational factors said O'Neil, reporting on research on the use of video games. The motivation to play a computer game comes from the challenge, the complexity, and the fantasy it creates.

It is important to understand what a video game is, said O'Neil, who explained that they usually have four components: settings that are real or imaginary; roles or agendas for the participants; rules; and scoring, recording, monitoring, or other kinds of systematic measurement.

The purposes of testing and assessment of learning in video games are both system-oriented and individual/team oriented, therefore calling for "a new set of boxes in the CRESST model of learning," O'Neil said. In playing a video game, learning and assessment factors arise including test anxiety, affective learning, effort and self-efficacy, and academic goals. To use video games to support learning and assessment, it is essential to specify what is being taught and what is expected to be learned, said O'Neil, who said, "It is a shock to game developers to have instructional goals and objectives."

Although the research indicates that computer games are potentially useful for instructional purposes and may provide multiple benefits, O'Neil could find only 19 usable studies done since 1990. Part of the problem is the different mental models between developers, "the bottom line is if it makes money", and the educators/trainers who believe that learning and evaluation are key goals. Major

research questions include whether more game-based training leads to better game performance and if video games improve cognition.

Research indicates that when instructional strategies are embedded in a game as an inductive discovery approach, they usually don't result in learning. Another finding from the literature is that games are not a good strategy for individuals with low prior knowledge. A third finding is that when examining collaborative problem solving in game play, then one test item may be enough. If it is teaching a concept, however, then more items are needed to produce a valid result.

O'Neil and other CRESST colleagues have been working with the U.S. Navy on the use of video games in distance learning. The research team developed a core set of research-based distance learning guidelines for such components as instructional design, multimedia use, learning strategies, assessments, management strategies, self-regulation and motivation.

O'Neil presented an example of a CRESST problem solving study in a gaming environment. The video game was a puzzle called SafeCracker©. As part of the game, participants role-played "cracking" various safes in a mansion. The assessment involved the measurement of adult student content understanding using knowledge maps, domain-specific problem-solving strategies using essay questions, and self-regulation using a questionnaire.

There was an increase in problem solving performance as a result of the game, O'Neil said, but it was small. Existing instructional strategies (discovery learning) in the game were not effective.

The "walk-away, elevator speech" on this research, O'Neil said, is that "off-the-shelf games do not teach anything." More research is needed on a game designed with effective research-based instructional strategies, he added, in order to determine if game technology can be leveraged for training.

Accelerating the Future of Technology-Enabled Measurement

The future of technology-enabled measurement of human performance began in the 1960s, said Greg Chung of CRESST/UCLA, when NASA began using performance sensors to check the health of astronauts during shuttle flights. Today, sensors can be used to detect almost anything, from the temperature on Mars to the number of steps and time it takes a student to link the nodes on a complex computer-based knowledge map. Technology-enabled measurement can "clear up

the details" and increase the scope, frequency and resolution of what can be measured and learned about performance. It is an attempt "to open the black box of learning," Chung said.

Chung presented results from a current UCLA study where a research team is investigating technology-enabled measurement in a large electrical engineering course. This is a class, said Chung, in which the instructor usually "marches along in his or her direct teaching, with little direct feedback from students." In the study, students were given computers with instant-response technology, during which they could anonymously send feedback to the professor on his or her understanding (or lack of understanding) of material as it was taught. The instructor, ranked among the top 10 professors at UCLA, was astounded at the clear feedback and number of questions from the students. Even students in advanced classes asked basic questions, which convinced the instructor to change his instruction. The instructor also found that the technology-enabled measurement "exceeded the interactivity of an hour-long, one-on-one office visit with the student."

Chung presented findings from second CRESST study, a prototype effort related to U.S. Marines marksmanship training. To better assess Marine marksmanship and inform training, the researchers developed a method to observe fine motor control of marksmen in the critical six seconds before shooting. Putting a sensor on the rifle's trigger allowed trainers and researchers "to see what you can't see in conventional observations," such as if the marksman yanked the trigger, which greatly affected the accuracy of a shot. The key finding was that the sensing data appeared to track well with the construct being measure. For example, we could observe the difference between a slow steady squeeze and a quick yank—this difference showed up in the waveform and the shape was obvious. Of course, the next step is to validate the machinery under more rigorous conditions. There was no impact on USMC training, as the research wasn't far enough to even show them stuff.

The prototype system demonstrated the potential of using sensors to provide diagnostic information during the crucial seconds before firing, Chung said, and the potential utility of this kind of information is to provide support for individualized diagnosis and prescription in ways that went far beyond mastery of learning. Admitting that it is a really bold assertion, Chung predicted that technology "will revolutionize the assessment and learning field, and broaden the way we think about how we measure human learning." Much as the microscope or telescope are

married to the science they are investigating, technology-enabled measurement will become the observation for assessment in education, he said.

Making Evidence-Based Practice a Reality in Classrooms

"If the American curriculum is at fault for being a mile wide and an inch deep, then how deep should it be?" asked David Niemi of CRESST/UCLA during his conference presentation. Displaying one of the most famous equations in mathematics, $e^{1}i + 1 = 0$, he asked the concrete question: What would you have to know in order for this equation to be meaningful and useful? Among other things, he suggested, you would have to be able to answer questions such as: What do the symbols mean? What is this equation about, what can you do with it, what's important about it, and how does it connect to other topics in math? If you could answer these questions, he said, you would have some understanding of the equation, but most students in most classrooms do not have anything like this kind of understanding of the mathematics they are studying.

To develop better strategies for teaching and assessing understanding in mathematics and other curriculum areas, CRESST researchers have been testing new methods for analyzing and mapping subject domains. Based on research on subject area expertise, we identify central principles of the domain, Niemi explained, then build maps of other kinds of knowledge around those principles. These maps provide a foundation for advanced learning and assessment design. "What is known about the domain," said Niemi, "drives performance and assessment." He illustrated this point with an expert map, showing the principles of mechanics, that has been used to develop assessments and instruction in middle school science.

Experts in many fields Ñ science, math, history, writing Ñ have highly structured knowledge, he said, and their knowledge is organized around a small number of concepts and principles, the big ideas. Teachers may not always be familiar with the big ideas in the domains they are teaching, but experts have no trouble identifying them. For example, biologists say that the principles of evolution are among the big ideas in their domain, geologists mention plate tectonics, and physicists, Newton's laws. By working with experts to determine the big ideas and related ideas and skills in a domain, CRESST builds maps, or ontologies, that visually lay out this information for teachers. Sharing CRESST's ontology for Algebra I, Niemi said that the map displays all of the central principles with the concomitant ideas and skills for each principle.

Using Assessment to Improve School and Classroom Learning: Critical Ingredients

The current standards-based reform and assessment environment has its roots in the work of psychometricians and psychologists at least as far back as Robert Thorndike in the early part of the century, said CRESST co-director Joan Herman during her presentation. A more recent basis for standards-based reform draws from W. Edward Deming's work in the 1980s—establish goals, assess progress relative to goals, use results to inform planning and decision making, implement plans, measure performance, and keep recycling the process until all goals are met.

While setting rigorous standards and measuring progress is not a revolutionary idea, some new aspects have developed in the past 20 years, said Herman. One change has been greater acceptance of the idea that assessment not only provides information about learning, it also drives learning. NCLB holds everyone accountable including students. Another change is that assessment today is part and parcel of the learning process, said Herman. "Feedback can be an occasion for meaning-making, she said, "and for kids to make connections between old and new knowledge." Assessment also is being used as a technical system embedded in socio-political and learning systems. In short, "assessment today plays many roles in a complex system of learning and accountability," she said.

But what is good assessment? Herman asked. "Good assessment starts with, communicates, and enforces meaningful goals for student progress," she said. Good assessment operates in an aligned system, and for it to work, not only must the assessments be good, but results must be well analyzed and used to do something to improve learning for kids. "If you have lousy data, no matter how well massaged it is, it is not going to be useful," she said. "If the data are not going to be used, what's the point?"

"Too often in the field today," Herman explained, "we have some people concentrating on data quality and paying attention to whether and how the data get used, while others are concerned promoting data use and not sufficiently worried about the quality of the data." Both are important said Herman, adding that quality data begins with quality assessments that sets clear and accurate expectations, are valid for intended purposes, are aligned to learning goals, and are fair. Are these the kinds of assessments teachers are using informally?, Herman asked. In her opinion, we are not quite there: "These types of assessments are often nonexistent in

curriculum materials, in teachers' repertoires, and not often found in benchmark tests." Quality data use supports student learning, said Herman when it is timely, provides quality feedback, strengthens teaching and learning, and involves students in accountability. Quality interpretation of assessment results requires appropriate criteria and sound analyses. "A major shortcoming is that many teachers tend to not review student work on a timely basis," said Herman, "which is a critical component of high quality data use."

CRESST's new Powersource grant addresses the policy and practice implications of developing a quality assessment system that supports student learning, Herman said. It will link assessments to key principles in the discipline (middle-school math), provide professional development on content, pedagogical and assessment knowledge; conduct research on ways to reduce the analysis and scoring burden; provide models for teachers' routine practice; and utilize technology to bring the assessments to scale. The ongoing support for teachers, Herman said, is an important component. Teachers do not have time, she noted, to devise these types assessments on their own.

Visions from the Vortex: What Evidence Classroom Teachers Need to Clarify State Standards and Assessment

"I'm here to tell you why evidence-based practice is desperately needed in classrooms," said Geno Flores, then deputy superintendent for assessment and accountability at the California Department of Education. He described a data gathering and reporting system that could not help but overwhelm, confuse, and frustrate classroom teachers. His description underscored David Niemi's contention that teachers are awash in data but too often unable to use data to increase learning.

The California assessment system, Flores said, consists of STAR (Standardized Testing And Reporting), established by the legislature in 1998 for all students in grades 2-11. The state tests 5 million students in English language arts and math, in history in some grades, and in science in some grades.

From these tests, the state reports results by district, school, grade, sub-groups of students, demographics, program participation, five levels of proficiency, scale score and percentage correct. From the WMB—weapons of mass bureaucracy—we report trends, strengths, problems, and acknowledge achievement gaps," he said. The state rolls out a series of reports—cluster reporting, average percent correct for minimally proficient, minimally advanced and all students statewide. But, Flores

said, we caution teachers (on this web-based reporting system) that some results are based on small numbers of test items and, therefore, may not be reliable or generalizable. The state gives an analysis report on students with the percent correct obtained by a student on the cluster score, along with the class score, "but we tell teachers to not compare individual scale scores." In other words, "don't analyze the data."

The situation is "utter confusion," according to Flores, who inherited the reporting system. The tools the state gives to help teachers are broad but insufficient and include academic standards, state-adopted instructional materials, state-approved professional development institutes, state test blueprints, and released test items.

Flores compared the environment for teacher use of data to the pronouncements on the first day of pro football. The analysts have gone over all the data and made clear predictions based on their data. In education, he said, teachers need the same kind of evidence—"to tell them if kids are ready for the big game." The kinds of bureaucratic data the state provides is not enough, he said, and called for resources to help teachers clarify the standards and to understand the expectations (as measured by tests), clear ideas of required grade level knowledge and skills, local evidence that students have the skills (formative assessments), and evidence that student status will meet end-of-course expectations.

Commenting on the presentations, Robert Glaser of CRESST/University of Pittsburgh wondered: "Why does this panel make me feel depressed, thinking of all the work that needs to be done?" The panel revealed excessive burdens on teachers, he said, and indicated that teachers are not clear on what instructional strategies are open to them. Geno Flores said that state education agencies would improve the expectations because standards are politically attractive, "but they need to decide what they mean by rigor and levels of expectations." According to Joan Herman, the problem is to get specific on learning goals. When that happens, she said, "it is possible to align instruction with goals, and instruction and assessment activities can be the same." CRESST's research suggests, she said, "that one of the reasons teachers don't use assessment outcomes is because they feel incredible pressure to move on, and they either don't have the authority or decision making capacity to stop and go back and reteach."

David Niemi pointed out that in middle schools, the typical teacher has more than 100 students, and teachers initially need to teach where most of the group is at. Brian Stecher, addressing the "span of teacher action," said one of the issues not addressed even in break-the-mold schools is the limits of the traditional schedule, "so we keep assigning teachers into the same model."

A Century of Testing: Ideas on Solving Enduring Accountability Assessment Problems

The final conference panel turned to what other countries are learning as they refine their accountability systems. Barry McGraw, director of the Directorate for Education in the Organisation for Economic Cooperation and Development and a leading expert on Australia's assessment system, said much has happened in the field internationally, especially on linking assessment to teaching and learning. Countries are refining their system monitoring, he said, especially with regards to effectiveness (quality), equity, and efficiency (value for the money). International studies, such as TIMSS and the OECD Program for International Student Assessment (PISA) are relating assessment results to the social backgrounds of students, where gaps are "much less steep in some countries, showing that we can achieve equitable results," said McGaw.

A key problem in measurement remains, however: the point of reference for judging individuals. The international assessment community is searching for an external criterion, McGraw said, "that is less interested in judging one person against another than whether an individual is growing." He described the technical aspects of measurement approaches with candor: "Norm-referenced testing is gung-ho psychometrics for what is often not important, while criterion-referenced testing has less robust psychometrics for what is important."

High stakes assessments are used world wide, according to McGraw, and a substantial number serve dual purposes. The purposes include certifying what students know and are able to do at the end of secondary school, and university admission. McGaw believes that such tests should measure what students have been learning in school, not the content as defined by a college entrance test such as the ACT or SAT. McGaw argued for a common curriculum across schools, pointing out that only a very small percentage of students take highly competitive courses.

Grade distributions are another important international issue as they traditionally have been used to monitor standards. This means, said McGaw, that

"the failure rate is used as a measure of maintaining high standards." A common opinion is that if more students take more challenging classes, grades should decline in order to assure than an "A" is still an "A." In other words, do enough students fail?

McGraw described the assessment systems in two countries, which are marrying criterion and norm-referenced assessments. England uses criteria set by a committee to define some grade boundaries, based on three pieces of information each year. These include a review of the previous year's scripts at grade boundaries, reference to prior grade distribution, and reference to evidence of change in the student cohort to justify shifts in grade distribution between years. Officials claim that the process provides consistent reference over time. But the media believes that if performance goes up, it is only because adjustments have lowered the standards, that is, students are not really improving.

McGraw provided details, including examples of reporting forms for the assessment system in New South Wales, Australia. The scaling process for the New South Wales Higher School Certificate has three stages. In Stage I, experienced examiners independently form an "image of bands" and set cut marks for each band boundary on each question. In Stage II, examiners work together to reach agreement on band locations for bands on each question and for total scores. In Stage III, student work (at boundaries) and total scores are reviewed, cut points determined, and boundaries are located on a marked scale. Student results are reported on a Higher School Certificate that lists all high school certificate courses with assessment marks, narrative descriptions of what students know and can do, exam work, and performance bands. The report also shows how all students performed on the assessment, the minimum standard expected, and the school assessment mark.

Despite these sophisticated refinements, the work "does not always change the debate," McGraw noted. A federal minister found an English paper that was awarded a passing mark despite some inadequate expressions within it and concluded that "far too few students were being failed." The debate again is about desirable failure rates, he said, "and it is important for such debates to be constructed as a debate about the nature of performance that is judged inadequate."

The Impact of State Accountability Systems on Classroom Practices: Successes and Enduring Problems for Teachers and Schools

States have been focusing more on the quality of their tests than on the tests' effect on classroom practice, according to Hilda Borko of CRESST/ University of

Colorado at Boulder. To fill the gap, Borko and colleagues at the University of Colorado, in collaboration with RAND researchers, analyzed classroom-level accountability effects in two states. Kentucky was chosen because it was an early implementer of comprehensive standards-based reforms and in the forefront of standards-based assessments. On the other hand, Washington State implemented its accountability system later and much more gradually, slowly phasing in the Essential Academic Learning Requirements (their state standards) and the Washington State Assessment of Student Learning.

The research team looked at writing and mathematics in the elementary and middle schools, focusing on the tested grades. RAND conducted surveys while Colorado researchers conducted case studies analyzing school and classroom factors linked to improvement.

Borko found three major areas of school and classroom practice that changed as a result of the new accountability systems. First was the allocation of instructional time. In Kentucky, schools and teachers allocated more instructional time to those subjects and grades that were tested compared to the subjects and grades that were not tested. For example, 92% of 5th grade teachers increased their math instructional time (tested at this grade level), telling researchers that a primary reason for the change was the state assessment. Similarly, Washington State teachers increased their instructional time in reading, writing, math and communications and decreased the time allocated to untested subjects.

Teachers also made changes in instructional practices. In writing, Washington teachers shifted their instruction to better align with the new state standards, for example, they made sure to cover writing for different audiences, purposes, and in different genres. In Kentucky, students had more conferences with teachers and more discussions about their portfolios that were part of the state assessment. Traditional practices still prevailed in math, but the biggest changes were on reform-oriented math content, such as more emphasis on probability and statistics, and on algebraic ideas. Teachers were still directing most math activities, but Kentucky students made increased use of manipulatives and real-world applications, and in Washington, students explained their math thinking rather than just solving problems.

Third, both Kentucky and Washington principals and teachers felt increased pressure from the new tests, and teachers spent a fair amount of time on test

preparation. In fact, Borko said, "they didn't want us in the schools from March to May," the usual testing window.

For each of these changes, Borko asked if they represented "success or problems." Was emphasizing one subject over another a good thing or not? The researchers did see patterns of change, "but the change was gradual and on the margins," said Borko. "Teachers were not doing major overhauls but adding to existing repertoires," Borko added. In most cases, the changes were the easiest ones to implement, not necessarily the best. Borko noted that the findings were consistent with other research—that classroom instructional reforms are usually limited and piecemeal.

To produce substantive changes in practice, Borko said, "we need to increase school and teacher capacity and target it to the reform agenda." Educators need strategies to address the negative aspects of teaching to the test, including testing in more grades, testing more subjects per grade, and changing the content and/or format of the tests from year-to-year. They must enhance their repertoire of instructional strategies for helping students develop deep conceptual understanding of mathematics and literary skills to produce sophisticated writing in multiple genres. Teachers and principals have professional development in each state, but it is inadequate to meet the desired outcomes of the intended reforms.

A Futures Market for Educational Assessment?

Drawing from his Iowa setting, Stephen Dunbar of CRESST/University of Iowa, noted that "where I come from, people bet on corn that hasn't been planted and on hogs not yet born," a fitting metaphor for his discussion of possible breakthroughs in testing.

First, however, Dunbar added to the historical picture drawn at the CRESST conference by giving a "lesson learned" from the Iowa Test of Basic Skills experience with innovation. In 1955 it introduced a series of assessments for grades 3-8 that were published in a single booklet as paper/pencil tests. It used an optical scanner for grading. All levels were in the same test booklet and could be assigned to individual students, a form that is called a flexi-level test today. This was cheaper for schools because they could order only as many copies as the largest class.

The flexi-test form faded away when test providers realized they could make more money if the tests were separated. More important, however, was the

influence of the optical scanner, which, Dunbar said, "meant that in Iowa, testing went from open-ended questions to multiple choice." Technology, he added, "doesn't always lead in directions you might hope," and in today's environment, technology for delivery systems has not had as much influence as expected. In high-stakes situations, he explained, the demands on the item pool limit computerized adaptive testing. But the message from his Iowa examples was this: the hardware, the scanner, survived, but flexi-level test design did not because "it required sophisticated understanding of assessment."

As to the "assessment futures" of the present, Dunbar said that online assessment is producing many examples of requests for proposals where some kind of administration via the computer is needed such as adaptive assessments and make-ups. Nearly 100% of the resources are going into the platform for online assessments "but not into assessments themselves. We are getting less complex items because those are the ones most easily adapted to the computer."

Discussing challenges for new test design, Dunbar said that for the last 15 years the attempts to develop elegant theories of learning and assessment "have tended to be in a very narrow content domain, typically aimed at adults or areas where we have expertise." Applications in early reading or math skills have been lacking. . He called for an acceleration of technology into domains of traditional school learning, where group differences tend to be vast, "but we lack experience in how to scale up this area." All areas, he added, "require us to find the right distance beyond our professional comfort zone."

Dunbar predicted "assessment futures here and now" in: technology-based database development and management; online reporting and software tools for presentations and use of test results "to give teachers something understandable;" standards-based reporting with running records from so-called formative and benchmark assessments; and a "dramatic" effect in the next five years on the ability to track students, "true longitudinal reporting for the second generation of NCLB and its AYP."

As for games and simulations, Dunbar said the big question is whether they can be used for assessment through embedded tasks, sequences, scoring units, and overcoming the "signal-to-noise" and its effect. Principles of test design and test theory are needed for effective assessment in game and simulation contexts, which

he believes would happen in the next 10 years, but "the really hard part will be in getting content that is useful."

So, Dunbar asked, is a breakthrough coming? "If I were projecting," he said, "I wouldn't feed a big risk in saying that a lot of paper in assessments will be removed." Existing tests will be put on line, technology will be used for immediate reporting, and there will be enhanced capabilities to use tests for accountability, but there will be limited use of test results to inform instruction and learning. The problem, he said, "is that we will have such data capabilities that we will forget what the data are for."

There will be no real breakthroughs, Dunbar said, unless there is "something we haven't heard about." As for a 25-year Futures Market: the breakthrough will come "when a new generation decides what we can do with assessments in the context of technology-based learning."

No Child Left Behind: Accountability and Assessment of Science Achievement

Richard Shavelson of Stanford University/CRESST framed his description of work on the upcoming NAEP science assessment, scheduled for 2009, with a philosophic exploration of the democratic foundations for accountability. Officials, he said, "are responsible for and are obliged to be accountable for their actions and outcomes." Underlying the notion of accountability, Shavelson said, is the belief that rationality "can be imposed on social and political life through planning, intervention and monitoring. Beliefs such as these make reasonable the expectations of accountability." In other words, "accountability creates the presumption of capability and causality and a presumption of choice, or the freedom to act intentionally."

There are complications in accountability, however. "When actions cannot be observed directly, conformance must be inferred from observing outcomes," Shavelson said. "The quality of inferences that are possible is limited by the complexity of the relation between outcomes and actions." With this context, Shavelson discussed the process that he and Senta Raisen have been involved in—creating a new NAEP framework for the 2009 science assessment. The content areas to be covered in the NAEP science assessment will include physical, life, and earth/space science. "The most significant thing we had to do," he said, "was to reduce the amount of content coverage. There will be a substantial focus in the NAEP on areas that are important."

The assessment also will deal with processes, what kids do with content. The assessment will measure four student outcomes: being able to identify various pieces of information; using scientific knowledge that will draw on students' mental models of how the natural world works; using models of real world situations; and inquiry—not only how to conduct inquiry, but also how science justifies claims about knowledge.

The item types will include multiple choice, cluster items (addressing a particular idea), and prediction, observation, and explanation items, Shavelson said. In addition, constructed response items will include concept mapping tasks, hands-on performances and interactive computer tasks. Computers also will be used for information searches, performance assessments, and simulations. Shavelson noted that, prior to this new science framework, NAEP does not used computers in its assessments.

Returning to his principles of accountability, Shavelson said that NCLB accountability holds a potential for harm that is greater than its help, "if we are not careful." Accountability is a very delicate concept, "and in education we are dealing with an environment where we do not have adequate control over the educational, social, economic and political environment." He said he was raising the issue because "we need to look at accountability as not only holding schools accountable, but as a way of creating improvement." With accountability comes the notion of sanctions, and NCLB is "heavy duty" on sanctions. "It seems to me," he said, "that we have a problem because accountability assumes that we have control over our environment."

Shavelson concluded by stating the problem as he sees it: "What can we reasonably expect when education is a political activity in which public officials do not have complete control over production processes? Should the focus be on actions primarily? On outcomes? What balance is needed?" He also wondered if NCLB has the sanctions issue right—"what kinds of sanctions will lead to improved performance? Rather than beating a horse, what alternatives are there for improvement, such as the combination of audits with reasonable assessment?"

Shavelson noted that "you can punish a system, and it will respond to the punishment, but as soon as it is released from the punishment, it will spring back to where it was before." A lot of thought is needed, he said, "about what we can do as an incentive system to improve education and learning."

Ongoing Research, New Understandings an Essential Part of the 2005 CRESST Conference

The 2005 CRESST national conference celebrated the history of assessment research as part of its 20th anniversary program, but also reported on current research that is contributing to policy and practice on issues confronting American education. In small concurrent sessions and expert panel discussions, participants at the conference explored findings on major challenges brought about by the intensified national interest in assessments and test-based accountability. The quality of the research presented points to the confidence shown in CRESST's work and its continued support from the U.S. Department of Education through a recent contract that will study formative assessment strategies in middle-school mathematics.

One of the most important needs across the country is the development of fair and valid assessments of English-language learners and special needs students. This is a continuing interest at CRESST, and in a session at the 2005 conference, Jamal Abedi, who has been involved in the center's ELL research over many years, presented the latest knowledge about opportunity to learn policies and assessments of English language learners. Now a professor at the University of California/Davis, Abedi described CRESST studies that examined differences between eighth graders who received accommodations and those who did not.

Providing accommodations is important, Abedi said, because assessments may be culturally and/or linguistically biased and students' English language limitations may prevent them from demonstrating how much they know. One of the least effective accommodations found in this study and in others, was giving a test in a student's native language when the language of instruction and language of assessment are not aligned. Similarly, modification of the test items in this study did not produce a significant difference in ELL performance when compared to their non-ELL peers. However, we found that on a smaller group of items with more complex linguistic structure the performance gap between ELL and non-ELL student was reduced significantly, said Abedi.

Abedi warned that accommodations used for students with disabilities may not be appropriate for all ELL students, despite a relatively widespread practice of doing so. A review of state-approved accommodations several years ago found 73 types in use, but two-thirds of them (47) were not relevant for ELL students, he said.

A key finding, said Abedi, was that ELL students with high opportunity to learn, that is, having a teacher with high content knowledge and being in a high ability algebra class, outperformed similar students with lower OTL, that is, having teachers with less content knowledge and in a lower ability class. This finding has important implications for improving OTL for English language learners, said Abedi.

Research on accommodations for students with disabilities is continuing at the same time that the field is moving toward universal design of assessment, which implies that the need for accommodations will decline, according to Martha Thurlow, director of the National Center on Educational Outcomes. They will not be entirely eliminated, though, because universal design cannot address all accommodation needs. Thurlow cited research showing that teachers were not well informed about the use of accommodation policies. They often were unable to determine if an accommodation was standard or non-standard, according to state policies, she said, and one-third did not know the effect on students' scores of using a non-standard accommodation.

Because of validity issues and the interaction of specific disabilities with the focus of certain test items (e.g., students who are deaf and items that require them to select words that sound the same), Thurlow recommended that policy decisions depend on being able to answer five questions:

- Does the skill in question reflect a standard that is to be assessed?
- Is there a clear match between the breadth and depth in the standard and how the standard is reflected on the test?
- Is there an accommodation that can be used by a student, even though it might not be used for other parts of the assessment?
- Is there an alternative skill that could be used for students whose disability precludes performance of the skill?
- Is there a way to score the assessment so that the student and the school are not punished because the student's disability precludes performance on the skill?

Academic language is different from conversational English, previous research has shown, so an opportunity-to-learn issue for ELL students is gaining a knowledge of the academic language, Christy Kim Boscardin and Barbara Jones of

CRESST/UCLA, said in introducing their research. ELL students, for example, may understand the forms to be used in essays, but they lack the understanding of the language that need to be used for a characterization study. The researchers developed a teacher training program based on the Functional Linguistics Model, which organizes language functions around such tasks as description, explanation, definition and persuasion. The purpose of the program was two-fold: to improve teachers' understanding of key components of academic language and use in instruction; and, to give teachers tools to give ELL students direct instruction on academic language.

The experimental training program involved teachers from three urban middle schools in California. A four-day workshop was followed several months later with a two-day observation and follow-up session. The study used a performance assessment, a teacher survey, classroom observations and teacher interviews. Students whose teachers implemented the academic language training in their instruction had higher performance on district tests and also had more access to classroom discussions than students whose teachers did not have the training. The trained teachers also provided better feedback to their students on their writing. The researchers also learned that more experienced teachers—those who had been in the classroom more than 10 years—were less likely to be flexible in their instruction.

Innovative Measures of Learning and Instruction

CRESST researchers are exploring several challenging ways of measuring learning and instruction using both technology and a focus on "big ideas."

Two members of the Technology-Based Assessment of Literacy and Language (TBALL) project team, Margaret Heritage and Markus Iseli, reported on the development of speech-recognition software to assess early literacy skills in both native English speakers and Mexican English-language learners. The purpose of the software, said Heritage, is to reduce the time it takes for teachers to perform one-on-one literacy assessments in early grades, and to reduce the variability in teachers' scoring. The development process is interdisciplinary, drawing scholars from three universities (UCLA, UC/Berkeley and USC) and experts from five disciplines (electrical engineering, computer science, education, psychology and linguistics). "One of the most amazing parts of the project," said Heritage, " has been its capacity to get people across disciplines to discuss each other's areas competently."

The software must meet challenging criteria. It must quickly provide instructionally useful information to teachers, -be sensitive language knowledge and to variations in pronunciations, and automatically score and analyze children's responses to the assessment tasks. . A second TBall member, Markus Iseli, described the challenges in creating speech recognition that is sensitive to a variety of differences such as gender, dialects, classroom noise and even health. "A child with a cold doesn't sound the same," he said. Children's speech also is not like that of adults because the physiology of speech is still developing. The design of the software is taking into consideration who is going to use the system, what tasks does the system need to do (recognize words and pronunciation errors), and the environment in which the software will be used.

TBall presents students with words, pictures, and sentences on a computer screen and then logs their responses into a computer. "All students take benchmark assessments," explained Heritage, "and some students will get a 'drill down' if the teacher finds that the student is not meeting the benchmark standards so that teachers can have diagnostic information. . Because some teachers do not know what to assess and when, the team is developing guidelines, using the basic reading skill components from reading research.

Another CRESST project is addressing the limits of only aligning assessments to standards and, instead, taking a "Big Idea" approach. Researcher Terry Vendlinski, who also teaches middle school algebra, brought a practitioner's view to the standards movement, saying that it had positively affected teachers' instruction by moving teachers to align their instruction with standards, "but that has not always been the only cure that is needed." He noted that standards are not at the same at each grade-level across the country, a lot of the standards reform "is just about knowing, not about understanding," and often it limits teachers' focus because the standards address learning in a specific grade or area and don't always have a clear link to how students must actually use or build on their learning in the future. Similarly, when teachers organize instruction around a specific test, "kids often don't grow in knowledge as much as is thought," he said.

"The question I always hate the most as a teacher is: "How am I going to use this?'," Vendlinski said. Actually, it is an important question to answer not only for students but also for teachers "in order to get them to look ahead and not just at grade level standards." CRESST researchers, he said, are developing an assessment model based on students creating their own understanding and connecting new

ideas to what they already know. "If they don't construct the meaning," he said, "they often don't retain it."

The research, according to Vendlinski, is looking for organizing features, and the cognitive demands inherent in understanding a concept. He noted that experts "tend to connect things in ways that get at key organizing principles, not superficial facts." Novices, on the other hand, tend to memorize facts and not make connections. In education, he added, "we find that assessment and instruction often focus on disconnected facts too, so how can we expect kids to do anything but the same?" When teachers are asked to develop a rubric for scoring the assessments they give, he said, "our research suggests that their understanding of what they are assessing also begins to improve because they are not just giving a test to kids and hoping something happens before understanding it themselves."

Being attentive to the "Big Ideas" would allow standards to be attached to a framework (an ontology) with assessments included in the overall organization of concepts. In addition, once a Bayesian Network is overlaid on such an ontology, everything could be tied together, he said. When analyzing assessment results, Vendlinski said, it should be possible with some certainty to distinguish between passing a test, to the actual understanding of a concept. Bayesian ontologies, for example, allow "the system to make probabilistic calculations that suggest to teachers what to assess next." One other advantage, he said, is that an ontology approach not only points to different ways of instruction, it can inform teachers how standards are interconnected and in which order they might best be taught.

Recent Research on the Evaluation of Accountability Models and Educational Innovations

With so much attention to the AYP status model of accountability, there is now less confusion between it and the growth model of accountability, Pete Goldschmidt of CRESST/UCLA, told a concurrent session. At the same time, however, confusion still exists between growth models and value-added models, and the choice often comes down to how a state wants to define performance.

A status accountability model, he explained, determines how a school is performing at the time of the assessment, "irrespective of everything else that is going on." It assumes that all student success is due to what has been taught and learned in the current school year, that students come to school with no "human capital" influences and that the school would perform the same irrespective of the

students who attend. Also, there are no compositional effects, that is the effect of the enrollment characteristics of the class or school on individual student achievement.

A second model, the conditional accountability model, also attributes student success to what has occurred in a school in the year tested, but it acknowledges that students bring "human capital" characteristics, which may not be able to be measured precisely but have proxies that are fairly close. As in the California similar school index, the use of student inputs allows the state to compare performance among schools with similar characteristics. Education policy makers are concerned that including student background in modeling school performance will change performance expectations for students. The fear is that performance conditioned on background necessarily leads to expectations based on background, but performance standards are (under NCLB) set irrespective of background - i.e. a student must be proficient in mathematics and reading.

A third accountability model, growth models, assume that a school's ability to make academic progress is a better indicator of performance than a static determination of the performance of a student—and school—at a certain time. Growth models also can consider the "potentially negative, spurious relationship between status and growth," said Goldschmidt, "and also account for the effect of status on growth and of student characteristics on growth."

"Value-added models," Goldschmidt explained, "are growth models, but their distinction is that they look at the difference between how a school's students actually perform and how they are expected to perform." The major differences among growth models, he said, relate to the time frame used for analysis, how the expected growth is estimated, whether a series of gains eliminates the need to include student inputs, whether initial status is included as a predictor of growth (the CRESST model) and assumptions about teacher effects.

A true growth model uses panel data to look at gains and growth for the same students over time, but some state models are asking different questions such as: How do schools fare when considering year-to-year changes in performance in the same grade (in sequential cohorts) rather than how schools fare in improving the performance of the same students over time? Both approaches are valid, Goldschmidt said, and the choice depends on a state's objective.

Showing ten years' worth of data, Goldschmidt pointed out that results based on individual longitudinal student data will differ from results based on

longitudinal cohort data. The data demonstrate that there is some relationship between cohort growth and individual growth within schools, but that schools are likely better able to affect subsequent cohorts over time than to alter individual student growth trajectories. Goldschmidt said the panel growth model can be used for more than accountability such as in-depth analyses of sets of schools to determine which schools are doing better and why.

Kilchan Choi described a new CRESST value-added model that uses longitudinal, multiple cohort data. The CRESST Growth Model Based Accountability System incorporates repeated measures of student achievement, and estimated school and mean gain, comparing the school growth estimate to district or state average growth. Using a universal identification for every student, it assesses multiple cohorts of students and can track how much the initial achievement gap has magnified, diminished or not changed over time.

Illustrating the model with data from an 11,000-student Washington State district with 74 schools, Choi presented data on five cohorts, with the first one representing students before NCLB took effect. The data show no difference in the estimates of value-added performance among the cohorts, he said.

Turning to the evaluation of programs, Mike Seltzer of CRESST/UCLA said the "black box" of educational experiments needed to be opened to find out what conditions enable programs to succeed. "Are programs more successful if they are implemented in a certain way, and if so, why?" he asked. Much of the emphasis in multi-site studies of programs in recent years, he said, has been on getting single-number summaries of overall differences in outcomes between those assigned to treatment conditions and those assigned to comparison conditions.

To make studies of the effects of educational programs as useful as possible to an array of stakeholders, including policy makers, teachers, and parents, Seltzer said it is necessary to move beyond single-number summaries to investigations of how differences in level of implementation magnify or dampen the effects of programs, and to test whether certain program components are critical to a program's success. However, he went on to say that the "rub" is that teachers are not randomly assigned to different levels of implementation. The differences with which teachers implement programs can depend on a whole host of factors, including differences in their training and experience, in their pedagogical content knowledge, and in the prior preparation of their students. And so in studying connections between various

aspects of implementation and the effectiveness of programs, we need to be alert to factors that may be confounded with differences in implementation.

Despite our best efforts to anticipate and measure possible confounding variables, teachers who differ in terms of the quality with which they implement various program practices may differ in important ways that have not been measured, giving rise to possible hidden bias. Building on the work of Ken Frank of Michigan State University, Seltzer outlined an approach to assessing the impact of omitted confounding variables in multi-site evaluation settings. In an example, Seltzer noted that work conducted by Saxe, Gearhart, and Seltzer (1999) showed that the degree of alignment of teachers' mathematics instruction with practices called for in reform-mined documents such as the NCTM standards, is strongly related to how well upper elementary students are able to solve challenging problems involving fractions. Despite controlling for an array of factors, one may still wonder whether perhaps there is a possible confounding variable that has not been controlled for. Seltzer showed that the correlations of a confounding variable with alignment and with student problem-solving scores would need to be exceedingly high to impact the results to the point where alignment is no longer a significant predictor of problem-solving outcomes. That is, it would take an extremely strong confound to alter our inferences regarding the effects of alignment with reform-minded practices on students' problem-solving outcomes. Seltzer concluded by noting that the ability to assess the sensitivity of results concerning practices of interest to unmeasured possible confounds can help us gauge the sturdiness of our results.

EVALUATION OF INNOVATIVE PROGRAMS

LA's Best Program

In the first of three presentations on evaluating innovative programs, Denise Huang of CRESST/UCLA described the evolution of evaluation of the Los Angeles' BEST program over 16 years. Initially, the evaluation focused on the conceptualization and design of interventions, then moved to evaluating implementation, and finally to assessing program effectiveness and efficiency. At the same time, the LA's BEST program was growing from 10 pilot sites to 133 sites serving more than 23,000 students.

Currently, CRESST is in the third stage of evaluation, using experimental models of comparison with both cognitive and non-cognitive measures, she said.

There have been some challenges including "the difficulty of collecting consent forms, locating comparison groups, working with a self-selected sample, dealing with a high transient group and working with incomplete archived student data," she said.

In general, the findings show no difference in academic performance between LA's BEST and non-LA's BEST participants, but those in the program either maintained or improved their standardized test scores and improved their attendance rates. Their social skills also improved, Huang said, including improved conflict resolution skills, better study habits, and self-reported increased aspirations. Various sub-groups also benefit more from the program, she said, including females, those who attended regularly, those who initially scored low on attitudinal surveys, English-language learners and students with low test scores.

These positive outcomes also held up on a longitudinal comparison study of LA's BEST participants and non-participants whom are now in high schools. The study cohort was the 1993-94 student group. LA's BEST participants reported more lasting friendships, knowing more about the importance of school, receiving opportunities for mentoring, and better school attendance. Another finding was that LA's BEST parents had a higher degree of involvement and higher expectations for their children. The next steps for the evaluation project, according to Huang, is to participate in a Department of Justice Long-term Effect Study and in a national partnership that is developing a web-based resource for after-school programs across the country.

The Shoreline Science Project

The Shoreline Science project conducted by CRESST researchers focused on the effects of the literacy and science integrated units on elementary students' interest, motivation, and learning. Both quantitative and qualitative analyses were used in the project, according to Jia Wang of CRESST/UCLA. Quantitative analysis examined student data on literacy and science pre- and post-tests. Teacher interviews provided the qualitative data. The evaluation included about 350 students in the science component, with 16 treatment teachers and eight control teachers; and, 237 students in the literacy component, with 12 treatment teachers and 3 control teachers.

Students in both groups improved in the post-test, but the treatment group students' gain score was higher than the control group students. The differences

were statistically significant for one of the two science measures, and for all four literacy measures that were expected. Teacher interviews found teachers appreciate the content of the guides and of the assessments. Although all students benefited, teachers said the program was particularly successful with second and third grade students who were at or above grade level. They said they would use the program again because of heightened student engagement, the mix of activities, the logical sequencing of lessons, the use of group and paired activities, lessons that encouraged conceptual understanding, and the integration of science and literacy. Teachers recommended that materials be developed for students below grade level, priorities be set on coverage of the material and the materials be differentiated for learning styles.

CRESST researchers recently completed a three-year evaluation of the Artful Learning Program, which uses art as the entr•e to all curriculum areas. Teachers design each unit around a central theme, or masterwork, using four phases: experience, inquiry, creation and reflection. Based on the work of the late composer Leonard Bernstein, the program is sponsored by the Grammy Foundation and is now being implemented in nine states and 23 school sites.

The evaluations consisted of a teacher implementation survey, a pre- and post-test survey and teacher and administrator interviews, according to Noelle Griffin of CRESST/UCLA. The findings included: improved teacher attitude and teacher practice and more teacher collaboration, which led to research on collective efficacy; increased student motivation; and improved retention of material by students. Limitations of the program, Griffin said, included inappropriate testing (it was focused on reading and math, but the arts program stressed interdisciplinary learning) and a lack of suitable assessment instruments. Also, the evaluation did not cover teacher transformation.

Griffin said the challenge to the Artful Learning Program in the future is the development and implementation of portfolio assessments. The rubrics need to be usable, she said, and there should be a more organized plan for selecting student assignments and greater attention to the interscorer reliability. Also, the project needs to provide rater training and ongoing technical support for sustaining the portfolio assessments, she said.

Issues in Teachers' and Schools' Assessment Practices

The research on classroom practice presented at the conference included two projects focused on science instruction. One was an evaluation of science teachers' uses of assessment portfolios that grew out of the professional development program sponsored by the Center for the Assessment and Evaluation of Student Learning (CAESL). CAESL teachers collaborated in cross-district, grade-level teams to develop and implement assessments for inquiry science curriculum units. In their portfolios, teachers documented assessment plans and assessment implementation, especially interpretation of student work as a guide for instructional improvement.

To examine teacher growth, CAESL researchers Shaunna Clark of UCLA and Maryl Gearhart of UC Berkeley analyzed the portfolios of 10 teachers who completed a series of two or three portfolios, and found some general patterns. Over time, teachers shifted from using general to more specific assessment criteria; made a greater effort to capture student understanding instead of focusing on a more superficial assessment of student work; and shifted from a binary (right/wrong) grading system to one that allowed for varying levels of student understanding.

In surveys and focus groups, teachers reported that the portfolios were a useful resource for learning more about classroom assessment. They appreciated the opportunity to learn to clarify their learning goals and develop appropriate assessments, though some teachers felt they did not have time to create good assessments and would have preferred to refine existing ones. While teachers felt they learned a great deal in institutes, teachers would have appreciated onsite collaboration and coaching in their schools or districts.

Formative assessment is central to good instruction in several ways, including focusing learning activities on key goals; providing students feedback so they can rework their ideas and deepen their understanding; helping students develop metacognitive skills to critique their own learning products and processes; and providing teachers with systematic information about student learning to guide future instruction and improve achievement. Yet teachers typically have little preparation in assessment or how to use such information well. Pam Aschabacher presented the challenges and results of a professional development (PD) intervention to help elementary teachers use science notebooks for formative assessment purposes.

Using a quasi-experimental design the Cal Tech research team studied the effects of professional development on teachers' formative assessment beliefs and practices, their general science teaching, and student achievement in Grades 4 and 5 science. Data on implementation and outcomes came from 12 novice and experienced teachers and their classes, before and after the professional development. Data sources on teacher knowledge, practice, and beliefs included class observations, surveys, interviews and focus groups, a content knowledge test, and analysis of class sets of notebooks. Data sources on student achievement included analysis of notebook entries, pre/post tests of unit concepts, and embedded performance assessments.

"Most teachers began to see science notebooks as formative assessment tools after the professional development," said Aschbacher, "and made attempts to use them as such." However, many teachers still struggled with knowing how to evaluate student work, making time to give feedback, and thinking about how to revise practice. "Unfortunately, few strong incentives for long-term change and low priorities for elementary science in the schools," added Aschbacher, limit teachers' development of deep content knowledge and use of formative assessment, despite positive results for students.

A third research study looked at the use of data in an urban school district that was using a value-added assessment system. The premise of many data-driven reforms is that schools with access to more data will improve, although there is only limited evidence to confirm this. CRESST researcher Kyo Yamashiro of the Long Beach Unified School District, with Joan Herman and Kilchan Choi of CRESST/UCLA, analyzed data transformation plans from the Long Beach schools, choosing 13 schools with higher than average low SES populations. Case study site visits at four schools, including interviews, focus groups and teacher surveys were included in the research.

"Over time," said Yamashiro, "some schools developed data review groups and all schools increased their use of assessment results." The researchers found that the use of data for instructional decisions depended in large part on the leadership at the school and the school climate, and that many schools felt that the amount of data was overwhelming. The school data review and transformation planning process gradually became more centralized in some schools. The schools generally felt that the state testing data was not useful, valid, timely or interpretable. District data were seen as more detailed and useful. In general, Yamashiro said, data use that is

comprehensive, integrated with instructional decision making, collaborative, and periodic appears to be correlated with high student performance, but more research is needed to investigate this link.

EXPERT PANEL DISCUSSIONS

A new feature of the CRESST national conference in 2005 was an opportunity for open discussion of issues with expert panels. Instead of long presentations, the experts summarized the research and challenges, then led participant discussions.

Robert Linn of CRESST, University of Colorado at Boulder and Daniel Koretz of CRESST, Harvard University guided a conversation on accountability policy and the effects of NCLB, in particular. The way accountability is playing out in schools results in unintended consequences, the group emphasized. "Demanding a rate of student improvement that is impractical increases the rate of teachers taking short cuts," Koretz said. These include focusing instruction on content areas that are being tested and ignoring all else and adjusting instruction to focus on details of individual test items or scoring rubrics when they are available. As a result, said Bob Parker of the Clark County School district in Las Vegas, Nev., "Teachers do not tend to look at big ideas, or a framework, or fundamental knowledge in the discipline."

The group identified two major problems with accountability under NCLB: the requirement for 100% proficiency and the desegregation of sub-groups.

The consensus was that the proficiency goal is not achievable. One problem is that the definition of proficiency varies from state to state so the rhetorical goal of 100% proficiency or above lacks credibility. In low-performing schools targeted by the sanctions under NCLB, teachers face a quandary, said Bella Rosenberg of the American Federation of Teachers. The majority of their students, she pointed out, "started so far behind the starting points of No Child Left Behind that you cannot make that kind of progress in any given year. So, teachers are being put in an impossible situation. Nothing fundamental is going to change so long as the incentives are lined up to basically commit immoral acts," she added.

The group also discussed several issues raised by disaggregated reporting of scores. They concluded that disaggregated data ought to be reported every 3-4 years instead of annually, as required by NCLB. Statistically, it is more reliable to look at the distribution of performance for each sub-group over multiple years. In time, the group should move up. In addition, because the size of the achievement gap is

dependent on the political decision of where the cut score is, it would be better to look at the entire distribution rather than at the percentage proficient to see if the gaps are narrowing. Finally, the group noted that another problem with the sub-group desegregation is that students who are in more than one group may be counted more than once, increasing the probability that the school will not make adequate yearly progress.

As to the substitute of value-added systems for the status system under NCLB, it would be an improvement but has its own drawbacks, Linn said. Tracking students over multiple years is difficult because of student mobility. Also, the content of what is tested matters more in value-added systems than in other systems, he said. If 7th grade honors students are studying algebra and 7th grade under-performing students are studying multiplication, group comparisons of value added present problems because the content differs. Value- added models are better for comparing students who start at the same level and where the content and instruction are comparable.

Policymakers in the United States need to see models for better accountability systems, suggested a participant from Australia. Currently, information gathering for system use in the U.S. is separated from information gathering for teachers. One system would be better, and should include professional development for teachers that helps them integrate instruction and assessment.

Formative Assessment

Experts at this discussion framed the conversation by presenting what they believed were the three most critical issues in formative assessment.

Determining how to connect "our" ideas to the materials of the teachers was the first issue discussed by Eva Baker of CRESST/UCLA. This is difficult, she said, because there is no unified curriculum or instructional program in American schools. Her second issue was the concern about the quality of intervention following formative measures. "There are a huge number of options available for teachers (resources), but mismatches between the materials and desired intervention may still exist," she said. Teachers have access to materials that lack validity for instructional purposes, which they get from the web or other sources. Another "connection" issue is that what works in one school may not work in the others. In Baker's view these issues have sufficient resources for solutions, but "we must create

systems that refresh people's confidence. Schools that don't do well may be struggling because they have no collective efficacy," she said.

Ensuring assessment literacy was the first critical issue cited by Margaret Heritage of CRESST/UCLA. More attention needs to be given to assessment quality, teachers' understanding of the purpose of assessments, and to alignment issues, she said. Secondly, teachers often lack sufficient depth of content knowledge to effectively assess students, she said. "In such a situation, should we expect teachers to know what misconceptions and skills students have or don't have?" she asked. Her third issue was professional culture in schools. Teachers and administrators, she said, need to create a culture to use the data provided by assessments, especially combining data from both formative and large-scale assessments and producing an action plan for learning.

The need to bring concerns of diversity into assessments was the first critical issue mentioned by Jerome Shaw of the University of California, Santa Cruz. He also cited data use as an issue including access to information, disposition to use the data and whether teachers have the skills to use data. "If educators have access to data and the disposition to use it, do they have the knowledge and skills to make use of it?" he asked. His third issue was the need for both federal and state policies that promote gathering and use of assessment data that are not based exclusively on standardized tests.

The participants noted that teachers often prefer checklists to "big ideas," and because of external pressure, formative assessments start to look like the external measures. Some discussion focused on the need in this country to revisit the use of portfolios created by groups of teachers, or to model efforts in Japan to collect and synthesize knowledge about students. The participants concluded that there was a need for a framework to systematically help students, which would include both informal assessment strategies as well as formal formative assessments. Teachers need examples or models of how this system would work.

Measuring Progress

The technical aspects of various accountability models were discussed by the measuring progress expert panel, based on research by Pete Goldschmidt, Kilchan Choi and Kyo Yamashiro. Goldschmidt and Choi drew substantially on a chapter of their work with Yamashiro which appears in the book, Uses and Misuses of Data for Educational Accountability.

"Controlling for student differences can be done by the use of initial status (a student's starting point of achievement) or by background characteristics," Goldschmidt said, and the choice basically is a policy issue. Choi's work examined the role of adjusting student and school characteristics in measuring progress. He used both status and background in research to predict later performance and found that adjusting for socio-economic status (SES) is not necessary if information on initial student status is available. School-level predictors generally consist of school mean initial status and mean SES (percent of students in the school lunch program). Contrary to the findings at the student level, status was unimportant in adjusting for school differences, but SES played an important role. Special education students and English-language learners were not included as predictors in this study in order to avoid shrinkage to the district mean and multi-co-linearity in the hierarchical linear model.

The group discussed the issue of choosing metrics for measuring progress, with Goldschmidt urging caution about using rank correlations. Another study found that normal curve equivalents and scale scores yield highly correlated results based on mean initial status ad growth estimates in hierarchical growth modeling.

These presentations are available on the CRESST website at www.cresst.org.

The next CRESST conference is scheduled for January 2007.