1998 CRESST CONFERENCE PROCEEDINGS
Comprehensive Systems for Educational Accounting and Improvement: R&D Results

CSE Technical Report 504

Anne Lewis

June 1999

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

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education, and under a subcontract to ISX (letter subcontract 95-D-8644-01) funded by Defense Advanced Research Projects Agency and administered by the Navy Research and Development Center.

The findings and opinions expressed in this report do not reflect the position or policies of ISX, Defense Advanced Research Projects Agency, the Department of the Navy, the Department of Defense, the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

1998 CRESST CONFERENCE PROCEEDINGS

COMPREHENSIVE SYSTEMS FOR EDUCATIONAL ACCOUNTING AND IMPROVEMENT: R&D RESULTS

Anne Lewis

Fifty sessions and nearly 70 presenters or discussion leaders thoroughly examined all aspects of comprehensive systems for accountability and improvement of education at the 1998 CRESST national conference, and it was distilled in a final comment. As chair of the last panel, Robert Glaser of CRESST/University of Pittsburgh noted that the panel's title - reinventing assessment and accountability to help all children learn - "is the real topic of the whole conference."

That topic was all-inclusive, indeed. It ranged from national assessment studies, to innovative uses of technology, to deeper understanding of the assessment of students traditionally left out of accountability, to benefits from partnerships between researchers and practitioners.

The CRESST national conference is a checkpoint in the continuing search for better assessments and school improvement strategies. The 1998 gathering began with a national context for work on assessment and school improvement, moved to specific issues, and devoted time to new and/or continuing research, often resulting from working relationships between CRESST and district and teacher partners.

The National Overview

The recent appointment of a new assistant secretary of the federal Office of Educational Research and Improvement, Kent McGuire, was the impetus for an opening session on "new thinking" about research issues at OERI. Combining McGuire's outline of the issues with his own commentary, Joseph Conaty, Director of the National Institute on Student Achievement, Curriculum, and Assessment at OERI, discussed what his agency sees as the next challenges for it from school reform initiatives. These include the following.

- **The need to understand the importance of different uses for testing systems.** Tests for instructional improvement and tests for accountability are not and should not be the same.

- **The necessity for tests to be relevant.** "What are the big things children should learn and how can we adequately measure them?," Conaty asked. Moreover, research should show how these "big ideas" are implicitly developmental, such as the algebraic principles students should know at certain stages of learning.

- **The growing interest in program effectiveness and evaluation.** The new Reading Effectiveness Act and the Comprehensive School Reform Act are basically about research on effectiveness of programs, Conaty said. They reflect a trend toward a much closer relationship between research and program development. "People want to know how effective these programs are and how effective they are compared to other interventions they could adopt," he explained.

- **Legal issues related to assessment.** Testing theory and methodology generally were developed with groups in mind; that is, they deal with inferences about central tendencies in the distribution of scores within groups. However, where assessment meets the law is on the issue of what the test does to individuals. Court cases challenging assessment policies are often about individual children, so policymakers need to know the effects of assessment on individual students.

- **Attention to the public and private interests in the field of testing and assessment.** The large public interest in testing creates a demand for better partnerships, such as one between a center like CRESST and publishing companies that need to market a product. Proliferation and marketing of many tests "leads to lack of comparability and some loss of the public interest," Conaty said.

- **Issues around who is "inside" and who is "outside" the testing field.** Attention is needed to more than who is included in testing programs and what accommodations are used. Research is needed on who is included in the conversations around testing such as the design of test items and decisions about accommodations. This area of research needs to "think of inclusion writ large."

- **The relationship between education reform and assessment.** What is needed is assessment that supports reforms. Research needs to provide more information on how assessment functions in a whole system and not focus only on assessment results, said Conaty.

Program effectiveness, Conaty added, is a political and a research issue. It is one that will play into reauthorization of major legislation by Congress. He cited the current U.S. Department of Education's study of charter schools that found 30 different forms of tests being used in the sites chosen for study. With so many different tests, how do we capture the effect of charter schools compared to other interventions? The models mentioned in the Comprehensive School Reform Act (Obey-Porter) “were selected purely for political reasons,” he said. What researchers need to do is develop some guideposts or principles for educators to use when selecting models for adoption / adaptation, leading them to ask good questions about the programs’ claims and suitability. Further, researchers need to not only identify successful programs but also to produce information on what pieces of programs are most effective.

Accountability and Assessment Systems

Citing Conaty’s discussion of the distinction between testing for instructional improvement and testing for accountability, CRESST Co-Director Robert Linn said a balance between the two is lacking in current assessment policies. “As is often the case in this country,” he said, “education policy has swung radically instead of making minor adjustments.” The swing in assessment policy, said Linn, is toward greater emphases on accountability, holding schools responsible, and providing rewards and sanctions for student performance. This is a move away from using assessment as an impartial monitor of progress and a means of diagnosing system strengths and weaknesses.

Linn discussed elements necessary to create a more coherent conceptual framework for accountability systems. One is a database for measuring progress (emphasized in Title I), and he mentioned as an example the value-added data collection in Tennessee. However, because it requires annual testing of every student, the value-added model creates pressure “to become less ambitious on what is measured.” Methods for establishing meaningful performance standards present another important issue on which there is insufficient guidance from research. Some of the anomalies that occur in setting performance standards, he said, include substantial differences in difficulty from one subject to another and from one grade level to another. Accountability that depends on assessments also must take into consideration:

- student mobility, both from school-to-school and from one district to another;

- the treatment of students who are not tested in an accountability system; and

- the disaggregation of assessment data.

Research Results on National Test Issues

Linn's cautions about using assessment data for major accountability decisions served as an introduction to the proposal to develop voluntary national tests. The 1997 CRESST conference spent considerable time on the pros and cons of President Clinton's national testing idea, presenting more opinion than research-based substance. A year later, the research community had produced a research base for discussion, centered on three studies Congress requested from the National Academy of Sciences through its National Research Council. Many assessment researchers connected to CRESST participated in the studies. The framework of the studies went beyond technical issues, and the research reports presented to Congress made a statement about the proper role of testing.

National testing was mired in political contests between Congress and the White House, noted Michael Feuer of the National Research Council, but it also opened up discussion on fundamental tensions in education, such as belief in local control versus new national efforts to influence education.

The final reports had been delivered on time - with only nine months of study - just days before the CRESST conference. That was possible, Feuer acknowledged, because of the "intellectual firepower we were able to muster that was truly astonishing. We could not have had such rigor in the studies without the caliber of people like Bob Linn and others working with us."

Researchers who worked on the three studies presented findings.

Voluntary National Test Evaluation, Phase I (presented by Lauress Wise). The Congressional legislation required the study to report on the quality of the test development but at the same time did not allow any field-testing. Not having any data created a challenge, Wise said. Groups of students were brought in to try out the tests, and the study also included sensitivity reviews and workshops with expert panelists. The study addressed questions set out in the legislation.

- **Test specifications.** Actions to align the tests with the NAEP frameworks have succeeded in building on NAEP efforts to achieve broad consensus, maintain independence from specific curricula and instruction, and maximize the possibility for linking the test scores to NAEP. In other words, attempts have been made to avoid reinventing NAEP. More detail is needed on the accuracy targets for resulting scores and links to NAEP achievement-level descriptions. Efforts should be made to promote acceptance of the tests in order to ensure high participation rates.

- **Item development.** A sufficient number of items have been developed in a remarkably short time (2,000 in reading and math), but additional items are likely to be required in specific content areas including short literary reading passages, algebra and functions, and geometry and spatial sense. The bias review was thorough, balanced, and should be repeated as additional items are written.

- **Field-test plans.** Generally appropriate; the sample size is consistent with common practice although there is possible concern with representation of small schools. How tests will be scored should be decided on before deciding on the final design. Moreover, the rationale for some elements is not clear; e.g., there are larger samples for some forms than for others, hybrid forms model, and an equating cluster design. Targets for equating/linking accuracy should be added.

- **Pilot test plans.** Generally appropriate; the number of items piloted appears adequate to support the creation of six field test forms, and the overall and sub-group sample sizes appear reasonable for intended analyses. However, more specifics are needed on how the data will be used to screen items, and the inclusion of hybrid forms is not well supported and may create problems.

- **Inclusion and accommodation.** The National Assessment Governing Board has approved a policy for comprehensive inclusion with appropriate accommodations. However, opportunities to advance the state of the art may have been missed because of the compressed schedule. The study found no evidence that accommodation options were being considered during the item development and there was no attempt to exploit technologies, such as computer-based testing, to achieve valid measurement with accommodations. Plans for testing the validity of scores with various accommodations should be included in the pilot and field tests.

- **Reporting plans.** NAGB clearly intends to report achievement using the NAEP achievement levels. However, several other reporting issues need to be resolved. How will items be related to the achievement level descriptions used in reporting? How will measurement errors be communicated? How will scores be reported for schools, districts, and states? What supplemental information will be reported, particularly for students who score below basic?

  It is essential that a clear vision of how results will be reported drives, not follow, other development activities.

Feasibility of an Equivalency Scale that Would Allow Test Scores from Commercially Available Standardized Tests and State assessments to be Compared to Each Other and to NAEP (Paul Holland, chair; report presented by Feuer). This study was necessary because of several tensions in educational policy:

local control of education versus national standards, diversity and innovation versus uniformity and coherence, and the debate over the feasibility of the voluntary national tests; that is, can they be valid and are they practical? Also at issue were the kinds of burden that such an equivalency scale would create and what would need to be done to create it.

The interim report focused on a definition of the problem and concluded that: *comparing the full array of currently administered commercial and state achievement tests to one another through the development of a single equivalency or linking scale is not feasible; and reporting individual student scores from the full array of state and commercial achievement tests on the NAEP scale and transforming individual scores on these various tests and assessments into the NAEP achievement levels are not feasible.* The final report discussed the following.

- Factors that affect links include test content; test format; measurement error; and diversity in tests, uses, and consequences (different test instruments, inclusion rules, administrative practices, content emphases, and reasons for testing).

- The challenges of linkage include the fact that choices made at each stage of test development affect validity, and linking magnifies challenges to validity because different tests reflect different choices made at each stage of the test development process.

- Linking to NAEP is challenging because of its distinctive characteristics that include content coverage, item format distribution, test administration conditions, use of results, achievement levels, and individual level scoring.

Thus, the conclusions remain the same: *neither linking to currently available tests nor to NAEP is feasible.*

The researchers, however, also “stepped back and asked what if not a full array of tests were linked, but only some?” said Feuer. Under limited conditions it may be possible to calculate a linkage between two tests, but multiple factors affect the validity of inferences drawn from the linked scores. These include content, format, and margins of error of the tests; the intended and actual uses of the tests; and consequences attached to the results of the tests. When tests differ on any of these factors, some limited interpretations of the linked results may be defensible, while others would not.

The researchers were more cautious on NAEP. Links between most existing tests and NAEP, for the purpose of reporting individual students’ scores on the

NAEP scale and in terms of the NAEP achievement levels, will be problematic. Unless the test to be linked to NAEP is very similar to NAEP in content, format, and uses, the resulting linkage is likely to be unstable and potentially misleading. Future research should focus on criteria for evaluating the quality of linkages, the level of precision needed to make valid inferences about linked tests, and reporting of linked assessment information to parents in a useful way.

High Stakes: Testing for Tracking, Promotion, and Graduation (Robert Hauser, chair; Robert Linn presented). The third Congressional charge to the NRC was to recommend appropriate methods, practices, and safeguards to ensure that existing and new tests are not used in a discriminatory manner or inappropriately for student promotion, tracking, or graduation. Additionally, existing and new tests should adequately assess student reading and mathematics comprehension in the form most likely to yield accurate information of student performance.

The research committee went slightly beyond the charge, Linn said, to consider “a professionally good set of uses for tests.” It explored several questions about high-stakes testing. What constitutes appropriate, nondiscriminatory use of existing and new tests - not just the voluntary national tests - in decisions about student tracking, promotion, and graduation? How can the participation of students with disabilities and English-language learners in large-scale assessments be maximized while ensuring the comparability of test results for all students? How can we ensure that test makers and test users will abide by norms of appropriate, nondiscriminatory test use?

There are several goals in testing for tracking, promotion, or graduation including: setting high standards, raising student achievement, ensuring equal educational opportunity, fostering parental involvement, and increasing public support for the schools. However, testing may have negative consequences for individuals, so policymakers should be sensitive to the balance between individual and collective benefits and costs.

Linn discussed a framework for high-stakes testing established by Sam Messick. The framework includes measurement validity (is a test valid for a particular purpose? Does it measure knowledge in the intended content area?); attribution of cause (does performance reflect knowledge and skills based on proper instruction or does it reflect poor instruction? Or does it reflect irrelevant factors such as language barriers or unrelated disabilities); and the effectiveness of treatment (are the consequences more beneficial educationally than other available treatments?).

The committee set out four principles of test use.

- Tests have validity only in relation to specific purposes.

- Tests are not perfect, and neither are the alternatives.

- No high-stakes educational decision about a test taker should be made solely or automatically on the basis of a single test score; other relevant information should also be taken into account.

- Neither test scores nor any other kind of information can justify educational decisions that are not beneficial for students.

Linn presented seven selected findings and recommendations from the committee report.

1. Accountability for educational outcomes should be shared among states, school districts, public officials, educators, parents, and students; not by students alone.

2. Tests should be used for high-stakes decisions about individual mastery only after students have been taught the knowledge and skills on which they will be tested.

3. Consequences of high-stakes testing are often "either-or," but this need not be the case; tests and other information can lead to early diagnosis and effective intervention when students have learning problems.

4. Some educational practices are typically bad for students, including placement in low-track classes and retention in grade; tests should not be used for such purposes.

5. All students are entitled to sufficient test preparation; e.g., familiarity with item format and appropriate test-taking strategies). At the other extreme, educators should avoid narrowly teaching to the test.

6. High-stakes testing programs should include a well-designed evaluation component, and the consequences of high-stakes assessments should be measured for all students and major sub-groups of students.

7. The Voluntary National Tests should not be used to make decisions about tracking, promotion, or graduation of individual students.

The report also addressed students with disabilities and English-language learners, seeking to ensure their increased participation in large-scale testing programs so that districts will improve accountability of those students, Linn said. The committee called for greater inclusion and use of accommodations without jeopardizing the validity of the scores.

The three studies were notable because “they said in simple language that maybe the voluntary national tests are not such a good idea and need to be reconsidered,” noted Penney Sanders, head of TORST, Inc., of Sulphur, Kentucky, and a designer of the former Kentucky assessment system. The studies’ conclusions were in line with the attitudes “outside the Beltway” where parents “essentially just want to know how their child is doing in school. Can my child read, do math? Is my child getting the tools to be successful?”

The dilemma is that very complicated issues must be explained to constituents, who only want simple answers, said Sanders. She said that measurement experts and policymakers must be able to communicate concerns about the appropriate use of tests but at the same time not be seen as making excuses.

Asked about the reaction of Congress to the studies, Feuer said that some members were disappointed because they did not find a silver bullet to support the test. While there were no firm conclusions one way or the other on linking to NAEP, he said the committee struggled with the general idea, and the achievement levels in particular, as well as the uses of the test. Wise added that committee members were concerned about a district adopting the 8th-grade test score as a graduation requirement, which would be “a bad idea.” That’s why the high-stakes committee did not avoid the issue of use, Feuer added, emphasizing its recommendation that a voluntary national test not be used for making individual decisions about students.

Assessment of Early Literacy Skills

At previous CRESST conferences, assessment of literacy skills in young children had been seen as primarily a research and technical issue, but the 1998 conference made it obvious that this, too, is in the political arena. Public demand for better student reading skills puts increased pressure on how we assess reading.

Alison Imbens-Bailey of CRESST/UCLA, reported on the development of a literacy development checklist for kindergarten teachers. The checklist includes such items as how well children are telling stories, how clear their narration's are, and their progress. The assessment is being used as a diagnostic profile to determine whether and what interventions are necessary to support student progress. She said her early results reveal that the basic issue may be teacher skills, not necessarily how we assess students.

"From the pilot study, we already are realizing that we are dealing with professional development issues ... and learning as much about teachers as about children," Imbens-Bailey said. The assessment's value will be in helping teachers understand they can work with children with literacy problems rather than refer them to out-of-class help.

Similarly, Samuel Meisels of the University of Michigan/Center for the Improvement of Early Reading Achievement reported on the use of the Work Sampling System in Pittsburgh Title I schools. WSS is a curriculum-embedded continuous progress assessment for preschool through grade 5. It is not used to sort or classify children, he said, but to help teachers document the progress of children's skills. It includes a checklist of academic accomplishments, portfolios of students, and summary reports - all in seven curriculum domains. Three observational periods involve teachers, children, and families.

Meisels said the WSS has proven to be a valid means of evaluating achievement and progress among young children. The correlation between WSS ratings and standardized test scores (Woodcock/Johnson assessment) is moderate to high, and by 3rd grade the students in classrooms using the WSS were generally at national expectations on achievement. In addition, teachers and families are satisfied and become more so the more experience they have with WSS, he said.

"There is no single way to demonstrate accountability," Meisels concluded. "Accountability does not need to be reported solely in terms of results data, but also can be thought of instructionally and used to improve performance. That way, we end up opening school doors for children instead of closing them."

H. D. Hoover of the University of Iowa confronted the idea that early literacy assessments are not valid. Data from the Iowa Test of Basic Skills (ITBS) and other standardized literacy tests, he said, show an internal consistency in grades 1-3 comparable to average reliability for tests in grades 3-8. Despite the rapid change in the skills of young children, the consistency holds up regardless of when students are tested, spring or fall.

Criticizing the recommendation of a National Education Goals Panel on Early Childhood Education against high-stakes early testing, Hoover said, “as a parent I would be offended at not having external information on my child’s strengths and weaknesses until the 4th grade.” The main point, he said, is that different tests give different kinds of information, and the norms used by external tests are the only way “to enable you to look at strengths and weaknesses.”

In a later discussion, Meisels said that his experience with data from ITBS on young children is not consistent with that presented by Hoover. “If we start to link early childhood programs to the content of standardized group-administered tests, we start to narrow the academic program seriously,” he said.

Politicians and parents need information on the achievement of young children, said Adrienne Bailey of the Council of the Great City Schools. She warned that the fate of Title I might hang on evidence that Title I helps needy young children succeed academically. Representing 49 districts enrolling 5.5 million grades 1-3 students in Title I schools, she noted that almost half (48 percent) of the districts are using a norm-referenced test at grade 3 with 30 percent administering such a test before grade 3 (districts primarily are using SAT-9 and -8, ITBS, MATS, and Terranova).

Studies show that there are low expectations and mediocre strategies used in many urban schools, Bailey said, and it is in the early grades where the diet of “drill and kill and dull routine” is the strongest. Using standards to drive instruction and assessment practices raises a number of critical issues and questions, Bailey said, especially because standards-based reform so far is more rhetorical than real.

A Council of Chief State School Officers study group on early childhood assessment has come up with two crucial questions: What is the appropriate role of academic content standards to early-grade monitoring and assessment?; and, Is it possible to reach a convergence or negotiation of some type between the two camps, one backing developmentally appropriate education and the other backing accountability and assessment?

The National Goals panel statement, for example, is understandable from an educational perspective, Bailey said, but “from a practical perspective, I am afraid that we may be backing ourselves into a corner without adequate rationale or justification." Although Title I regulations do not support large-scale, group-administered tests for young children, it still requires that programs be assessed and evaluated. The law permits flexibility in selecting an assessment approach, including the idea of backtracking and using 4th-grade level for accountability. Still, she pointed out, "as we ready ourselves to go before an already skeptical Congress to ask for a continuation of this $8 billion initiative, what evidence will we have that the program is working for the kids served by Title I in grades 3 and below when there is no requirement for statewide accountability?"

Bailey called for an open dialogue among state and local stakeholders, parents and professionals, policymakers and practitioners on the issues of accountability and assessment in early grades, suggesting that CRESST be the conveyer. "It is high time that we open the discussion and ask ourselves the basic question: who really speaks for these children?"

Much of the discussion on this issue concerned interpretation of the National Goals Panel report. Lorrie Shepard of CRESST/University of Colorado at Boulder and chair of the National Goals Panel on early childhood assessment said its recommendations had been misquoted. The report said that assessment should support instruction, including reporting to parents. It opposed "high-stakes" testing in the context of using results to rank teachers and schools through the media and, thus, "track kids and keep them down." Hoover replied that he was not in as strong a disagreement with the Goals panel as it may seem and that he had not meant to omit other goals of assessment but had taken up the issue that was presented. However, he stated that he had *not* misquoted the report.

Bailey again called for a conversation with a wide range of people on how to merge the ideas of using assessment for instruction and reporting to parents but at the same time use it for larger political accountability. "We have to figure this out," she said.

Concurrent Groups

Language Accommodation

It is admirable that the country is now in an era of inclusion, according to Arnold Goldstein of the National Center for Education Statistics, but the variability in the criteria states are using to include LEP students in NAEP presents serious problems. Moreover, a significant number of LEP students still do not participate in the assessment.

Inclusion needs to be rethought in several ways. Its philosophy is needed in the design of assessments, not as an after-thought to looking at data. Inclusion needs to be considered in item development, in pilot testing, and in reporting of results. There are conversations, for example, over whether to report results of limited-English-proficient students as disaggregated, in the mix, or not reported at all. Each of these issues, Goldstein said, has its own assortment of technical and political questions that must be answered.

The issues are broad: validity and comparability, fairness and equity, credibility of the test, utility and use of results, practicality of what is being done to include certain students, and the nature of accommodations and their cost factor. Policies need to find the right balance in decisions about inclusion, he said. Providing a translation or adaptation of a test, for example, helps increase participation in assessment, but it must be asked if the results are comparable, if they are the same as using an English version of the test. Furthermore, some subject areas, such as reading, are deemed not appropriate for translation, but that negates consideration of the language of instruction - and the debate continues. The bottom line, Goldstein said, is that no single policy fits all of the issues and that inclusion strategies have different effects on different sub-groups.

Speaking specifically on issues of NAEP redesign related to language accommodation, Steven Gorman, National Center on Education Statistics, said that the recommendations covered:

- including LEP students in the development of test items and in pilot testing;
- strongly considering the use of simplified language;
- allowing additional time for LEP students when possible;
- closer analyses of LEP student data compared to non-LEP student data; and
- further analyses by type of accommodation whenever possible.

These issues arose partially from research on sub-samples of test takers with and without special accommodations. Gorman also said that more information is needed on the use of glossaries by students, as well as differences in LEP student performance aided by accommodations by subject matter and grade levels. Each subject matter has unique characteristics, he said.

Addressing at least some of Gorman’s and Goldstein’s discussion was research by Jamal Abedi, CRESST. Using questions from NAEP assessments in mathematics separately administered to student samples in Los Angeles, Abedi examined linguistic modifications to accommodate students with limited English proficiency. He used four accommodation strategies: a) modified (simplified) English language of the test items; b) a glossary explaining potentially unfamiliar or difficult words; c) original English with extra time; and d) glossary with extra time. Students were assigned randomly to the different accommodation forms and to a comparison group within participating classrooms, to control for teacher and school effects. Abedi summarized his findings.

- Both LEP and non-LEP students showed the greatest increases in performance when provided both a glossary and extra time.
- LEP students’ scores were higher on all types of accommodation except Glossary Only.
- Students who were better readers, as measured by reading test scores, achieved higher math scores.

Most accommodations helped both LEP and non-LEP students, said Abedi, however, the only type of accommodation that narrowed the score difference between LEP and non-LEP students was Modified English. Abedi concluded that simplifying the NAEP items is probably the best strategy to accommodate LEP students.

The presentations on inclusion strategies indicate that much is underway by CRESST, AIR, RAND, NCES, NAEP, certain states, and other groups, John Olson of the American Institutes for Research, noted as panel discussant. “There are a lot of pockets of research,” he said, “but what is needed is a coordinated sharing of research ... ways of providing up-to-date data coming from various studies.” He suggested linking web sites among current networks that are conducting research, such as those at CRESST, OBEMLA, and CCSSO.

Building Teacher Capacity for Improved Learning

The experiences of two exemplary Kentucky elementary school teachers set the framework for a discussion of how standards-based reforms can create classroom improvements.

“Lots of things changed in our teaching,” according to Julie Wright of the Dixie Elementary Magnet School. The Kentucky Education Reform Act encouraged multi-aged classrooms, mandated high-stakes assessment with consequences for schools, created performance-based assessments of students, and strengthened site-based decision making that “put teachers and parents into roles that they had not played before.”

Wright praised state education policies and officials for preparing and supporting teachers as they adopted the reforms. Teachers were welcome to participate on curriculum committees that were set up throughout the state; were given professional development credit, stipends, and released time to visit other sites; and had resources tailored to help teachers meet state standards.

Kentucky also has been responsive to teachers’ concerns, she said. When it appeared that the assessment schedule loaded too much onto the 8th grade, for example, the state spread out portfolio assessments to include 7th grade in some subjects.

Accountability is at the school level, “and in a building that works well together, teachers will prepare students for the next grade’s testing,” she said. Schools not meeting their goals (“in decline”) receive additional support from the state through *distinguished educators* assigned to help them align their curriculum and instruction with state standards. In her school, results of the state assessments helped teachers focus on weaknesses in instruction.

A major result of the reforms, largely driven by the state assessment system, has been to greatly improve students’ writing skills, according to Wright’s colleague at Dixie Elementary, Adele Thompson. “Our students are learning to communicate very well,” she said, “especially because of the emphasis on practicing open response tasks in classrooms.”

Both teachers said the process of change was evolutionary, with teachers at first resisting transforming their teaching, then gradually understanding what needed to be taught and what had to go. As schools decided what needed to be taught at each grade level, Thompson said, “we boldly said to teachers that ‘this is your responsibility.’”

The use of 70 percent of professional development funds is decided upon at the school level, the teachers reported. At Dixie Elementary, they said, the monies have been used to create expertise across the school, enhancing the knowledge and skills of everyone. Hilda Borko of the University of Colorado/Boulder and panel discussant added that her case studies of Kentucky reforms showed that schools made very different uses of the professional development money, but they were all tied to meeting the reform objectives.

Ronald Stevens from UCLA’s School of Medicine reported that technology is the tool for changing classroom practice in Interactive Multimedia Examinations, Experiences, and Exercises (IMMEX).

Focusing on teachers as creators of technology-based curriculum, the program began with three guiding principles:

1. The potential of technology will not be recognized until teachers are confident in and committed to using technologies to enhance classroom practice;

2. Educational technologies and software cannot be crafted for a single class, single grade level or single school, but should be part of a broader design that articulates content and process skills across disciplines and levels of education; and

3. The need for in-depth assessment of students, of the curriculum, and of teaching practices requires that evaluation be a central component of all educational software development and implementation efforts.

Starting with 25 teachers, the program has grown to 100 teachers at a time, who spend a month with the researchers, developing Windows-based software. They also learn how to integrate technology with classroom practice, shape students’ metacognitive skills and become educational researchers.

Hardware is not a serious problem. “It can be fixed in days with money,” Stevens said. The same is not true of professional development. To integrate technology into the curriculum “requires extensive, multi-faceted and prolonged professional development,” he said, a fact now realized in a number of studies about technology use in the schools.

IMMEX-trained mentor teachers and even some students conduct background scientific research and construct problem-solving software for classrooms, “learning with technology, not just about technology." With up to 75 variations of a problem on each set, students who use the software get multiple opportunities to work through a problem primarily designed for chemistry and biology classes.

Stevens reported that the lag from when the software is developed to when it is used regularly in classrooms is about a year, "telling me that technology transfer doesn't happen overnight," he said. "When they do use it, teachers have an embedded assessment to rely on."

As students go through the problems and order menu items (data/information), their choices are recorded. Software recreates the sequence of steps students go through as they solve problems, and the researchers can record improvements over time. "For example," Stevens explained, "a student might start with lots of strategies, then refine them gradually. The software also can spot those students who always guess.

Stevens predicted that the program could scale up easily, handling 200 teachers at a time, "but institutionalization of the technology use is more difficult. Scaling out - using distance learning to extend the reach of the program - may be asking too much of the electronic medium. We don't have a feel for what proportion of the training has to be personal and what can be done electronically," he said.

The Placentia-Yorba Linda Unified School District has been able to substantially institutionalize the IMMEX project in its science classes at Esperanza High School, said Marcia Sprang, a high school science teacher at Esperanza High School. Five years ago teachers became concerned that high school students came to them with compartmentalized knowledge and skills, prepared, for example, to do algebra but not chemistry and unable to write across the curriculum.

Linking with the IMMEX program, the teachers transformed the curriculum in one year. They developed a workshop at the high school site, set up a demonstration in the library for all teachers to look at, including at the elementary level. In the 1997-98 school year, half of the science staff implemented the IMMEX program, using it also for assessment in AP chemistry. "The students were very focused," Sprang said, "and learned things I had not covered in class because they made connections."

Professional development at the school included a summer workshop of 15 teachers and 30 students who wrote a total of 20 programs. "Students were teaching teachers about the technology, and the teachers were teaching them about content," Sprang said. As a follow-up, all new IMMEX programs are being worked into the curriculum, the school is expanding its technology capacity, and administrators in the system are preparing to keep performance records on students from kindergarten up.

The Kentucky and IMMEX presentations presented school-level changes in two very different kinds of projects that stem from very different initiatives, Borko pointed out. In both of them, however, “I hear a commitment of teachers to new practices, an open-mindedness of teachers and willingness of them to consider change, a notion of confidence, a sense of leadership on instruction in the school buildings, and the central role of assessment.” They both also depended on professional development at the school site that could lead to home-grown expertise.

Assessing School Reform

Developing an indicator system to evaluate a comprehensive reform plan was explored by four panelists during the conference. Funded by $53 million from the Annenberg Challenge plus matching funds, the Los Angeles Annenberg Metropolitan Program focuses on deepening existing reform efforts rather than starting new ones, according to Maria Casillas, LAAMP president.

The history of most school reform in the districts was to plan over and over again, implement some of the plans, then never know what to evaluate because of the uneven implementation, Casillas said. LAAMP created a feeder-pattern of school “families” held together by a series of action principles that included: research-based changes, parent and community involvement, rigorous curriculum, time for professional development, and better use of data. An annual review process focuses on the impact of professional development and the changes made by the district in policies and practices.

Self-evaluation and public reporting moved up in priorities, Casillas said, “to better inform the ongoing process in more meaningful ways.” She predicted the efforts would pay off because “the school families are getting smarter, they will have evidence to report out, and a capacity to continue.” A new strategy uses a team of “critical friends” who receive special training to help teachers examine their assumptions about standards and student work, then come together with teachers several times during the school year to share what they are learning.

Discussing the indicator system that supports LAAMP, Barry Gribbons of CRESST/UCLA, explained that the long-range goal is for LAAMP to fade away, leaving the 15 districts involved with local ownership of their reforms and able to use indicator data for continual improvement. Teacher and administrator surveys, databases, and supplemental studies are the core of the evaluation approach.

The LAAMP action principles guiding the development of the indicator system are to:

- build stable learning communities with seamless education throughout the feeder patterns;
- decentralize control over resources and decisions;
- build a challenging and equitable curriculum especially for English-language learners;
- align professional development to school family priorities;
- build parent and community involvement;
- reallocate professional time to allow for more collaborative planning in order to promote articulation across schools, grades, and content areas; and,
- use data for accountability and an annual process of evaluation and improvement.

The indicator system project has had to deal with a great disparity among the 15 districts in their ability to maintain records, especially with those districts that do not collect records electronically, Gribbons said.

The indicator system is sensitive to several factors in the project, including the diverse efforts of 247 schools and the need to:

- be rigorous and creditable;
- look at school, family, and district levels;
- take a longitudinal picture showing progress of students as they move through the system; and
- be consistent with other indicator systems that are being used at state and national levels.

Some considerations of constructing the indicator system are technical in nature, Gribbons said assuring the integrity of the individual measures. At the same time, “we felt we needed to focus not only on discrete pieces but also on the big picture, trying to get an idea if the indicator system as a whole was capturing what we intended for it to do,” he said. Part of this check has included creating profiles for each school family and asking them to review it. The project also relies on intensive studies such as classroom observations and in-depth interviews with teachers.

The indicator categories include student achievement from the state mandated Stanford 9 test, disaggregated by language, special education and economic levels. Language arts and mathematics data are the focus. Data from performance assessments supplement the state test information.

Other indicators include SAT scores, four-year graduation rate, percentage of students passing AP exams, one-year dropout rates, redesignated limited-English-proficient students, student engagement (absenteeism, tardies, and homework completion); and social competence (school violence, teacher reports on behavior).

Still other indicators examine how challenging the curriculum is in the families, including course-taking patterns, progress toward course completion, access to AP courses, number of emergency credentialed teachers and which students they are assigned to; a “checklist” on the appropriateness of curriculum materials; and the degree of articulation.

Survey data, Gribbons said, is helping the project understand if the curriculum is articulated, “if teachers know what other teachers are teaching their kids.”

The family, or cluster, organization being used by LAAMP is unique to Los Angeles and grew out of concern about the instability of both the adult and student population. A family organization could create a more stable learning community, according to Priscilla Wohlstetter of the University of Southern California.

Some LAAMP families are more effective than others. Wohlstetter reported on a part of the project that is studying strategies that school families successfully put in place to implement the action principles and to create a sense of family. One strand of research is looking at what allows for conversation about teaching and learning to take place across the school family; another is looking at the processes that families take to facilitate the design and implement innovation.

"If we expect reform to go beyond single schools, we need structures and processes in place that allow for communication about teaching and learning that cuts across individual schools," Wolhstetter said.

The research methods for this inquiry are focusing on four sites in depth, two in the Los Angeles Unified District, one in Long Beach and one in Pasadena. Interviewing focuses on the members of work teams designing and implementing the reforms. The study also includes reviewing documents such as learning plans and self-evaluations.

"The first thing that has struck us," Wohlstetter reported, "is that the internal organization within families is very similar even though it was designed by each one individually." They have organized their work through teams and committees with every family having a management team. As the families developed greater trust level, they have tended to shrink the management teams because it is difficult to do hard work in large groups, she said. Also, every school family has a set of improvement teams, some relating directly to the action principles while others focus more specifically on learning plans; e.g., literacy teams, technology teams. A final group of teams focus on integrating the work, such as a team composed of principals from every school.

Wohlstetter hypothesizes that what makes a team successful is drawn from research on teaming and school-based management. These include control over budgets, good sources of information, attention to professional development, process skills, and accountability.

Lindsay Clare from CRESST/UCLA reported on another piece of the LAAMP evaluation: how it is affecting classroom practices and student learning?

Twelve schools are participating in this study, one school at each level in four families. "We looked at school family learning plans," said Clare, "interviewed teachers as to how the plans impacted their practice, conducted classroom observations, and collected assignments and student work in language arts at grades 3 and 7." Initially, Clare said, when they asked about the effect of LAAMP, teachers were unfamiliar with the term. Consequently, the interviewers had to ask about specific initiatives. They found a coordinated approach to the LAAMP reform initiatives at the elementary level but much less coordination at the secondary level.

The key factors that seem to impede successful implementation of the reforms include:

- lack of coordination on reform strategies;

- need for much greater focus on literacy at the secondary level;

- greater difficulty at the secondary level to coordinate assessment practices and principles;

- tremendous number of emergency teachers who took up time of veteran teachers; and

- lack of specificity of the standards with teachers saying they needed help on how to link standards to practice.

“One thing that seems to support change at this early stage of study,” Clare said, “is to have a principal who is an instructional leader and who presses on standards early.”

Another study examines one of the action principles - parent involvement - in three school families. CRESST researcher Denise Quigley is focusing on three aspects of the parent involvement: parenting skills, learning at home, and communication with teachers. According to Quigley, the study is most interested in parent support for academic achievement. The project targets both teachers and parents. It provides professional development for teachers on how to work with parents, how to bring parents into the classroom, and how to involve them in academic concerns. Parents are offered workshops in parent skills, educational techniques, school policies and standards, and strategies for communicating with teachers about academics. The expectation is that behaviors and expectations of both groups will change.

Quigley described several results of the study addressing parent behavior. She found that a limited amount of informal training of teachers on how to work with parents existed at the elementary level and that this type of teacher training was almost nonexistent at the secondary level. Workshops for parents were widespread in elementary schools, but not at the secondary level.

“Quite a few parents knew the type of workshops that were being offered,” she said, “but only 20-40 percent attended any kind of workshop, and only 50 percent had visited or used a parent center.”

As for baseline data on parent-teacher relationships, teachers felt their relationship was relatively positive, but needed improvement. Most parents felt very positive and that they were kept informed about how well their children were doing. They were less positive on the question of whether teachers listened to them or that they shared common expectations for their child.

Results also showed that parents do a lot of reading with their children, indicating "that there is a base out there for teachers to rely on that they might not know about," Quigley said. And parents seem to provide some structure for studying at home, such as limits on TV and helping with homework.

In summary, Quigley said, teachers' professional development in communicating with parents needs to be intensified and more consistent, and parents need more strategies to better communicate with teachers. "Parents seem to be on campus, but schools need to figure out ways to engage them in the academics of their children," she said. Teachers need to help parents structure learning time at home.

Models for Assessing Problem Solving

Although student collaborative work is popular in the literature and in practice, not much data exists on how to measure achievement in group settings. Noreen Webb of CRESST/UCLA has studied the influence of group composition on both performance in groups and on what happens in groups when they tackled problems in science. She reported on a study that examined the affects of different student abilities within a three-person group.

Some consisted of low-ability students only, while others were high ability only. Most were mixed ability groups, with at least one high level ability student based on pre-testing. Webb reported the results of the low-ability students working in groups with at least one high ability student.

"There was a dramatic increase in the scores of low-ability students, not because they were copying the work of the high-ability students but because there was a lot of discussion going on," Webb said. The group experience carried over to the individual tests given the next day. Furthermore, achievement was positively correlated with the accuracy of answers, she said, with groups having a competent person included being more accurate. Students seemed to learn by being actively involved in group discussions, asking questions, and trying to paraphrase, she added.

The results presented a dilemma, however. High-ability students did better when they were in a group of other high achievers, yet their presence in a group of lower achievers helps the lower achievers do better. Heterogeneous groups produced much more incomplete and inaccurate information. Where there was rich discussion of concepts, Webb observed, “students did very well, but where the discussion was incorrect or incomplete, students scored much lower.”

The research led to asking further questions, such as the extent to which students challenge each others’ ideas, how students paraphrase each others’ work and synthesize different ideas, and to what extent do students bring in external ideas.

CRESST is working with teachers in Los Angeles and Chicago to design performance assessments that, among other things, focus on problem solving skills to find out if students understand what they are learning - if they are constructing the big picture. By bringing specifications to the table when working with teachers, “we are finding the models that allow us to achieve better alignments with standards in the districts,” reported David Niemi of CRESST/UCLA.

Discussing theories of learning as a background for the research, he said that past design research depended on behaviorist theories about problem solving which contend that if responses to stimuli are reinforced positively, the learner will more likely repeat the response. Consequently, the research focused on influencing behavior toward new tasks, such as how familiar a student is with problem solving, the number of words in a word problem or its length, the number of computations in a world problem, the sophistication of the vocabulary, the degree of concreteness and similar issues.

Most of this research, however, did not look explicitly at what students were doing when they solved problems, what mental processes they were going through, Niemi said. The research changed with the advent of cognitive sciences, which contend that learners “are born with certain kinds of competencies in language, certain basic principles that they know how to use. Newly added to that is the idea that high achievers have basic, important qualities that allow them to solve problems, not always consciously.”

As an illustration, Niemi described a math task given to both novices and those with higher levels of expertise, one which required them to decide which problems should be grouped together because they are similar. Novices, he said, tend to put together those that look alike or have some surface similarities, like blocks on an inclined plane. Experts put together things that look completely different but they interpret as belonging to a particular domain, a solution based on knowing how to solve these kinds of problems. More than just prior knowledge, the expert seems to have some “operational confidence ... that is highly organized around a relatively small number of principles / domains, an organizational elaboration around those ideas.” The novice problem solvers typically don’t have such flexibility, he said, because they “are stuck with trying to solve problems that are stuck in routines.”

With this background, Niemi described the research project with teachers, beginning with thinking through with them what the important ideas are that students need to learn in a certain area. Although a lot of school districts now have standards, many of them are stated in exclusively behavioral terms, he said, and the danger in depending on them is that “it will train kids to produce behaviors without developing the underlying understanding of the structure.” So, the project asked teachers to answer such questions as: What are the ideas kids need to know to do these problems successfully?; What kind of facts should kids be learning?; What should they be able to do with those facts?

Working with 150 teachers in four subject areas in Los Angeles, the teachers designed tasks integrated with the ideas they were trying to teach. In three days, the teachers developed a lot of tasks, but after a review one-third were discarded because they did not meet the criteria. Either they were not complex or they required too much work. Going through the process, teachers began to get the picture, he said. One conclusion from the early research is that there is no evidence of success with teaching generic problem solving skills. They need to be taught in the context of each subject matter, he said.

Speaking on models for measuring science achievement, Maria Araceli Ruiz-Primo described studies by her, Richard Shavelson, and others at Stanford University. She explained two current CRESST projects, the first with three purposes: to link assessments to cognitive components; explore technology for developing technically credible assessments; and advance the understanding of good assessment task characteristics; i.e., structure and nature of the response. As part of their research they analyzed concept maps and performance assessments, the domains and skills that they measure.

One of their findings was that multiple-choice tests and concept maps appear to measure overlapping but somewhat different aspects of students’ declarative knowledge (knowing that something is true). Their study led them to believe that several characteristics of knowledge should be considered in achievement testing and that such testing can help identify competent and less competent students in a domain. They also found that the characteristics of the assessment tasks influenced the sensitivity of the assessments to improvement in performance.

They also evaluated performance assessments and concept maps used as an assessment. Concept maps, she said, help students organize knowledge while performance assessments reflect procedural knowledge. An issue with concept maps is how structured they should be. “If they are too structured, we are imposing ideas on a student,” she pointed out. A separate issue revealed by research on performance assessments is why students are inconsistent from one performance assessment to another. Their research leads them to believe that students’ partial knowledge or understanding may be causing the observed performance variation.

Problem Solving and Technology

Concept mapping also was the core of discussions in another session, only in this one it was linked to technology. Computer-based concept mapping, noted Davina Klein of CRESST/UCLA is a way to represent information and to score assessments quickly and at less cost. Using some examples, she explained scoring methods, based on comparing student concept maps to expert concept maps (medical students had been asked to construct maps of what big ideas in science students should know, then teachers reduced them to a 4th-grade level). Future work on this study will consider if the concept maps are measuring what they are intended to and how the assessment can be linked to instructional purposes.

Another aspect of the study is looking at the improvement of student concept map scores over time and how students are using the maps to develop ideas. Ellen Osmundson of CRESST/UCLA reported on an incremental understanding scoring system, ranging from “illogical” to “grasping concepts over time.” The initial map scores were not significantly different between an experimental and a control group. The final map scores and essay scores of the experimental group, she said, were considerably higher. However, over time the experimental group developed greater science understanding.

Osmundson’s conclusion was that “participation in group mapping activities supported development of science understandings and connections across systems. The scoring system also shows that incremental scoring allows us to see students’ ideas progress." In the future, the researchers are interested in studying other science domains and scoring methods and doing further analysis of the group work.

Howard Herl of CRESST/UCLA reported on a project requested by Statistics Canada to study knowledge mapping and answer certain questions. What is the impact of providing certain kinds of content information to adults taking the assessment? Would telling them how it would be scored have any effect? What are reliability and validity issues? The study obtained information from participants beforehand such as spatial ability and background information. Some participants were then given content information, some not; some were told how it would be scored, some not. The task used the properties of a bicycle pump to illustrate the human respiratory system.

Results, said Herl, showed that participants benefited significantly when receiving content information; i.e., diagrams, on both the knowledge mapping and problem solving tasks. But providing scoring instructions to participants did not significantly affect their performance on knowledge mapping tasks. Latent variable analysis yielded high generalizability (.83) using multiple indicators to measure adults' problem solving skills. There were no gender differences, despite the bicycle pump example, which Herl had hypothesized might favor males over females.

Panel

Assessment of Special Students

There are both research and political reasons for including limited-English-proficient students and students with disabilities in assessment systems, Lorrie Shepard of CRESST/University of Colorado at Boulder said at the opening of a general panel discussion on the assessment of special students. To exclude them, she said, distorts the accuracy of assessment results and removes them from the accountability system. Therefore, attention must be paid to accommodations, which she defined as "adaptations or changes in how assessments are administered or the mode of response with the intention of removing irrelevant sources of difficulty."

The problem is that although the use of accommodations is intended to level the playing field, in some applications "it instead tends to inflate test scores and may undermine rather than enhance test results." For example, in some cases mentally retarded students in Kentucky were scoring above the state average on that state's assessments. An ideal study, Shepard said, would be to look at the effect of accommodations in a random assignment for accommodations and non-accommodations to see if they help all students.

But working in a “real world” context, she analyzed statewide data in Rhode Island, focused on math at the 4th-grade level. Statewide, accommodations had been offered for both a standardized test and a performance assessment. A Spanish translation was used for the performance assessment but not with the Metropolitan standardized test.

Among the findings Shepard reported:

- accommodations increased participation of LEP students (more for the performance assessment than for the standardized test);

- most common school practices were extreme; that is, in most schools all students were accommodated or none were accommodated;

- on average, LEP students performed better on the performance assessment than the Metropolitan.

The variance in school practices raises questions about teachers’ reasons for using accommodations, she said. Further, most accommodations were administrative, especially giving students more time. Spanish language translation was used rarely. Four schools accounted for most of the “incredible gains” in achievement.

In reflecting on the findings, Shepard said new questions have arisen on the topic of inclusion. Is it fair to ask whether or not every pupil tested with accommodations accomplishes the original purposes of increasing accuracy and assuring greater accountability of LEP students? “As exemplary as is the effort in Rhode Island,” she said, “I doubt if it is getting better comparisons school-to-school. Rather, what we are adding is a variance of whether the accommodation is done well or poorly, generously or not so.”

Shepard suggested as an alternative model that students be accommodated on a sampling basis under the supervision of trained administrators, “if the purpose is just to gather good data.” If the purpose is to affect instruction and get teachers to be better at accommodations in day-to-day instruction, “then maybe that needs to be put on the table as a primary goal and think about pre-training for assessment programs or clinicians visiting schools and modeling accommodations.”

Jessie Montano of the Minnesota Department of Children, Families and Learning described how the growth in the LEP student population in the state created a dilemma when statewide testing began in 1995. Students in the country more than 12 months are expected to take the tests, math and reading tests first administered in the 8th-grade and a 10th-grade writing test, all of which students must pass to graduate. LEP students did not do well in the first administrations of the test, she said, but focus group sessions with families from the five largest language groups indicated parents wanted their students to be held accountable for passing the high-stakes tests.

Her department put together a committee of bilingual and ESL teachers that is in the process of developing a test for LEP students that will measure reading and writing skills in non-native speakers. The test will measure acquisition of language, not listening and speaking skills, and be administered in grades 3, 4, 6, 7, and 8. It was to be field tested in October, and the designers are hoping to find out if these tests give accurate information on LEP students, accurately predict performance on statewide tests, and provide information about the quality of LEP classes.

In a continuing study of accommodations given in Kentucky on its statewide assessments, Daniel Koretz of Boston College and RAND reviewed some earlier findings and reported new ones. At last year’s CRESST conference, he reported that Kentucky schools had successfully integrated students with disabilities into their assessment, with nearly 90 percent included in the assessment system. He also found that most Kentucky students with disabilities tested, especially in elementary schools, had at least one accommodation. Nearly three-fourths of all tested elementary school students with disabilities had at least part of the assessment presented orally, and about half had some of the assessment paraphrased for them. Half had been provided extra time and half had part of the assessment paraphrased for them. Another interesting finding last year was that accommodations became less common as students progressed through school. But perhaps of greatest interest was that the scores of some groups of students with accommodations were implausibly high; e.g., mildly mentally retarded students with certain accommodations had mean scores roughly comparable to the average of non-disabled students, and learning disabled students with the same accommodations had mean scores well above average. These findings last year suggested that accommodations were being overused, generating biases instead of offsetting them in some cases.

In a follow-up study, examining data from two years later, when Kentucky had reintroduced multiple-choice items, Koretz and his colleague Laura Hamilton found that Kentucky still managed to include a large number of students with disabilities in their assessment system, about 90 percent. The use of accommodations also remained similar. Relative to students without disabilities, the performance of students with disabilities was roughly similar on the open-response and multiple-choice components of the assessment. But scores of elementary students with disabilities on the open response part of the assessment dropped sharply from two years earlier. This drop was attributable to students receiving accommodations.

"I don't have a clue why performance dropped so dramatically with students with accommodations," Koretz said. "It is tempting to look at data and suggest that later results are a more accurate portrayal. What we don't have is a good criterion, a trustworthy measure of what the true performance of these kids is."

Also, he continued, "the similarity of differences between the non-disabled and the disabled in two formats might be encouraging except there, too, we don't have a clue as to how much of it is due to use of accommodations, which in Kentucky is very liberal. If accommodations were controlled, we might be able to find out." Whether the open response format is comparable in difficulty to the multiple-choice format for these students. Nor, he said, are there good explanations yet for format differences that did emerge.

Koretz called for experimental data to separate the impact of accommodations from the characteristics of students because at this time researchers have no control over accommodations, the assigning of which depend on students' Individualized Education Plans. Also, researchers need a better criterion measure and a more targeted examination of accommodations, as suggested by Shepard. At this point, he said, "there is some evidence that use of accommodations in an unregulated system is very problematic."

Frances Butler of CRESST/UCLA described an exploratory study of large-scale test accommodations with 7th-grade English language learners (ELLs) in Burbank, California. Two types of accommodations were available to ELLs in six social studies classrooms: a) extended time and b) the reading aloud of test items and directions by the test administrator.

Student preference for the accommodations was also studied. After taking a standardized test in social studies without accommodations, students were asked which of the two accommodations they would prefer to receive if given the test again. On a subsequent administration of a parallel form of the test, one-third of the students received their accommodation of choice; one-third received the accommodation they did not choose; and the final third were randomly assigned one of the two accommodations.

Data analysis found that students did not significantly improve their performance on the accommodated test with either accommodation, even when given their preferred accommodation. Data analysis also found no significant relationships between accommodation preference and various background variables.

Follow-up questionnaires and focus groups revealed that unfamiliar vocabulary was the major problem for many of the ELLs in the study. A number of students indicated that having an English dictionary would have been more helpful than either extra time or having the directions and items read aloud.

The results of the exploratory study suggest the need for: caution in using accommodations, a better understanding of the impact of English language proficiency on the standardized test performance of the students, and attention to the issues of opportunity to learn and test reliability with ELLs.

This work is reported in detail in Castellon (1999).[1]

Using Technology to Improve Information Systems

How will technology transform large-scale assessments? This is a critical question as more attention is focused on high-stakes assessments at the K-12 level, said Randy Bennett of the Educational Testing Service. He outlined the issues that need to be addressed.

- **Test design.** A system is now under development which details the process and the tools needed for explicitly using cognitive principles as the basis of test development. A major issue is determining if these tools will be efficient.

- **Item creation.** Technology allows automatic item generation which is important because it allows tests to be created more efficiently and to be schema-based, “not as a large collection of unrelated problems but as a

[1] Castellon, M. I. (1999). *The performance of English language learners on large-scale academic achievement tests* (Master’s thesis, University of California, Los Angeles).

smaller set of more general problem classes that we want students to be proficient with." The issue is insuring that the tool will be used thoughtfully. "If used thoughtlessly, it could rapidly create bad tests."

- **Task presentation.** An example is computer-based simulation such as interactive case management simulations developed at Marshall University School of Medicine. This is important because the computer allows delivery and capture of performances, presenting tasks that contain many of the complexities in which students should be competent. The issues are deciding if the tasks capture the performance and skills required; and can the tasks be developed, delivered, and scored cost effectively.

- **Scoring.** This is the key to efficient use of simulation and other large-scale assessments. At least five programs exist to do automated grading of essays. This is important because it could allow large-scale assessment programs to deliver essay tasks efficiently and routinely. The issues are deciding if these programs really work, if the education community will accept use of these programs for high-stake purposes, and if the use of these programs will negatively effect instruction or staff development.

- **Testing purpose and innovation** (i.e., distance learning). This may make possible the simultaneous generation of assessment information. For example, homework could be submitted electronically and a proficiency estimate created based on multiple samples. An issue is the effectiveness of the strategy: "how will we know the person is who we think it is, will that person accept electronic scoring, how will the mass of information generated be used?" This may shift the focus of large-scale assessment in very fundamental ways: from selection to knowledge certification. It may change large-scale assessment from an isolated tool to part of an embedded curriculum. The fact that the Internet can fundamentally affect instruction and testing is not yet appreciated.

"The partnership of advances in cognitive sciences with technology will permit the transformation of large-scale assessment and the development of better information systems," Bennett concluded, "by allowing us to create tests more firmly grounded in conceptualizations of what student need to know to perform." The partnership also will allow performance assessments to be more practical and routine and change ways in which assessments are delivered. Furthermore, he said, while technology facilitates this transformation, it should not be allowed to drive it. "This technology must be driven by substance, by educational needs, and by cognitive measurement and domain-based principles required to respond to those needs effectively."

Testing is no longer a one-shot affair but a process with lots of feedback. With technology, item creation becomes much more systematic and includes scoring criteria. The advantage of developing tests in this matter, said Isaac Bejar of ETS, is that it can develop many forms that are similar to each other, sets that are of one schema. It can tell about performance without repeating the same items. There aren't many applications of this approach, but one standout is the licensing exam for architects in which a candidate could be asked to generate specs for a two-story building. Such models for computer-assisted assessment don't fit the least experienced teacher, but there are models for more experienced teachers. Computer-based assessment can help accelerate the use of technology in the classroom, especially its ability to meet some of the challenges in performance assessment. "As a result," he said, "I see no conflict between technology/accountability issues and education reform."

Eva Baker, co-director of CRESST, focused on technology uses in assessment in different ways. It occurred to us that we need to think of assessment from an indicator viewpoint and indicators from an assessment point of view." The longer-term goal may be the place where Issac Bejar ended, she said, "to use technology to reconcile perceived conflicts between accountability and instructional improvement uses of assessment." Right now, technology doesn't let us differentiate between purposes. So we need to make the proper decisions with available tools while looking for ways to strengthen the effort.

What is interesting in the K-12 environment, Baker said, is that there is little use of technology in systematic instructional improvement and almost no work in system monitoring, although the new NAEP contract to ETS should accelerate the latter.

Why is there an impetus for technology-based assessment, she asked? First, "we are thrilled it could happen." The other reason is economy and efficiency. And she took some issue with the argument that technology-based assessments should serve pre-existing requirements because "my view is that they tend to redefine the requirements as we go along."

Some of the issues surrounding technology-based assessments are not connected to education reform, such as whether there are enough examples for people to understand what is being talked about and not be fearful of it and whether or not some of these assessment systems are sufficiently connected to policy-relevant issues such as standards. In general, Baker said, the development of assessment systems must continue to look for domain-independent mix-and-match assessments that can be used up and down the system for classroom and accountability purposes. They should allow people to customize items, but still be conceptually linked to the standards.

CRESST is working on all of these issues, Baker said, naming many of the conference topics and presentations. Some are quite creative, including the development of a set of stickers to put on concept maps that can be scanned and provide an approximation of an automated scoring system.

The indicators work came about because people want more information, more trustworthy information, and more interactive information. Present indicators tend to be periodical and untimely, coming out in pieces that can't be put together in any particular way. For example, CRESST reviewed all of the report cards available on the web, she said, and found, in addition to the timeliness problem, that the report card indicators were not arranged in any priority and few differentiated between what was under the control of schools and what was not. Another area of indicator study is that of quality school portfolios which allow school-based and district-based data to be reported in multiple ways. Title I reports will shortly be automated in this form, she said.

Assessments and indicators developed through technology, must have a better focus on content and feedback purposes. Further, such assessments need summaries that are understandable to students and the public. All the while, they must provide valid comparisons and link clearly to evaluation and policy.

James Pellegrino of Vanderbilt University described another example of an effort to use technology to scaffold student learning with understanding. He and his colleagues are developing an integrated learning and assessment model in math and science that connects problem-based and project-based learning. The Science and Math Arenas for Refining Thinking (SMART) provides substance around experience. "It is not sufficient just to generate interesting materials and environments and then just put them out in the classroom," Pellegrino said. Learning from them "requires sophisticated and well-constructed scaffolding. We need technology to help students see what they are doing." Projects like testing for river quality may seem interesting, but students don't learn much from them because "the environment is not constructed for learning."

In the SMART project, the Web provides individualized feedback to students on their choices for information. The SMART lab collects data and analyzes their choices. Students not only choose from a catalogue of information, but they also must select a justification for their choices. The feedback directs them toward resources and further thinking about what they are doing. The whole idea, said Pellegrino, “is to move them along on conceptualizing.”

Student answers are collected and displayed through SMART, then discussed by students and teachers. Kids-on-Line presents vignettes by student actors that give explanations. A whole set of resources accumulates for students before they ever test the river water. The ability to learn from this activity greatly increases because of the scaffolding that precedes it, he explained.

SMART was designed for middle schools and has worked in a Title I school. According to Pellegrino, students show considerable gains from pre- to post-testing on several items. He concluded by emphasizing that “achieving high-learning outcomes requires well-designed learning and assessment environments. The goal must be to find ways of creatively using technology to give teachers and students the feedback they need.”

Breakout Session

Quality School Portfolios:

Reporting on School Goals and Student Achievement

A major problem in school reform is that information related to student achievement “sits on a plateau unrelated to the school world,” according to Derek Mitchell of CRESST/UCLA, “and is not used specifically to improve teaching and achievement at the school level.” The Quality School Portfolio project gives a school control over the information, an opportunity to examine the information often, a way to come up with indications of what it needs to do to improve achievement, and a means of infusing professional development with a capacity to ask the right questions.

Ordinarily, schools create goals that are either not very measurable in site-based plans or are not measurable with existing data. The Quality School Portfolio project, instead, hopes to do a better job of linking data currently available to goals schools have set for themselves, giving them a report on whether they actually are achieving what they said they were going to do. The project links schools to district databases through the Internet. They can dial up any information they want and store it, even down to the level of storing a paragraph written by a student that shows the learning wanted by the school. The portfolio can show data across years, student backgrounds, teachers of individual students, and results from multiple assessments.

Discussing the process used by schools to set goals that are reported in the school portfolio, John Lee of CRESST/UCLA said the project gets specific, allowing up to three targets for each goal. Researchers are helping schools select multiple measures for goals and connect information in the database with the actual goals to determine if the goals have been met. Some of the questions generated by this process: How do students with special needs compare on the SAT-9? What are the grade distributions on math for various feeder schools?

Another part of the quality school portfolio covers non-cognitive information, using surveys, questionnaires, observations, and interview group protocols. The information includes safety and security, parent involvement, curriculum and instruction, professional development, and technology and innovation. All the information is computerized. The next phases of the project will include translating some of the measures into Spanish and other languages, adding new measures, coordinate the portfolio with the school report card, continue quality control testing, and develop a quality portfolio for the district.

Models of Standards-Based Assessment in California

Having to produce multiple measures to satisfy Title I requirements is a major challenge, according to Lynn Winters, Assistant Superintendent for the Long Beach Unified School District. It is impossible to assess all of the California academic standards, so evidence of student work at the classroom level must be part of the data collected. Winters' concerns also include lack of time, technical issues in reporting school and student performance, and the misclassification scoring of schools.

On the other hand, Robert Ferrett of the Riverside Unified School District, believes that multiple measures are helpful. Riverside looked at what teachers already were using systemwide, then set end-of-year goals for each multiple measure and interim goals. At each juncture, the data are analyzed to help make decisions on how to get each student closer to the end of year goals. The starting point for reading goals are the Scholastic Performance Assessment levels and the Harcourt Brace Performance Assessments and levels. His district is using three measures for reading and math administered three times a year. No child will be reported in the assessment who has only one measure, nor will the district include data on a child who has not been at the school all year. He predicted that it would be simple to cross-reference the SAT-9, the writing test, the district's tests, and grades.

The Visalia Unified School District also is using multiple measures including the SAT-9 in grades 2-11 and a number of additional tests, according to Younghee Jang, director of program evaluation. As an example of the complexities of multiple measures, she showed how state 10th-grade SAT-9 scores, indicated that 121 students met grade-level standards, but by using grades only, that figure drops to 97.

Some of the discussion focused on scoring strategies for district-wide assessments, with each district taking a slightly different approach. Long Beach mandated a one-day scoring program for all teachers while Riverside had two leaders at each school assigned to recruit scorers. San Luis Obispo used experienced substitute teachers for scoring. Visalia trained teachers and paid them to do scoring in the summer. Winters noted the different approaches and said, "you cannot make everything high stakes and be loose about the scoring strategies."

Tempes noted that a tension exists around standards-based accountability because the state system is theoretically based on local standards. State standards are not mandated, but the trend will be to use them more and more, he said, because that's what the state is testing and reporting. Added Winters: "No one knows if the decisions we're making on numbers really have consistent meaning."

Asked how to use the data to motivate students, Tempes replied that "we're trying to create a structure. What people do with it is their own decision. Some treat it as an intellectual exercise, others are really using it to guide instruction and motivate learning."

Looking Closely at Teacher Expectations and Learning Opportunities

To provide better feedback to teachers by using indicators of classroom practice, CRESST researchers are looking at evidence of changes at the classroom level as a result of reforms instituted by the Los Angeles Annenberg Metropolitan Project (LAAMP). "What are teachers expecting students to do? Does it add value? What kinds of learning goals are there? These are some of the questions being asked in the project, according to Pam Aschbacher of CRESST/UCLA. As a piece of the evaluation, the researchers are taking a sample of teacher assignments to determine what the learning goals are, how students are graded, and “if there are mixed messages in all that,” she said.

In the process, the researcher team is trying to come up with rubrics for good assignments, but the process also raises issues. Aschbacher said they are wondering what assignment examples should be asked for and how often, how many pieces are needed from each person, how to make appropriate evaluations rather than work from their own biases, and how to minimize the burden on teachers? For the first effort last spring, the researchers asked for six tasks during a single month: classroom assignments, homework assignments, grading criteria, and data on how students performed at certain levels on the assignments. With the data in hand, they are now trying to answer those questions.

Researchers are collecting classroom work in another major California reform effort: reduced-class sizes in the early grades. The goal is to see if smaller classes produce differences in instruction. Brian Stecher of CRESST/RAND said that teachers in the study are asked to fill out daily logs over a 10-day period and answer closed-end questions like how they organize the class, what sort of instructional approaches they are using, and what work students do in class. Teachers also are asked to select student work completed in response to these assignments, that they would describe as reflecting high, medium and low performance.

“The sets of logs allow us to construct nifty activity profiles,” he said, “but we are struggling to interpret the profiles.” So far, class size does not “predict” specific pedagogy. “There are some provocative patterns in teaching, but nothing so far that is consistently different between small and large classes,” he said. They are finding great use of small groups in smaller classrooms, but Stecher expressed frustration “at the lack of clarity in teachers forward example description of the lessons particularly in terror of what they were trying to accomplish. Most of their goals were very low level, most of the feedback methods they used were checklists,” he said. “Their responses on the logs were terse and not very revealing. We didn’t get much insight yet, but that doesn’t mean it isn’t there.”

In the LAAMP evaluation, responded Aschbacher, teacher interviews produced greater depth than other documentation. She noted that the California teachers had a hard time articulating what their goals were, while Kentucky teachers she has interviewed “can talk the language of standards reform and know what an activity is aimed at.”

Helping the Media to Understand Quality of Practice

Bridging the culture of journalists who know the story they want to tell and the culture of ethnographers who are more interested in letting the story unfold was a challenge for Ann Mastergeorge and Lindsay Clare of CRESST/UCLA. Asked by the *Los Angeles Times* to prepare a group of reporters for classroom observations, they prepared a one-day crash course on ethnographic methodology. "We spent the time working through discussions of what is important and how do you find it, how do your know that you are writing about something that is quality," said Clare. Their sessions revealed several tensions between the media and researchers, such as the use of ethnographic field notes, documenting the whole picture rather than just a piece of it, and the importance of following just a few students rather than interviewing many.

One of the issues, according to Clare, was how to get comparable information with different reporters fanning out to different schools. To help provide some consistency in their observation strategies, the researchers and reporters visited a UCLA laboratory elementary school and then talked about what they observed. They also prepped the reporters on what to do before a visit and what to ask teachers before observing, such as: background information about the class, teacher goals of the lesson, how does the lesson fit into the curriculum, how are student groups formed, and what the teacher will be doing.

In the classroom, the researchers advised reporters to look at what resources are available to students, if the materials are up to date, interesting, sufficient, and include multicultural content. Reporters should determine if it is a teacher-led class, if the room arrangement encourages student interaction, and if students are engaged and focused. Classroom management factors include interruptions from PA announcements, warm relationships with the teacher, safety, and if the teacher is using a variety of strategies to encourage participation.

The researchers described several indicators of quality instruction for the reporters.

- Are concepts being made explicit, are teachers providing relevant background, are students being given explicit guidance on reasoning and complex thinking skills?

- Are students being asked to justify their answers based on content evidence?

- Are students being asked to describe their problem-solving processes?

- Are teachers asking the kinds of questions that require knowledge of subject matter but not necessarily a correct answer?

- Are students engaged in purposeful activities?

"Many of the journalists were very savvy about classrooms," said Mastergeorge. "What we tried to give them was a way to manage the data they were going to collect .... We encouraged them to let the story unfold. To look at the sub-cultures in the classroom, and if there is chaos, to ask what kind and why." She also said that if journalists are going to tell a story, they must have data points to support it. "They had the dramatic points they wanted to make," she said, "and we wanted to be sure those points were substantiated by data."

Lessons Learned from the California SAT-9

As part of a statewide educational accountability effort, massive testing of California students was undertaken during the 1997-1998 school year using the Stanford Achievement Test - Ninth Edition (SAT 9). Nearly all students in grades 2 through 11 participated, including many students with limited English speaking skills. CRESST researcher Rich Brown overviewed the testing program and discussed many of the political issues that arose from one of the largest assessment programs implemented in less than one year.

For example, how well California students did on the test largely depended on who was doing the talking. Democrat Delaine Eastin, the California Superintendent of Public Instruction said that "… the scores indicate that California students are coming back." Meanwhile, a spokesperson for Republican governor Pete Wilson said, "These scores clearly point out that we must continue our efforts to reform California's education system." *The Los Angeles Daily News* headlined student performance "DISMAL SCORES."

Brown also pointed to some unexpected results including several counties where bilingual students outperformed LEP and English-only students across grade levels and subject areas, including mathematics, with a significant drop-off in reading achievement between 8th and 9th grades. Statewide scores dropped more than 15 points between the two years, and dropped even more in 10th grade before increasing in 11th grade, he added.

Brown noted that the state would introduce even more testing in 1999, a result of the SAT 9 tests non-alignment to the state standards. Plans to imbed the SAT 9 tests with additional items linked to California standards were abandoned due to both technical and practical problems. The only way to measure those standards was with additional testing, said Brown. He also discussed other problems such as the exclusion of many LAUSD students on the test. In 1998, 48,000 LAUSD students, many of them with limited English proficiency, were excluded from the test compared to less than 1,000 the year before.[2] Another problem is that the norms for the Stanford 9, included only 1.8 percent of students who were LEP, while LEP enrollment in California is nearly 25 percent of all students. These are important factors that can significantly influence the test results.

Panel

Bringing it Together: School Quality Indicators

Which indicators are important to the public, to researchers, and to both communities? What are the limits and possibilities of test score data as an indicator? How do researchers make a technical indicator, such as nuances in test scores, intelligible to the public? These were some of the issues discussed in a panel on school quality indicators that included both researchers and a well-known journalist.

Researchers may fault the media for creating a horse race out of test scores, "but they do tell parents and the public something," Richard Colvin of the *Los Angeles Times* said to open the panel discussion. "I know there will be people who draw improper conclusions and who even make decisions about buying a house on test score data, but that doesn't mean we shouldn't report them and try to find out what they mean."

Digging deeper was an assignment that sent Colvin and other reporters to CRESST for help. Asked by an editor to analyze NAEP data and demographic trends in the state, the reporting team enlisted CRESST researchers to help analyze the data. An extensive amount of work basically only yielded performance for the college-bound, he said.

Part of the assignment included spending a week in seven high schools. "It is incredible what we learned," Colvin said. "In some places libraries are full of books

[2] As part of their new assessment system, LAUSD implemented the Stanford achievement tests one year prior to state adoption and implementation.

and students, in others, libraries are a wasteland. If you see a row of copies of *The Old Man and the Sea* that haven't been checked out in 15 years, whereas before that they were checked out regularly, that tells you a lot." The reporters learned that there was an "incredible difference in standards among schools, with some teachers not assigning homework because they know student's won't do it or don't have books, and other places where the pace was incredibly fast and full of high intellectual activity." The result was 20 pages of stories reported across three days of the *Los Angeles Times,* which, Colvin said, "I hoped made it clear that as journalists we were not only interested in horse races."

The public can't undertake such a study, but Colvin suggested a way that schools could convey their standards and expectations. By using the Internet, he said, schools could post what an A paper looks like compared to a C paper, present the syllabus for AP sciences, and other indicators of what the school is trying to accomplish. "There ought to be ways beyond test scores for the public to get some sense of what a school is about," he said.

Reinforcing this point, Yeow Meng Thum of CRESST/UCLA pointed out that test scores change over time, "but the change doesn't happen in a goose step," so it is necessary to find out how they change and the effect of reforms on them. "There are lots of methodological challenges in this area of research - we have to go back and a take fresh look at every possible source of this problem." Furthermore, he said, it is no longer possible to bury some of this knowledge in technical reports; it needs to be made public. He described a value-added definition of change that would be made up of a small number of indicators that could be grasped easily. Indicators can be either plain, such as a truancy rate where not many analyses are needed, or summative, which would be changes over time in truancy rates and require two or more numbers to be put together.

Previous studies looked at baselines and benchmarks, but his current research is looking at summaries of performance patterns, such as how the distributional patterns of performance for schools in a school system might be described and how changes in individual and overall patterns of performance might be detected.

William Schmidt of Michigan State University, addressing the policy implications of TIMSS, questioned how extensively a single test score can be used as an indicator of quality. In TIMSS, for example, rankings differed by how specifically the curriculum was defined. The 8th-grade results in math were an aggregate of all sub-groups of content, but to use the overall test as an indicator of school quality around the world is a misinterpretation of the data, he said. When looking at approximately 40 specific content areas in science and math, the United States' performance was not consistent, very high in some and low in others. The rankings depend upon which part of the curriculum is being talked about, he said.

"How can you use these data for decisions about school quality?" Schmidt asked. "Is someone going to say, I will move to that neighborhood because the structure of matter is well taught there even though the school falls down on genetics?" The American public does not understand that if simplistic test scores are going to continue to be used as an indicator, "we are doing a very imprecise job of measuring quality."

Such indicators are used, Schmidt charged, because "there is a fundamental belief in the American education system that content doesn't matter. As long as we are teaching something called math, it's okay. When people talk about quality, most exclude content, and refer to instructional practices or generic teaching skills." Instead, he said, content-specific notions should be included in the definition of quality - and that will happen only if the content is commonly specified and the test is lined up to the content.

Schmidt pointed out that almost every other country has clearly defined standards for all grades and schools. Therefore most school variation is very small. In the United States, almost 75 percent of the variation in math and science scores is attributable to school track or classroom differences, "which is at least in part a simple reflection that the curriculum is different."

The United States wants and uses achievement measures, but without a common content and as a consequence the results are terribly confusing, he said. He proposed as a first step to remedy the situation that the achievement measures be made curriculum specific and reflecting 10-20 different areas of content. "Let parents and the public be made aware of the specific strengths and weakness of the schools so that sensible changes can be made where necessary he said. Capping his remarks with reference to the release of the SAT-9 results in California, he asked: "Where were the indicators that showed exactly what children were being taught in each district?" Without those data, there is no way to interpret the meaning of the scores, he said.

David Green of the University of Pennsylvania has studied how to "know" schools in England, Scotland, and Australia, and most recently, Chicago. These experiences have taught him that "it is virtually impossible to know the whole truth about any school, but it is very possible to know critical aspects of the truth all about schools," a line of inquiry that he began developing three years ago with a set of schools in Chicago under a grant from the MacArthur Foundation. His purpose was to add value to the schools' work, but he realized that their experience with self-evaluation, primarily through regional accrediting agencies, was problematic "and less than rigorous." The accrediting agency, for example, had approved all of the high schools that were later put on probation by the district.

Quoting Cambridge Professor David Hargreaves, Green said that teachers are not trained in the skills of inspecting, auditing, and evaluating and are "inclined to parochialism." What they have to work with in evaluation is mediocre, such as simplistic videotapes that give them no sense of the reality of classrooms. "There is a difference between seeing what you know when you go into a classroom and knowing what you see," he pointed out, noting that many observers do teachers a great disservice when they assume they know what they are seeing.

To give more substance to self-evaluation, Green said, teachers need scaffolding that integrates the practice of teaching, student learning, and aspects of the school as a learning community. "There needs to be a connection between what teachers say they are doing and what students actually are taught," he explained. This effort at scaffolding in the Chicago schools began to focus on defining what is Challenging student work, leading teachers to evaluate whether students are getting the kind of structured experiences they must have if they are going to learn effectively.

The process, Green reported, is "causing disquiet" among teachers because it is becoming more obvious that "what is needed in these schools is not marginal change but seminal change." The project relied on many reliable sources, including the perspectives of teachers and administrators and data from the Chicago Public Schools and the Consortium on Chicago School Research, "but getting that data beyond the principal's door to the classroom was a tremendous challenge." The effort is necessary, though, because it helps schools find a way to tell their own story. Many others tell it, including test publishers and journalists, Green said, but schools need to represent their voice "with clarity."

Summing up the discussion, Joan Herman of CRESST/UCLA returned to the opening theme - that the public and parents want a simple answer - but pointed out that each presenter had elaborated on how complex the issues of assessment, quality indicators, and accountability are at this stage of their development.

Panel

Reinventing Assessment and Accountability to Help All Children Learn

Reinventing assessment entails five principles, Robert Glaser of CRESST/University of Pittsburgh noted to open the final panel of the conference. These include:

1. **Maximizing access to education.** In order to provide access to educational opportunities and lessen possible exclusionary aspects of evaluation, assessment practices must be primarily designed to survey possibilities for student growth so teachers can recognize and support student strengths and optimize learning skills. They should be built on three principles:

- community-based knowledge and competencies available through problem solving in everyday life (these can be related to school competencies and school content);
- abilities for self-regulation of learning by students; and
- redirection of student belief to the value of effort.

2. **Assessing achievement and attaining competence.** Not only the content of what is learned must be assessed but also the way in which information is used. The fundamental point is that we are taking on integrating the findings of cognitive psychology and psychometrics. Assessment systems, he said, depend intimately on the nature of human performance and how human beings learn.

3. **New methodologies.** The design of assessment needs to take on some of the characteristics of cognitive task analysis and the experimental study of assessment features in their settings in order to elicit target performances.

4. **Assisting instruction along with assessing teaching quality.** Assessment instruments can either be downgraded or enhanced by environments in which they are used and the skill of those using them. A fundamental issue is the integration of teaching and assessment in instructional settings together with coordinated use of

assessment to inform teachers of their own effective practice and the quality of their expertise.

5. **Reporting of assessment outcomes.** Much effort is being devoted to this problem and the need to put the most inventive minds to it. National and international assessments come up with interesting displays of outcomes, but we need new forms of reporting of assessment outcomes to students, parents, and the community. Displays are needed that not only refer to standards but also to patterns of growth and developing achievement and that give profile components of performance for discussion. Various media should be used for these tasks.

Is progress being made in these areas and in assessment, in general. Robert Mislevy of the Educational Testing Service drew parallels between architecture and assessment designers. In architecture, he said, expert knowledge is organized around deeper knowledge, which produces procedures that handle routine work but are not useful when architects deal with new materials or design new kinds of buildings. In the same way, "today, as designers of assessment, we are definitely encountering some non-routine challenges in assessment - and some opportunities." The opportunities include capitalizing on the advances in cognitive and educational psychology and new technologies. The challenge, however, is to decide "what the heck do we do with all these neat data?"

First, Mislevy said, what is important in the learning domain needs to be decided. There has been a focus on evidentiary reasoning because "what we work with is typically incomplete or amenable to more than one explanation," he explained. "Data only become evidence when they are related to some conjectures or some use. There is no magic bullet for reinventing assessment," he said. "Somewhere along the line someone is going to have to understand the evidentiary principles and work through them to get to the nuts and bolts of applications."

In this work, Mislevy continued, three basic models have to be present and coordinated in order to achieve a workable, coherent assessment. They include:

- the student model - what complex of knowledge, skills, or other attributes should be assessed;
- the evidence model - what behaviors or performance issues reveal those constructs; and
- the task model - what tasks or situations should elicit those behaviors.

The student model variables “could be one or hundreds, qualitative or numerical - whatever suits the intended use of the assessment.” Cognitive task analysis and instructional goals help the assessment designer determine the key aspects of proficiency to assess.

An evidence model expresses how what is observed constitutes evidence of student model variables. First is understanding how to identify and evaluate the key features of what a student says or does - that is, determining the values of “observable variables” from the students’ work products. Depending on the assessment task, this might be simple or complex, accomplished automatically or through human judgment. Then, a statistical model can be used to integrate this kind of information across tasks, by updating beliefs about the student model variables. The task model provides a framework for describing and constructing situations which examinees can act. The focus is on evidentiary value of tasks - just what aspects in a situation are necessary to give students an opportunity to show what they know. The flip side, Mislevy said, is understanding which aspects are can be modified or adapted, while still providing evidence about the important knowledge. This allows assessors to customize assessments; that is, to get different evidence from different students, but still interpret the data in the same evaluative framework.

Mislevy pointed out that there are some opportunities to apply these ideas in assessment. A conceptual framework allows “a powerful connection between assessment and learning to be forged by making explicit the standards of good work.” Another possibility is to develop assessment packages with built-in ability to customize the assessment, as in the Advanced Placement (AP) Studio Art portfolios. In the process of building evidentiary reasoning structures into education products, “you don’t have to know how it works,” he said. “It can be built in behind the scenes, making assessment an integral part of the instructional program.”

Over the years, there has been good progress made in the methods for gathering and using data in familiar forms of assessment, Mislevy said. Using off-the-shelf methodologies “actually can tackle some pretty tough evidentiary problems, but there are still gaps between assessment users, policymakers, innovators and test theory specialists.” Critical problems arise from new demands for assessment and expanding technology to capture data, both of which can outstrip standard practices. “We can’t solve these problems without thinking about them,” he said. “We can attack new assessment challenges, however, by working from generative principles of assessment design - principles of evidentiary reasoning, applied to inferences framed in current and evolving psychology, using current and evolving technology to help us gather data."

Speaking of a specific assessment of quality teaching - the effort of the National Board for Professional Teaching Standards - Lloyd Bond of the University of North Carolina/Greensboro outlined several characteristics of its assessments. They are performance-based, call for an appropriate balance between content knowledge and appropriate practice, must contribute to professional development, and must meet standards of reliability and validity.

Postscript ...

Repeating Mislevy's comment that "we can't solve these problems without thinking about them," Robert Linn ended the conference with the challenge to deal with the tension of crafting systems of assessment meant to improve student learning, but at the same time meant to satisfy the needs of accountability. "We have a lot of work to do," he said. CRESST is intensifying its work to bring the necessary tools to policymakers and practitioners.