Assessing Student Achievement: Search for Validity and Balance

The 1997 CRESST Conference

CSE Technical Report 481

Anne Lewis

June 1998

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
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 Center Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

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

ASSESSING STUDENT ACHIEVEMENT:
SEARCH FOR VALIDITY AND BALANCE

THE 1997 CRESST CONFERENCE

Anne Lewis

Introduction

With uncanny timing, the 1997 CRESST Conference confronted issues about proposed national tests at the very moment the proposal was being debated on Capitol Hill. While the east coast discussion was almost purely political, the three-day CRESST Conference made it obvious that such large-scale testing would raise challenging technical issues which, ultimately, also would become political ones. As the debate continues through several required Congressional studies, it will become even more apparent that the CRESST Conference agenda anticipated the large as well as less obvious concerns of assessment experts and policymakers as they tackle this new phase in assessment policy.

The CRESST presentations and discussions made the conference theme—searching for validity and balance in student achievement—even more urgent because **validity** certainly was on the minds of those who had examined the issues related to national testing. And **balance** was an issue for all who placed the idea of national tests into the context of local realities and community values. These issues "are essential to every testing program," CRESST Co-director Eva Baker noted at the opening of the conference, but they are particularly pertinent to the idea of national testing, she said.

The National Testing Proposal

In early September, an Administration spokesperson could open the CRESST conference with a report that there was "growing support for the tests from a range of groups," including education, business, states and school districts. David Stevenson, special assistant to Deputy Secretary of Education Marshall Smith, presented the Administration's case for Smith, who was called away from the CRESST conference by the need to testify about the testing

proposal before a Senate subcommittee. Stevenson laid out the why and the how of the proposed national tests.

At the time of Stevenson's presentation to the CRESST conference, the tests' design looked like this: voluntary, no individually identifiable data to be given to the federal government, all items to be released after the tests are given, linked to NAEP and TIMSS, use a NAEP framework but with different items and test specifications, take 90 minutes with about half of the items machine scorable, required inclusion criteria and appropriate accommodations, results reported within the same school year on a metric easily understood by teachers and parents, federal funding (estimated to be $8 to $12 per student) for the first year of administration and perhaps afterwards, trial tests in spring of 1998, and ongoing research.[1]

Stevenson went into detail about the inclusion and LEP criteria because of controversy over these aspects. Students identified as being disabled should be included except in limited circumstances. LEP students should be included unless they have been in LEP programs less than three years or cannot demonstrate knowledge of reading in English even with accommodations permitted by the test. The test plans include a bilingual math version. The LEP restrictions, noted a participant later in the discussion, will exclude about 90 percent of language-minority students from the reading test. However, Stevenson said the Administration **views the test as a test** of reading in English, he said. Further, there is no way to accommodate more than 100 different languages spoken by students in schools.

Why did the Administration select fourth-grade reading and eighth-grade math as the targets for the tests? "There is ample evidence that students need to

[1] Since then, the development contract for the test has been turned over to the National Assessment Governing Board, and Congress has mandated several studies to be conducted by the National Academy of Sciences. These include one to determine whether an equivalency scale can be developed that would allow test scores from already available tests and NAEP to be compared with each other. Another will evaluate the technical quality, validity, reliability, design, and racial, cultural, or gender bias of test items already developed. A third will recommend appropriate safeguards to ensure that tests are not used in a discriminatory or inappropriate manner. NAS's recommendations will be taken up by Congress during reauthorization hearings on NAEP and NAGB this spring. No pilot tests will be given in fiscal year 1998. A House action in February 1998, requires Congressional approval for any national tests. In April, 1998, an amendment, including language identical to that in the House-passed bill, passed in the Senate. However, in conference, the Senate and House agreed to remove the measure against national testing.

be able to read by the end of third grade," Stevenson said. "Basically, fourth-grade reading scores have been flat for the last 12 years. Forty percent of students are not able to read at the basic level." Eighth-grade math scores are disappointing, he said, because the standards are not high enough. The math curriculum at this grade for many students is similar to the math curriculum in the seventh grade and lacks rigor, he said.

In presenting the reasons for development and administration of national tests, Stevenson talked mostly about the messages such tests would deliver. The Administration expects national tests to improve student achievement because they will:

- improve the odds of success for all students at two critical points where strategic interventions might improve student learning;

- focus national attention on the need to improve skills in reading and math by making "standards come alive" through exposing actual student work;

- develop a national consensus on the crucial nature of fourth-grade reading and eighth-grade math;

- use a national effort to energize local efforts to improve teaching and learning; and

- provide teachers, parents and students with accurate and reliable information about student performance measured against national and international standards.

Stevenson acknowledged that the testing proposal has its critics. Some call it a federal intrusion into local control of schools, some fault it for introducing a major change in policy without sufficient consideration, and others believe it will undermine local Goals 2000 standards work. It could lead to a national curriculum, some charge; ultimately, it will do nothing to improve student performance, say others.

However, Stevenson countered most of the criticisms. "We've had state leaders tell us the tests are not undermining but complementing what states are doing," he said. Furthermore, this proposal is not simply a test. It is a national effort to focus attention on basic skills in reading and math and, thus, will have a greater effect on student performance than other testing efforts, he said.

Finally, Stevenson described the present situation as "unfair to poor and minority students who have not had access to good instruction. We are pleased that urban districts have signed on because they need such tests to hold up high standards."

Stevenson noted that he had been involved in many policy issues, "but I am really impressed with the seriousness and care with which the department has pursued this initiative. It is easy to be skeptical about something that has not been done before, but we need you to push us to be clear, to ask the right questions."

Responses to the National Testing Proposal

State Response

The CRESST conference participants began immediately to ask very fundamental questions.

Wayne Martin, assessment expert with the Council of Chief State School Officers, discussed several issues that seem simple on the surface but become complicated when put into a political context.

The tests are to be voluntary, but does that mean voluntary at the state, district or private school levels? Who has the authority to sign a state or district up for the tests? At the state level, the answer to that question is very murky, Martin said, because "within the Beltway there is a view that all 50 states are the same, but they are totally unique and represent seven to eight different governance models." A governor, state board, state chief or some combination of the three may decide on participation in the test. But what happens if they are in disagreement? If a governor signs up his state, is the next governor honor-bound by the former governor's decision?

Use is another question said Martin. States that have agreed to participate intend to use the tests in several ways. Some with new assessments will use the national tests as comparisons. States in the process of developing new assessment systems, will look at integrating the national tests into their own efforts. In other states, the tests will simply replace existing tests.

Yet, only seven states as of the CRESST conference had approved of the idea. Martin said the rest are reluctant for a variety of reasons—some are strong local-

control states, some fear the tests will undermine efforts already underway, some worry about future costs and about too much testing in the fourth and eighth grades already.

A Practitioner's Response

From a teacher's viewpoint, the commitment to improving instruction and raising the bar through national tests is welcome, but "think of the realities of the classroom," warned Sharon Draper, National Teacher of the Year. So many students come to school today with "baggage" that must be addressed, from teaching children to use toothbrushes to compensating for reading skills they do not get at home. Also, teachers must deal with constant turnover among their students and unequal resources. And then, Draper said, they "must give up on teaching in March because they have to prepare students for tests." If students do not do well on tests, "the public will say teachers are not doing a good job. And that is not fair." The focus of tests and other efforts, she said, should be on what will help teaching and learning.

Draper used a sample test item that Stevenson had presented earlier to illustrate an issue about what the tests really measure. Some students might do well on the *Charlotte Web's* test question, said Draper, not because they were at a higher achievement level, but because they had read *Charlotte's Web* at home or their teacher had read it to them in the classroom. This raises a question about what previous experience kids have had that help them do well on the test, not whether they can read and comprehend what they are reading.

National PTA Response

The national testing proposal is based on "quantum leap" theories, not on research, contended Arnold Fege of the National PTA. "As I listened to the presentations this morning," he said, "I didn't hear about any research that backs up the introduction of national testing." In his opinion, "no parent in the country is losing sleep because his or her child is not meeting NAEP standards," and even though testing is pervasive in American education, it seems to not have made a big impact on change.

What's missing, according to Fege, is capacity building at state and local levels to understand what test results already reveal and take action, especially

on equity issues. "I think there is very little that can be learned from a national test that isn't already known about what needs to be done," he said.

Fege also criticized the process for decisionmaking about the test. Few parents have been involved so far, he pointed out. "How can 25 people on the NAGB sub-committee develop tests for millions of students without involving people at the local level?" That might take years, Fege said, "but the only way to get consensus on standards is through a democratic process involving teachers and parents." When they come together on teaching and learning, then better assessments and higher learning will follow, he added.

Response from a School District

Representing a large school district with a considerable language minority population, Ruben Carriedo, assessment director at San Diego, approached the national testing proposal with caution. It would have to meet his four criteria:

- the extent to which the national test purposes are consistent with reform efforts in San Diego, especially its assessment plan, which is already very comprehensive;

- the extent to which the national test is part of a standards-based system (most test makers say their assessments are standards-based, he pointed out, "but one is not always sure");

- the extent to which the national test is connected to local and state policy and content; and

- the extent to which the national test is planned in a timely fashion for successful administration in a large district.

"San Diego is open to the idea of national tests, but we need answers first," Carriedo said.

Response on Equity

Edmund Gordon, professor emeritus from Yale University, brought up different and very specific concerns that focused on equity. He decried the misdirected attention by everyone from the President on down—"a serious misuse of power," he said—on standards and testing when those are not the core problems in society. He insisted that standards have always been present in

American education, although "perhaps they have changed, and we need to be reminded of that." However, a new infrastructure is unnecessary.

Rather, "we need leadership to do something about the results," he said. Ever since the Coleman[2] report in 1965, "we have been challenged to uncouple academic achievement from social divisions into which people are born. Academic achievement should not be distinguishable by gender, race or other such factors." The widespread use in this century of tests with their implicit standards, Gordon pointed out, "seems to have had little impact on the nation's academic under productivity."

Even if tests were an appropriate means of increasing academic achievement, Gordon questioned whether they are as good as they should be. Developing tests is hard work, he said, and developing one that will match pedagogical strategies that come out of the cognitive sciences is going to be even harder. Instead of investing in more of what already is, the development money should be going into tests that support analysis, construction, imagination, reasoning and synthesis.

The power of the federal government, Gordon said, would be better spent on:

- improving the competence and productivity levels of those who teach: "We need to increase student achievement four-fold, and we can't do that without improving practice";

- reducing the gap between the quality of pedagogical knowledge and the quality of pedagogical practice and educational policy;

- studying the implications for education and assessment of the juxtaposition of great diversity in human characteristics and the increasing demand for pluralism;

- developing educational assessment technology that reflects modern conceptions of learning; and

- finding ways for parents and others to actively support academic growth of students: "We haven't been successful at having schools substitute for the support that families provide to children," he said.

---

[2] The Coleman study used statistical analyses to relate school characteristics to student achievement. *Equality of Educational Opportunity,* Government Printing Office, 1966.

Response from a Policy Viewpoint

The final panelist, Lorraine McDonnell of CRESST/University of California, Santa Barbara, drew five policy stories based on different views about the proposed national tests. The Clinton Administration, as outlined by David Stevenson, believes parents and the public will insist on standards being raised if they have the kind of information the tests will deliver.

Another scenario, that played by Rep. William Goodling (R-PA) who is chair of the House Education and Workforce Committee, presents the tests as unnecessary and as federal intrusion.

A third view looks at disparate policy impacts produced by the tests and other initiatives. If the Office for Civil Rights, for example, investigates the affirmative action pullback at the University of California, but the Administration pursues a dependence on testing results, the policies could be at odds, McDonnell said. A fourth story concerns the defense of local control. And the fifth story is about public attitudes toward the test, which have been high but decreasing in the most recent polls. Another "public" issue is the potential loss of public support of standards if the released test items and their emphasis upon NCTM standards, for example, become controversial. "If there are big debates over pedagogy, it would take political skill to get standards back on track," McDonnell said.

"It is clear from these stories," she added, "that the major policy issues related to the testing proposal are not about the tests per se but about fundamental issues that have animated debate on education reform for the past few decades." She ended the panel discussion by asking four questions that summed up the policy issues:

- How can top-down policy with limited incentives and limited sanctions change what goes on in classrooms?
- Do political leaders, educators and the public have the political will to afford all children an equal opportunity to learn?
- Whose values should define what is taught?
- Who should control the content and practices of American education?

Emphasizing that the testing proposal is a logical outgrowth of reforms begun in the 1980s, Joseph Conaty, Office of Educational Research and Improvement, closed the session by pointing out that improving students' learning is a worthy goal and "to have a President and a country talking about improvement of education for all children," is an important consequence of the current proposal.

Other Conference Sessions

While the political rhetoric around the national testing proposal appears to speak clearly about a simple idea, researchers who have been working on aspects of assessment policy and technical issues brought additional insights. Concurrent sessions presented a number of research studies that directly addressed issues about national tests.

Inclusion of Limited English Proficient Students

An issue that rippled through many discussions at the CRESST Conference was the inclusion of Limited English Proficient (LEP) students. Discussions ranged from strategies for accommodating LEP students to the comparability of tests in different languages.

Arguing that several federal statutes provide a legal rationale for inclusion of LEP students in the proposed national tests, Richard Duran of CRESST/ University of California, Santa Barbara contended that excluding LEP students would bias the results of large-scale assessments. Some districts in the Southwest, for example, enroll extremely large percentages of LEP students, most of whom would be excluded in the proposed national reading test.

Duran suggested that the national testing design could include a variety of accommodations—extended time/multiple testing sessions, one-on-one testing, small group testing, bilingual dictionaries, word lists and reading directions aloud in English. Research is still needed, he said, on allowing responses in a student's primary language and on the comparability of performance when accommodations are provided. Furthermore, he insisted that selection of LEP students for inclusion in the testing program, such as the cutoff dependent on number of years in a language program, "ought to be a locally informed decision."

So much of the data, commented Lorrie Shepard of CRESST/University of Colorado at Boulder, deals with counting students when "we actually know little about what it means to teach LEP kids." She described her CRESST project which is using LEP students' classroom work to identify performance benchmarks and comparing scores on classroom work with standard achievement results.

Shepard also noted that the conditions of the administration of tests are often unfair to some students. "We need to ask if such accommodations as paraphrasing and glossaries matter, and then we need to change assessments on the basis of what we learn."

It worries her that states do not indicate what their large-scale assessments are intended to do. "If they are clear on the purpose, then they don't need to test every child," she said. The national testing proposal's purpose is political, Shepard added, but it would be useful to think of several ways the data would encourage other questions. For example, data on LEP students might lead to asking who was schooled here or elsewhere in their native language.

John Olson of the American Institutes for Research summarized a report on accommodation issues for LEP students based on NAEP bilingual math and science assessments given in Puerto Rico. He agreed with Shepard and Duran that data are lacking on LEP students because they have not been included in testing programs and because of the lack of uniform definition of an LEP student.

Before NAEP became more focused on inclusion, he said, states were excluding more than one-half of LEP students from NAEP testing. Guidelines for including LEP students in testing programs have great variability, and accommodations are not likely. According to data from the 1996 NAEP tests, the most commonly used accommodations were to give LEP students extra time and to administer the test in smaller groups.

Jamal Abedi, CRESST/UCLA reported results from a study that analyzed the effects of different accommodations provided to LEP students on a test composed of NAEP test questions in mathematics. Abedi found that simply translating questions from English into Spanish did not significantly increase LEP student performance.

"It is clear," said Abedi, "that if students have not learned the math concepts in their native language, translation is of no benefit."

The researchers did find that LEP students benefited when the language of math questions was simplified. While all students' performances increased, the LEP students performance increased more than native English speakers, said Abedi.

Kris Waltman, CRESST/UCLA also found evidence of the complexity of translation issues in performance assessments being developed for the Los Angeles Unified School District. In interviews with teachers after pilot tests that even though instruction may be in Spanish, some key vocabulary words occur only in English. For example, such words as "Gold Rush" and "pioneer" were not understood by students in Spanish, but they did understand the terms in English.

Furthermore, the goals of native language instruction may differ from those of a standard English curriculum, resulting in differences in opportunity to learn similar language skills. Reflecting on students' writing performance, Waltman noted, "It appears that students in Bilingual classrooms are not taught academic written Spanish or, if so, the goal is not to learn Spanish but to make a transition into English," Waltman said. Her findings emphasize the need to provide *all* students the opportunity to develop the knowledge and skills upon which they are assessed.

Including Students with Disabilities

The national testing debate also illuminated the need to understand better the current state of assessment of students with disabilities. These students "serve as the canaries, the worry signals" that are needed "to help build the learning communities we need," commented Kris Guiterrez of UCLA, chair of a panel discussion on the impact of standards and assessment on special needs and language minority students. In addition, changes in federal and state policies will require most students with disabilities to be included in large-scale assessments.

Reporting on the details of a survey of current state practices regarding inclusion of students with disabilities in standards-based assessments, James Ysseldyke of the National Center on Educational Outcomes said the participation of such students in state-wide assessments has increased. Thirty-nine of the states have provisions for accommodation of students with disabilities in their assessment programs, but the accommodations are not consistent, he said. Also,

23 states specify what needs to be in assessment reports, but almost two-thirds of these specifically eliminate reporting on students with disabilities.

He concluded that states are unclear about policies on the issue of participation in assessments but predicted that as educators "learn how to deal with kids at the margins, we will create more diversity in assessment and instruction for other kids."

Studying the effects of Kentucky's provisions for inclusion of students with disabilities in the Kentucky Instructional Results and Instructional System, Daniel Koretz of CRESST/RAND found that Kentucky had successfully included almost all students in the statewide assessment (only 2 percent were excluded). However, a disproportionally high percentage of disabled students failed to try some of the mathematics questions or scored zero on them, suggesting that part of the assessment was too difficult for many students with disabilities.

Also, even though the state has stiff restrictions on accommodations, they were used more frequently than would be expected. Testing accommodations were to be consistent with individual educational plans and classroom teaching accommodations. Yet 57 percent of students with disabilities were accommodated through paraphrasing of test questions. Another issue was that the use of accommodations was very different across classrooms and schools, raising considerable doubt into the comparability of the assessment results. That the scores of students receiving some kinds of accommodations, e.g., dictation, appeared quite high also raises validity questions.

Koretz contended that researchers and policymakers need to be clearer about what accommodations are appropriate for specific disabilities and which really change the nature of what is being assessed. Research is needed to find out what accommodations work best, and more careful monitoring of results is needed, Koretz said.

Assessment and Classroom Practice

Another theme of the national debate about the testing proposal concerns the actual impact of assessment on classroom practice. CRESST research presented at the conference suggested that assessment reform produces important changes in instruction, but the results are heavily dependent on implementation.

Investigating assessment reforms in Kentucky, Hilda Borko of CRESST/University of Colorado at Boulder and Brian Stecher of CRESST/RAND, found that schools increased their use of reform-minded mathematics instruction while continuing to use traditional instructional approaches.

Schools with the highest test score gains on the Kentucky state assessments used both types of practices. Exemplary teachers, reported Borko, tended to use "more of everything," mixing reform-based and traditional instructional strategies.

One of the telling points is how well exemplary teachers weave it all together, she said. They typically use portfolios as culminating assessments within their ongoing instructional programs and have an instructional approach to making portfolios successfully work in the classroom. They tend to take leadership roles in the state's reforms, seeking to benefit their school rather than just their own teaching. These teachers were positive about the reforms but not reluctant to suggest how to improve them.

The road to change is a difficult one, as found in another CRESST project. Megan Franke of CRESST/UCLA, studying a group of California teachers engaged in mathematics reform, found that teachers were committed to change and that their assessment practices were indeed changing over time. Teachers were moving toward more use of open-ended questions and beginning to use rubrics to evaluate student performance; many found value in these assessments, but they also found such forms of assessment a challenge to use and integrate in their classroom routines. "Making sense of what students are learning from open-ended results is very confusing," she said. "Teachers need a lot of content knowledge to make good use of rubrics, and we need to look beyond whether teachers are saying they are using something and see how they actually are using it."

Of course, principals play key roles in changing assessment practices in schools. To gain an understanding of the views of principals, CRESST researcher Maryl Gearhart surveyed principals who had shown an interest in school improvement by attending events at the UCLA Principals' Center. Surveys were returned by 96 principals from 35 public school districts. Gearhart reported that these principals were generally committed to standards grounded in new views

of math education. With regard to testing, principals were likely to favor performance-based measures, although a large minority favored the use of both norm-referenced and performance tests, and some principals favored norm referenced tests. "This study identifies common ground among the views of forward-looking principals," Gearhart commented. "To embrace and honor the diversity of their views, states and districts need to develop standards that capture a breadth of knowledge and skill, and a standards-based assessment system that integrates multiple measures."

Scoring Assessments and Classroom Practice

One of the proven methods for jump-starting what teachers know and believe about instruction and assessment is to include them in the scoring of standards-based assessments. Using observations and surveys of about 200 teachers and interviews with a smaller sample, Beverly Falk of the National Center for Restructuring Education, Schools, and Teaching at Teachers College/Columbia University, found that scoring standards-based assessments was a very useful experience for teachers.

"The teachers felt that looking at student work with colleagues in relation to standards supported their learning about the standards, clarified their goals for student learning and deepened their understandings of the discipline," she said.

The teachers, who were scoring a state assessment, also gained insights about what students know and can do and began to see that students have a variety of ways of learning and thinking about learning. A number of the teachers said they intended to change their teaching, using more group work, making sure students understand concepts and studying how students arrive at their answers. Finally, the teachers became much more supportive of the standards process.

Falk cautioned, however, that these benefits will be evident only if the standards and assessments are rich and support meaningful learning. Providing opportunities for this kind of teacher work needs to be ongoing in schools, she said, and it should be focused on supporting students and learning, not on high-stakes consequences.

CRESST researcher Robert Land reported on the results of teachers scoring language arts tasks assessing student revision skills. Can useful revision tasks be developed? Can they be scored reliably and in a timely fashion?

Land said that he found that such tasks could be developed and scored with reasonable reliability. The scoring took no more time than any other scoring task once the scorers were familiar enough with the original draft. He also found a reasonable spread of student performance. What was most interesting, said Land, was how little change has occurred in how teachers teach writing strategies to students; they continue to focus more on the drafting process than prewriting, revising or editing.

Community Involvement in Standards-Based Reforms

Charlotte Higuchi of CRESST/Los Angeles Unified School District, also reported results from a language arts project, emphasizing the value of including parents and the community as well as teachers in standards-based school reform. Involved in the development of language arts standards and classroom assessments for LAUSD, Higuchi noted that the process produced great debates, including the phonics controversy (parents' insistence that phonics be included won). Another key issue was that parents at first did not understand teachers' attention to student "voice" in writing; they just wanted students to be able to spell. But when teachers explained they wanted to be able to pick up a student's paper "and know that this is Sara's or Jose's writing," parents caught on, Higuchi said.

"We learned there is a real hunger out there among teachers, parents, the community and administration" to know more about standards and assessments, according to Higuchi. "Teachers were exhilarated by the dialogue and reported that their teaching is now standards-based."

Systemic reform efforts also talk a lot about parent involvement, but the research has tended to focus on parent involvement practices and not on the actual effect of such practices. Denise Quigley of CRESST/UCLA is helping the Los Angeles Annenberg Metropolitan Project implement and evaluate parent involvement programs using a different approach to assessment. The project is working at 29 schools enrolling more than 31,000 students.

Parent involvement efforts usually provide education for parents but not for teachers on how to work with parents, she pointed out, but "teachers also need help, especially as parents become more engaged in school." The Annenberg effort is focusing on forging closer relations between teachers and families in support of student learning, providing a learning environment at home and at school, sharing knowledge to better understand student needs and abilities, and jointly holding high expectations for students' behaviors and achievement.

The focus is on students, she said, but the evaluation is looking at interim changes that happen before changes in student achievement. Parent involvement is an incremental process, she said, and it may take years to see an impact on students.

"We are concentrating our early efforts on documenting changes in parents and teachers because that has to come first before changes in students," Quigley said. Her project is triangulating information, matching what parents say about contacts with schools with teachers' logs on their parent contacts and with student achievement.

A key lesson from the classroom assessment and practice sessions is that although reform is driven by policy makers, its support comes from parents, teachers, and schools working together.

Evaluating Title I and Other Reform Programs

The 1994 Improving America's Schools Act supported schoolwide reform programs generally requiring new assessment systems to evaluate progress. But systemic reform is a moving target, challenging traditional ideas about evaluation, according to Elois Scott of the U.S. Department of Education. The flexibility given Title I schools to use funding for schoolwide reform, for example, makes it especially difficult to provide policymakers with comparable information about the efficacy of individual programs. However, OERI's panels on promising practices in reading and math have been able to identify research-based programs with significant impact on student learning. Furthermore, the researchers found that implementation was the key factor, she said. "It doesn't really matter what the program is, it will not work unless it is implemented in

the way intended." Thus, evaluations need to support the improvement of implementation quality, she said.

To look at how change occurs, researchers conducting a longitudinal study of systemic reform divided states into levels of reform activity. Within the states, they selected high- and low-reform districts. But what they discovered was how political processes can quickly change where a state falls on the reform continuum. Two years ago California would have been in the high-reform category, Scott noted, but not now. Similarly, Kentucky once would have been out in front, but it is going back to more standardized forms of testing, she said. These shifts make evaluation very difficult, but the systemic reform research is beginning to develop some successful evaluation processes.

Evaluators are using a range of methodologies to examine implementation issues and to document actual practice. "Examples," said Scott, "include collection of artifacts, use of focus groups and observations to investigate alignment of curriculum and instruction with challenging standards."

Scott also noted the difficulties of making sure that the data collected from districts and states are comparable. Title I researchers are administering a common instrument across the projects it is studying, but the law encourages states to develop their own evaluation instruments, which they are doing. "We are comparing our instrument to state ones so that we can legitimately make an appropriate comparison."

Linking Tests

Comparability—in this case, of results from different tests—was the burning issue at another conference session. President Clinton originally proposed that the voluntary national tests would be developed by a contractor and linked from NAEP and TIMSS tests, such a student performance on the national test could be used to estimate students' performance on NAEP and TIMSS. Congress, however, was interested in knowing whether a new assessment system was really necessary and whether or not existing tests could be used to provide such a link. To do so, existing commercial tests would need to be linked to NAEP and TIMSS, upon whose frameworks the national tests were to be based.

A feasibility study, presented at the CRESST conference, which attempted to link test results between TIMSS and NAEP, demonstrated just how difficult it

would be to link commercial tests to NAEP or TIMSS. Eugene Johnson, Educational Testing Service, Don McLaughlin American Institutes of Research, and Sharif Shakrani from the National Center for Education Statistics, discussed the study findings.

Johnson, who investigated procedures for linking TIMSS and NAEP, found that the linking worked acceptably at the eighth-grade level but not acceptably at the fourth-grade level. A requirement for an acceptable linking is that the two assessments measure the same content. Unless the two assessments are built to the same framework, it is unlikely that a linking is going to support many of the types of statements people would like to make, noted Johnson. Evaluating the goodness of the link was hampered by the fact that no student was administered both assessments. Consequently, it is impossible to assess the degree of correlation between scores on the two assessments.

Johnson noted that content analyses and validity studies supported the grade eight link for the purpose of comparison of state-level predicted TIMSS results with actual country-level TIMSS results but not for comparisons below the state level. The grade four link has proved much more problematic and is still undergoing review by NCES.

Don McLaughlin said that NAEP and TIMSS are generally sufficiently similar to warrant linkages for global comparisons "but not necessarily so for detailed comparisons of areas of student achievement or processes in classrooms." The studies essentially came to the same conclusions for both math and science.

McLaughlin concluded that the study raised more questions than it answered, such as how they should account for the use of calculators on NAEP but not on TIMSS. Or why students in Minnesota and Colorado outperformed much of the rest of the United States on the NAEP open response items, but not on TIMSS open response items.

Shariff Shakrani, whose office funded the AIR study, acknowledged the linking problems and said that there were policy concerns as well.

The study of linking NAEP and TIMSS "was supposed to be a feasibility study," Shakrani said, but some people wanted to use the study results to compare state NAEP scores to TIMSS, thereby comparing state performances to

countries such as Japan and Korea. But the validity of the linking study does not support such uses, said Shakrani.

Technology and Assessment Advances

One important strand of CRESST's recent research in problem-solving skills has focused on the use of measuring problem solving using a combination of measures. "Our problem-solving studies," summarized Harold O'Neil, Jr., of CRESST/USC, "attempt to measure what we believe to be the three most important components of problem solving," including:

1. Domain-specific knowledge (content understanding);
2. Problem-solving strategies specific to the domain;
3. Self-regulation, which includes metacognition and motivation.

The focus of the current CRESST problem solving studies has been on assessing students' understanding of how a cause-effect system works (e.g., mechanical systems such as pumps). A particularly novel approach to assessing domain-specific knowledge is the use of functional representations. This approach uses concept mapping to depict student understanding of information. For example, in a mechanical pump system, one component of the system is the handle. The concept handle is depicted as having two possible states—up or down. All components within the system are represented with such state information. Students are then asked to construct a knowledge map that explicitly relates the states among components (e.g., [pump handle DOWN]—causes—[cylinder pressure HIGH]). The power of this new approach is that the task measures not only if students know what a concept is, but more importantly for problem solving, how well the student understands how the concept or component *functions* within the system. We are using knowledge maps to measure procedure knowledge. Most prior applications of this technology measured declarative knowledge.

CRESST research suggests that measuring each component is feasible and "can be accomplished with reasonable validity and reliability in a short period of time and not a great deal of cost," said O'Neil. The studies are still working on the best methods for reporting the information.

Howard Herl, a member of the UCLA team investigating technology and assessment, described lessons learned from the implementation of computer-based concept mapping tasks. Herl discussed theoretical and applied aspects of online concept mapping and how the technology provides capabilities not practical in other formats. The design of concept mapping systems can be "open" or "closed." Open systems allow students to create their own map using their own concepts and links while closed systems provide students with a fixed set of concepts and links. Each approach has implications for scoring and system design. Open systems, typical of paper-pencil formats, are problematic for scoring. Because of the uniqueness of each student map, there is no standard against which to judge performance. On the other hand, open systems allow students to represent their understanding in a way that best fits their own conception of the content.

Closed systems, where concepts and links are provided to students, constrain students on what they can express but offer a large advantage in terms of scoring and automation. A fixed set of concepts and links means that all possible concept-link-concept propositions are known prior to the test. Thus, scoring is carried out by directly comparing student maps with expert maps. For assessment purposes, computer-based concept mapping can only be done using closed systems.

Herl also described results from a recent study using the CRESST Internet-based assessment system. "Our system effectively evaluates students' use and judgment of Web information sources," said Herl, "in real-time and within seconds. Our system provides students with the option to obtain feedback about the quality of their maps at any time during the task." Another benefit, said Herl, is that "students get excited about the visual representation and become engaged in the task, even though the task is difficult and open-ended."

According to CRESST researcher Greg Chung, the design and implementation of online performance assessments requires clear assessment goals and a shift from single-user to multiple-user system designs. Some successes of the CRESST Internet-based assessment system were real-time scoring of concept maps, platform independence using Java, and the measurement of searching behavior. A major shortcoming was slow performance while accessing the concept map database.

Chung stressed, "Despite the power and seemingly unlimited potential of online environments, the measurement issues remain unchanged—particularly with respect to validity and reliability. We still need to know, for example, the extent to which online systems can measure complex performance, and the extent to which online systems provide gains in terms of cost savings, measurement fidelity, engagement, or access."

CRESST/UCLA School of Medicine researcher Ron Stevens has applied computer technology to the assessment of high school and college students' problem solving skills. With a computerized data base as an information source, students try to unravel a genetics mystery, demonstrating various ways of using information. As they gain competence in their investigations, they learn to conduct thorough information searches first, then perform confirmatory tests, rather than ordering up multiple genetic tests first. The computer records student searchpath maps, showing strategies that students use and how they change over time, he said. The assessment improves their problem solving skills and helps to explain why some students miss problems. Although problem solving has different patterns in different subject areas, simulations such as the genetics assessment are applicable to other subjects, Stevens said.

Assessment in the Early Childhood Years

Interest in the use of assessment of young children remains high, although it has a history of misuse.[3] Many private schools still use assessment as the primary or single screening method for school entrance, despite questionable dependability of many measures. They have similarly been used to retain children in kindergarten or to place them into special programs despite the possible negative effects from such actions. The CRESST session on assessment of young children focused on whether or not there is a legitimate role for classroom-based instructional assessment in systems of accountability during the early childhood years. Samuel Meisels of the University of Michigan asked the above question to launch a panel discussion on the issue and immediately answered "yes"—under certain conditions.

Early learning assessments must support improved learning first and foremost, even as they may be designed to also serve accountability and public

[3] See Shepard, L. (1994). The challenges of assessing young children appropriately. *Phi Delta Kappan, 76(3)*, pp. 206-213.

reporting services, said Meisels. To serve such purposes, such an instructional assessment requires standardized administration, reliable application of evaluative criteria, evidence that it is a valid indicator of learning, and the potential for aggregation for public reporting.

The Work Sampling System Meisels developed fits these criteria, he says. For pre-school through grade five, it is being used with 200,000 children in 8,000 classrooms. Research data have been published so far only for kindergarten, but data for K-3 are about to be released. The System focuses on individual students, covers all aspects of a curriculum, is aligned with national standards, and requires teacher training. Teachers, children, and the family are involved in the assessment.

Meisels presented preliminary data from one school district that showed students in the Work Sampling System demonstration had an average growth in reading on the Iowa Test of Basic Skills from one year to the next of 22 percent, compared to a district average increase of 10 percent.

UCLA researcher Carollee Howes described several other assessments appropriate for very young children, all using extensive observations. These include the Attachment Q-Sort, a rating of the child-teacher relationship. A 20-minute Snapshot Observation of children's play activities provides information related to the complexity of cognitive activity and complexity of peer interaction and adult-child interaction.

There is still a gap however in our knowledge about appropriate assessments of early literacy achievement, said P. David Pearson of Michigan State University.

"We have paid a price for directing most of our reading assessments at the third- and fourth-grade reading level," Pearson said.

That's about the age where most experts agree that assessments are valid, but by then it's often too late to help those children who need it the most. An additional problem is the general dissatisfaction with standardized tests used to assess early reading skills.

"We have lots of good psychometric data on a whole range of assessments that are not highly regarded by teachers and not used by them," said Pearson. In interviews with 100 teachers, he could find none who liked standardized tests in

K-1 grades. Furthermore, teachers are justified in being suspicious of test items that, in the name of maintaining the scoring ease and reliability of the multiple-choice tradition, no longer engage students in the cognitive behaviors they were designed to measure. Teachers think these tests are just plain silly. On the other hand, because so many classroom-based assessments used by teachers have not been subjected to the scrutiny of reliability and validity analyses, we have no idea about the quality and appropriateness of the information they provide.

To respond to these problems, the U.S. Department of Education has funded a new reading research center (Center for the Improvement of Early Reading Achievement) which will study current early literacy assessment practices and create and validate new early literacy assessment tools.

Michael Nettles of the University of Michigan, suggested that we need more information than we get from just assessments. Author of the three-volume *African-American Data Book*, Nettles said that over one-half of young black children are in some form of pre-school every day (30 percent are in Head Start programs) and begin kindergarten excited about going to school. But by the fourth grade, more than two-thirds are below basic in practically every subject. Prior to the fourth grade, the only national measure of achievement is a national longitudinal survey that provides data on the vocabulary of 4- to 5-year-olds.

"We don't know what these kids' day looks like," said Nettles nor "what instructional strategies teachers use." No information is available on the pre-school environment for African-American students, such as the training of teachers or the curriculum, and there is a dearth of data about what happens in Head Start programs, he said.

The public, parents, and policymakers need reliable information, Nettles insisted, in order to be able to act on problems. This may become available through the new NCES longitudinal study of young children, which will follow them from early childhood through fifth grade.

The subject of early assessments is urgent, according to Beverly Falk of Teachers College because districts are using test results now to make decisions about special education placement, even though the test results are questionable. She cited studies by ETS and Teachers College on the reliability of tests in New York City that found only a 50 percent correlation between the tests and scores on books verified as on the children's reading level. The studies compared scores

children received on reading tests in the third grade and their scores on tests based on books verified as readable. The biggest discrepancy was among children in the lowest quartile on the tests, who were scoring higher on the texts than test scores showed, she said. Nettles said his data show that individualized standardized tests are used often for decisions about kids getting into and out of special education, but research is lacking on whether these assessments are valid for such purposes.

According to Lorrie Shepard, an upcoming guide for state policymakers on early childhood assessment will help them understand that the same instrument cannot be used for classroom assessments, special education placements or other high-stakes decisions.

Q&A for Technical Issues Expert Panel

Billed as a discussion of technical questions about assessment systems, a panel of experts once again underscored the importance of the political and value issues raised by the national test and that the development of the test will require political and value decisions as much as technical ones.

Answering questions presented by panel chair Ed Reidy, one of the architects of the Kentucky assessment system, and others, the panel ranged over a number of topics, including linking, norming, and reporting.

Linking: Can We Compare Results from Different Tests?

The national tests bring to fore any number of comparison issues. While responding to parents' and public's apparent desire to know how their children stack up against the common standard defined by the national tests, the national test proposal also raises a number of thorny technical issues, e.g., How can results from year to year be compared if the tests are released and thus have totally new items annually? How can results from the national test, which occur at only one grade level per subject area be integrated with results at other grades levels, locally, and/or at the state and national level, so that local communities can judge what progress their children and schools are making? How can results from the national test be used to infer performance relative to an international standard, such as TIMSS?"

Linking," said H. D. Hoover, "is a new term we use to describe our ability to create comparable scores from different tests. Equating is one of the methods that can be used for linking test results," he added, "which would require all tasks to be fundamentally normed."

While policy people are trying to find ways to make different tests the same, differences in the content of tests have important implications for how students perform, as the NAEP/TIMSS linking session had pointed out, suggested Dan Koretz. Just changing the content mix of the TIMSS items, for example changed international rankings, sometimes substantially. The same problem is true of state NAEP ranking.

Regardless of the technical appropriateness, the proliferation of state and national standards has created pressure to link assessments, observed Rich Shavelson. People want those kinds of broad comparisons that will let them know how their children stack up to other children in the state, or nationally, not just at the local level. Given this public demand, someone is going to link the tests to provide them answers.

"I think we have made mistakes in considering many of our questions technical," said Lee Cronbach from Stanford University. In California, the legislature suggested that many different tests could be linked, without consulting the psychometric community. I think the consensus was that what the California legislature said would be done, couldn't be done.

Norming

"Given the nature of tests and limited number of tasks, the only way to equate from year to year is for all tasks to be fundamentally normed," said H. D. Hoover of the University of Iowa. That means they will be equated on a nationally representative sample. "The biggest technical problem is how are you going to convince enough schools, a representative sample across the United States, to participate in the norming process," he said.

Hoover also pointed out that in Iowa the rank order of students is virtually identical on the ITBS and NAEP, and nationally, the percent of students below the Iowa median on the ITBS and on NAEP is about the same. Other conference participants noted, however, that the results are not similar in other states, such

as Massachusetts, where the ITBS and NAEP showed significantly different results.

"No matter how far we go," said Koretz, "we will come back to norms because it's an essential part of making sense of student performance."

The national tests, explained Hoover, fail to take advantage of two major benefits of norms. Because the tests are given only once, you won't know if the individual student is getting better or worse from one year to the next. Second, the national tests won't detect differences in student skills, such as math computation, spelling, or reading comprehension.

Reporting

At present, commented Koretz, "it is abundantly clear that no one understands test scores," not even after efforts to educate people about achievement levels. "What you see over and over again is an enormous simplification of complicated data.... If the purpose of the national tests is to get people to judge their schools on national and international standards, then we have to decide what kinds of information will be useful and not misleading. Rankings are not helpful. It would be more useful to have descriptive data, but that is not the way the national tests are going."

If the voluntary national tests are reported using achievement levels such as used in NAEP, then they may become known as the "bad news tests," said Ed Reidy, because the NAEP achievement levels continually show that kids aren't doing well. The National Assessment Governing Board (NAGB) ignored test researchers who recommended against the use of the NAEP achievement levels for many reasons, including that even experts had difficulty in deciding which items constituted a basic, proficient, or advanced level.

For Hoover, NAEP's use of achievement levels has never been convincing. "What is adequate or proficient varies from state to state and from one locality to another. It includes a little bit of Wizard of Oz."

"Judgment plays a role in everything we do," said Rich Shavelson. "Setting standards is extraordinarily complex, and we really haven't given it sufficient attention. We are setting performance standards that are highly inadequate, and .....we have not come to grips with the judgment process around indicators of what is good enough. "

"We're talking about communication systems (grade levels, standards, etc.)," said Cronbach. "If we send messages that are too complex, we are not going to communicate. There is no way to say that because we think complexly, the policymakers in state capitols must think complexly, too." Yet we must be concerned with the messages we are sending, with the bottom line being "How is it [this test or assessment] affecting the educational system?"

National Tests: Specific Issues from Working Groups

Conference working group sessions considered the impact of the proposed national tests from several viewpoints, such as special issues for at-risk and language-minority children, reporting to parents, linking assessments, effects on classroom and school assessment initiatives and the integration of national assessments with state assessment systems.

Each group considered the benefits and the concerns about the national tests, relative to their particular subject, as well as the key research and technical questions.

Despite an individual focus for each working group, the discussions tended to be similar. The participants generally groped to find a strong rationale for the tests. It was acknowledged that the national tests could provide a common yardstick for discussions about instruction and learning and help raise teachers' expectations of students. Some states without a statewide assessment might be able to piggyback on the national tests to save time and money.

One group redesigned the national testing proposal, wanting to add a national standard for the quantity and quality of school resources as well as for student performance. The test could then serve a policy purpose, the participants said, that would help equalize resources.

The working groups spent more time on their concerns about the test. Central to most of the discussions was finding an appropriate purpose for the tests. As described, the test is too complicated to guide instruction and overpromises. As with other large-scale assessments removed from teacher decisionmaking, participants felt the results would not substantially impact teacher practice. Other objections included the loss of a sense of ownership of the assessment process by the states, overkill on assessments in schools already, the lack of attention to equity and minority-language concerns, the potential for

biased results because of teaching to the test, and the strong possibility that the public will not be interested in or get much out of the testing program.

The research questions pinpointed by the working groups included:

- the consequences of excluding LEP students on both students and test results;
- the dangers of over-generalization of data or misuse of data because of the large data base created by the tests;
- validity issues related to accommodations;
- how to measure adequate yearly progress, as required by Title I, using the national tests;
- what the tests can reveal about best practices and exemplary schools;
- how various publics use and interpret the test data;
- what meaning to draw about the other grades from testing at the fourth and eighth grades; and
- the relationship between student achievement level on the national tests and resources spent on staff development.

As one group summarized the discussion: "The new national tests should fit into what already exists rather than adding an additional load to the current testing situation.... The tests should include as many students as possible and should provide instructional feedback that can be addressed in teacher training programs as well as provide an indication of which schools and districts are in need of educational resources. If the vision for the national tests can be made more specific and address some of these issues, many constituents may feel more comfortable with the move towards national tests and standards."

Making Standards and Assessment Work Through Curriculum and Instruction

Ultimately, national tests, or any assessment system for that matter, are intended to influence classroom practice. When the whole issue of national goals and assessments began to develop more than eight years ago, some thought they were needed as an "alarm bell" and a motivator, but there was another

group which believed their major purpose would be to set targets for what was worth teaching and learning.

Lauren Resnick was a leader of the latter group, and for the opening of the last discussion of the CRESST conference she presented new research findings to support that idea. She calls her approach effort-based education, an antidote to the standard operating procedure in schools that considers aptitudes for learning to be inborn or acquired very early in life. In an effort-based system, students actually create knowledge and an ability to learn. It is created when:

- students know what is expected;
- evaluations are fair and credible (you can study for them, know what is coming up and organize instruction around them);
- there is a real recognition of accomplishment;
- the results are fixed,
- the time needed to accomplish them varies; and
- there is expert instruction all the time in every subject.

Under this idea, grading is based on absolute standards, not on a cut score. Curriculum and assessments are aligned, public accountability and instructional assessment are aligned. And good professional development becomes the heartland of reform.

Resnick presented early results from the California Math Renaissance professional development program, whereby middle school teachers receive direct assistance to improve math instruction. As measured by New Standards math reference exams, students in the program, though somewhat more at risk, outperformed students in a control group. Fifty-five percent of Math Renaissance students scored above average compared to 40 percent in the latter group. The results indicate "we are able to bring kids up from the basement," Resnick said. Similar results are coming in from other studies she cited.

"The point of this early evidence," she said, "is that the idea of very clear standards, assessments referenced to professional development based on those standards and a driving commitment to kids can start to make a difference with some of the lowest achieving students."

The opposite of Resnick's experience with the Renaissance Math teachers—a lack of investment in teaching—characterizes American education for at least the past 15 years, noted James Stigler of UCLA, who directed the videotape study of teaching and classroom learning practices for the Third International Mathematics and Science Study (TIMSS).

The videotape survey in which American, Japanese, and German teachers participated, showed that American teachers rarely change their teaching methods despite widespread attempts at educational reform. For example, 70% of the teachers viewing videotapes of recommended changes in teaching math said they were implementing the reforms in the tape. However, "when we looked at their teaching, we found superficial changes, sometimes for the worse," he said.

This lack of impact on teaching is not surprising, noted Stigler, because, oddly, "teaching has not been a prominent part of the reforms." However, he believes things are changing.

One of the values of the videotape survey, he said, is that it can introduce some good examples of various modes of teaching into the culture of American schools. Teachers can see what different kinds of teaching really look like, in contrast to the out of context training teachers in this country generally receive. Everything known about learning would suggest that such decontexualized training does not work, he said. What is needed is a system that will enable teachers to share "good stuff"—not a reform system, he said, "but a system that supports gradual and continuous improvement of teaching and learning."

Philadelphia's school reform similarly has been accused of failing to make a significant difference in teaching and learning, according to Warren Simmons from the Philadelphia Education Fund. When Simmons arrived in Philadelphia to join its reform effort, a union official summed up the general attitude—"we've been doing reform for 30 years, and it doesn't work." One of the problems with school reform said Simmons, is that decisions are oftentimes based on politics, not on research or empirical data.

The most recent Philadelphia reform project is Children Achieving, Philadelphia's response to the Annenberg Challenge, which is partially supported by the Philadelphia Education Fund. One of the positive outcomes of Children Achieving, said Simmons, is that parents and teachers are now asking the right questions about school reform that get to the heart of teaching and

learning issues, such as: How well are assessments and standards measuring the strengths and weaknesses of students? What should the curriculum be? How do I know textbooks and materials are aligned with the Philadelphia standards? How do I know staff development is aligned with the standards?

What is discouraging, said Simmons, "is that we do not have answers to those questions." The District attempted to help teachers develop units of study, for example, but the units were of low quality, and district officials realized they had to provide teachers with a lot more guidance on what a standards-based classroom looks like.

"We are going to spend the year talking explicitly about the theory of learning underlying standards," he said, "and make a stronger connection between standards and assessments and what changes they require in pedagogy." Teachers are talking about student work and assessment tasks and scoring rubrics, "but we have to surround that with a broader conversation about what it means to have standards-based teaching."

While the immediacy of radical changes in assessment policy confronted the CRESST conference, it ended with a less contentious, and extremely thoughtful adventure into a different future.

Reading a "letter" that CRESST Co-directors Eva Baker and Robert Linn might receive, Robert Glaser of CRESST/University of Pittsburgh described a future world in which standards and assessment are part of the job of curriculum and instruction. There are two overlapping classes of standards. One concerns general competence and knowledge for participating in society, and the other are standards that are a matter of regional expertise, reflecting local excellence and specializations communities adopt for cultural and regional well being. Communities develop pride from the particular human capital that their educational goals provide. Glaser said about this future scenario, that "standards are legal entities." "That became necessary," he explained, "after below-standards protest marches were held around the country."

The use of cutoff scores for levels of performance have been deemphasized, and assessments instead are used to survey possibilities for student growth rather than to designate students who are not ready for certain academic opportunities. The analysis of item difficulty requires attention to the characteristic of the assessment situation and the nature of performance before an item is designed.

Basic skills are integrated with complex levels of performance, and much of learning is under the control of the learner.

In this assessment future, there are three conditions:

- assessments are socially situated—students contribute to and assist others, they develop and question their definitions of competence;
- competence is displayed using advanced technology so students and parents can see performance; and
- the assessments have cognitive significance, covering content but not ignoring cognitive competencies such as asking questions and planning prior to problem solving.

Concluding Remarks

This year's CRESST conference, Eva Baker said in summary, shows that the same technical and policy issues resurface in any assessment system; problems of validity, balance, and political support. In the case of the proposed national tests, the conference produced a consensus that while the tests pose challenges of technical merit, the debate has more to do with politics and who will win an exercise in power. For CRESST's work, the issues brought out by the debate over national tests implied that assessment researchers "must be sure we are not just talking to each other, but finding a way for our community to have a voice and a strategy for communicating what we think is important in the policy arena."