2000 CRESST Conference Proceedings

Educational Accountability in the 21st Century

CSE Technical Report 549

Anne Lewis

October 2001

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 in part by the Educational Research and Development Centers Program, PR Award Number R305B960002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

2000 CRESST CONFERENCE PROCEEDINGS

EDUCATIONAL ACCOUNTABILITY IN THE 21ST CENTURY[1]

Anne Lewis

As a mentor rather than advocate, CRESST provided an opportunity at its 2000 annual conference for more than 200 practitioners and researchers to debate and to develop consensus about how to build a quality K–12 accountability system. Part of the September 14–15 conference focused on changing the design in the middle of the construction process.

Participants listened to competing approaches, heard about the most current research on aspects of accountability, weighed the mistakes that have been made, and concluded that accountability also extends to policymakers at state and national levels.

The conference opener—a panel discussion about the Texas accountability system with a presentation by one of its leaders and a response from researchers on some of the pitfalls of systems like Texas'—reflected the issues to be debated for the next two days. The discussion, "Improving State and District Assessment and Accountability Systems," presented a basic tension in accountability, said the panel chair, Joy McLarty, deputy superintendent of the Austin (Texas) Independent School District. The panelists, she noted, had competing models—"do no harm or do the greatest good for the greatest number."

The Texas accountability system, according to John Stevens, of the Texas Business and Education Coalition (TBEC), resulted from a "stealth" campaign that attracted relatively little attention when compared with more comprehensive, one-time reform efforts like that in Kentucky. The essential components were put into place incrementally over several years, and the national media, for the most part,

[1] Although we have made every reasonable attempt to include all CRESST conference presentations, a few may not be mentioned here due to technical recording problems. Our apologies to those authors whom we may have missed. All presenters are asked to share their overheads or other conference handouts with us after the conference. Please see the CRESST Web site, http://www.cse.ucla.edu for additional information about this conference and other CRESST conferences.

did not take notice until Governor George W. Bush began his campaign for the presidency.

The current Texas reform movement began in the mid-1980s as an effort of business leaders led by Ross Perot. The group has continued to support education reform to this day. The TBEC Board of Directors includes about 60 business and education leaders who work together to develop recommendations for state education policy and who support community-based school improvement efforts. Another important element of reform was a landmark Texas Supreme Court decision that required the state to equalize financial resources available to school districts.

The elements of the state's accountability system, Stevens said, are student learning standards, statewide student assessment, results-based accountability, decentralized decision making, and feedback systems. Beginning in 1986, the Essential Elements laid out the material teachers were expected to present in their classrooms. These rather general standards were replaced in 1997 by new student learning standards, called the Texas Essential Knowledge and Skills (TEKS), that are considerably more precise and rigorous. At the same time the TEKS were adopted, the State Board of Education established a schedule through the year 2020 to review the standards one subject at a time. School district data collection began in 1986, with additional requirements in 1991. Statewide testing began in 1980. The current test, the Texas Assessment of Academic Skills (TAAS), is a criterion-referenced test created in 1990 to be aligned with the student learning standards. A new generation of tests aligned with the TEKS version will be administered beginning in 2003. The state also is requiring all districts to adopt an instrument for evaluating reading development for students in kindergarten, and first and second grades. In addition, Texas is implementing a Spanish-language TAAS and a Reading Proficiency Test in English for non-English-speaking students.

At the heart of the Texas accountability system, Stevens said, is the annual rating of schools as Exemplary, Recognized, Acceptable or Low Performing. The rating is based on the lowest group performance in reading, writing, math, attendance, and dropping out of school. A series of interventions is required for schools that are persistently rated Low Performing. Though schools can be closed by the state, "that has never happened because, in most cases, local school districts have taken action," he said. The accountability system initially required that 25% of the students pass the tests for a school to earn the Acceptable rating. The Acceptable

standard was raised 5% a year and is now at 50%. The Recognized rating initially required a 70% passing rate, and that has been raised to 80%. The Exemplary rating requires a 90% passing rate. To make the system more inclusive the state began including the results for students with disabilities two years ago and for ESL students in 2000–01.

"As cohorts of students have passed through the public schools" Stevens noted, "more are able to meet the standards each year, and the achievement gaps are closing." In math, for example, only 40% of minority students received passing scores when the current system was put in place in 1994; 80% are now passing, he said. Texas has ended social promotion beginning with third grade students in 2003, it has provided resources to schools to help students who are failing, and it has funded a massive professional development effort in reading with 14,000 of 17,000 kindergarten teachers participating in a week-long institute on reading in 1999. Similar training will be provided to first-grade teachers in 2000 and to second-grade teachers in 2001. Some school districts are analyzing test results by classroom and selecting successful teachers to mentor others, provide in-service training, and develop curriculum.

Still to be addressed are the number of dropouts, increasing the percentage of students at the top end, and better performance in math. The accountability system, however, has increased informed decision making from the statehouse to the schoolhouse, Stevens said. More and more people are using information generated by this system to help shape their decisions about how best to move the public schools forward. "The system is not perfect," he concluded, "but I really don't think another approach anywhere can document these kind of positive results for kids."

The "Texas miracle" may not be all it seems, countered Dan Koretz of CRESST and RAND. Texas' efforts are part of a national wave of testing and accountability that began in the 1970s and 1980s and that has gradually ratcheted up the stakes for educators (with rewards and sanctions) and for students (with exit exams and promotional gates). On the plus side, the newer accountability systems focus on outcomes rather than inputs, take a commonsense approach to incentives, can be powerful influences, and are cheap to do.

On the other hand, Koretz said, test-based accountability systems provide incentives to raise scores *per se*, resulting in a narrowing of the curriculum, cheating, and inflated scores and gains. Often, erroneous inferences are made about the effects

of an accountability system, and such systems also involve very high rates of misclassification of students.

Taking apart the Texas results, Koretz predicted they would follow the same pattern of score inflation that has been seen in Kentucky, where the focus of instruction is so riveted to the content of the test that scores inevitably go up. Even more disturbing, he said, was evidence from 1994 to 1998 of differential score inflation among the lowest performing students. At the same time that the Black-White gap on TAAS was decreasing substantially, he said, the gap was growing slightly on state-level administrations of the National Assessment of Educational Progress, "which confirms our suspicions of differential test preparation for TAAS." When people are encouraged to pay attention to the particular curriculum of a test, "scores become seriously inflated," he said.

Acknowledging that the pressure for greater accountability will persist, and testing will remain central to it, Koretz said accountability systems need to be developed that capture more of the benefit and less of the deleterious effects. The research community, he said, "has been walking behind an elephant with a broom, and we need to get out in front and help create more effective accountability systems."

To lessen score inflation, research is needed to address modifying testing programs and accountability systems, and to undertake major efforts to generate systems that are supported by evidence. He outlined several steps:

- Standards must span the domain well, be specific, and be clear. Teachers have to know what they are expected to teach students outside of a test. "If you look at state standards, they are anything but clear in many instances," Koretz said.

- Tests must do a reasonably good job of reflecting the standards. Right now, he said, "we don't look at the degree to which tests meet standards, which becomes evident if you take a test and try to envision the standards on which it is based."

- The standards and the domain should be clear from the test. "We must convey to people that what is going to be evaluated is what they are teaching, which means higher professional development costs."

An effective test-based accountability model ought to improve both information and incentives, Koretz said. It would use tests as a trigger for additional evaluation, focus directly on changing incentives so that "teachers do what you

want them to," use a broader set of outcome measures, include direct measures of teaching practice, and directly measure resources and opportunities.

A more complex accountability model would set realistic goals, with the targets for rate of gain based on research. "States are setting goals that are unrelated to what students actually do, and as a result, performance increases that are required are far beyond what teachers can do," he explained. The targets should be based on starting points. Replying to the criticism that such a strategy would condemn students to poor performing schools for a long time, Koretz noted that there is "enormous variability" in student performance even after all the inequities between groups are accounted for. Educators must be prepared "to do very hard work to bring all students in low-performing schools up but realize those schools will be lower scoring for a long time." The variation in performance needs to be acknowledged in the design of the test, in reporting of results, and in rewards and sanctions.

Programs should never be evaluated with simple observed score gains, Koretz advised. There should be routine validation of gains through audit testing, other validation studies, and the monitoring of test preparation and administration.

A research agenda for accountability would include assessment design focused on accountability, methods for validation of gains, value-added modeling, mechanisms of score inflation, multiple-measure systems that include proximal measures (find out which kind of system works best), systems for evaluations of teacher practice, and effects of incentives.

Drawing primarily from long-term research on the Kentucky accountability system, another member of the panel, Brian Stecher of CRESST and RAND, began by describing the "logic" of standards-based accountability. Standards drive policy at the district and classroom levels, which influences curriculum content and instruction. These lead to student outcomes, some of which can be measured with tests. The accountability system attaches incentives such as financial rewards and public reporting of results. This model does change behavior, he admitted, and can lead to some positive results.

However, it also has major shortcomings. In addition to the score inflation discussed by Koretz, there is evidence of undesirable changes in practice, Stecher said, not only in Kentucky but also in other states included in his studies (Vermont, Maryland, and Washington). Generally, the measures being used are imperfect and "corruptible," practice is driven by test scores more than by standards, and local

strategies in response to the results can be narrowly focused or even counter-productive.

Specifically, Stecher said, teaching practices and the curriculum changed in the tested grades more than in the nontested grades, class time was reallocated away from nontested subjects, and instruction became focused in narrow ways. These findings were based on information from both principals and teachers.

The research shows that tests often overshadow standards. "What's driving instruction is not standards but the tests or scoring rubrics attached to the tests, because these are the closest thing to day-to-day instruction."

Stecher proposed two strategies to counteract the negative results of test-based accountability. First, broaden the content representation in tests by strategies such as varying the subjects covered, changing grade levels over time, and using matrix sampling within tests. He also suggested that districts could use classroom-based assessments, informing teachers about when to use them, how to use them, and ways to make the information public. Writing prompts, for example, could be used at certain times of the year and the results shared with parents and teachers in a nonpunitive way while at the same time giving districts some evaluation data during the year.

The other strategy would be to reward practices as well as performance. Accountability systems, Stecher said, should be moving to incorporate measures of appropriate practice. For this to be done, it would be necessary “to create more elaborate standards about what ought to be seen in classrooms” and to create longitudinal student information systems.

In the discussion that followed formal presentations, Koretz was asked about the value-added model, as used in Tennessee. It assumes an accumulated curriculum, he replied, but different groups of students have different trends over the summer, “which can distort scores unless you test both in the fall and in the spring.”

In Texas, Stevens said, the Texas Education Agency also reports annually a value-added measure called “Comparable Improvement.” Every school, he said, is grouped with 40 other schools that serve similar student populations. Schools are ranked within their group on the basis of year-to-year gains in math and reading. This measure is useful in identifying the schools that are serving various student populations most effectively.

Quality Standards for Accountability Systems

Informed by advocates and researchers on the successes and shortcomings of several state accountability systems, the conference participants turned to consideration of what quality standards might be.

The previous panel reinforced the perception that there is a gap between intentions and reality, CRESST Co-Director Eva Baker noted as she opened a second panel focused on "Standards for Educational Accountability." CRESST is more interested in exploring what should be done up front than in discussing what should not have been done, she said, and will focus on supporting the validity of existing and emerging accountability systems.

Baker listed three improvement strategies that could be used: ad hoc analysis; creation of models, taking some of what Stevens, Koretz, and Stecher recommended; and developing and promulgating standards for educational accountability. There is a partial precedent for the last strategy, she said, referring to the *Standards for Educational and Psychological Testing* that the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education agreed to and published in 1999. These standards represent a consensus of the field; they are referred to in statutes, regulations, and case law; they guide test developers; they include requirements for users, test takers, and those with authority to mandate the tests; and they are intended primarily for technical and legal audiences.

"We are proposing standards that are similar but also different," Baker said. They will represent a minimum set for guidance, reflection, and review of accountability systems; they will be intended for a public audience, including legislators and the news media; they will reflect research or best practices; and they will be used as endorsements rather than a consensus.

The principles guiding the development of the standards will include (a) a developmental perspective that allows modifications as systems evolve, (b) the use of system information that accurately represents the state of affairs, (c) the use of indicators that are substantially under the control of the accountable institution or personnel, and (d) fairness to all parties.

Baker listed suggested descriptors: system validity; measurement concerns; accuracy and technical quality; special needs populations; incentives, sanctions, and stakes; and evaluating effects. She also presented examples of proposed standards:

- Information is given for planned purposes intended for a test or indicator (validity).
- Tests should minimize factors irrelevant to the domain assessed; for example, language complexity with a science knowledge test ought to not put special populations at a disadvantage.
- High-stakes decisions should not be made using results of only one measure.
- A test (or system) used to judge the growth of individuals, programs, or institutions should show evidence that change is the result of educational interventions (e.g., instruction).
- When high-stakes decisions use multiple measures, the method of weighting all measures must be made public.
- Studies evaluating effects should be conducted to determine whether the system supports high-quality instruction, promotes student access to education, minimizes corruption, affects teacher quality, and produces unanticipated outcomes.

Where is CRESST now in the development of standards? It is working toward agreement on what the characteristics of standards should be, Baker reported, including language that is appropriate for the audience, is commonly understood but not overstated, and is parsimonious—"the basic things we think everyone can believe in" and give high priority to developing. "We hope the standards will provide guidance to the Department of Education as it reviews Title I assessment plans," she said, "and be available for ESEA reauthorization." CRESST is expanding opportunities for policymakers and educators to talk about standards, utilizing Web contacts as well as continuous evaluation of accountability efforts.

As part of its work under a contract with the Office of Educational Research and Improvement, CRESST hopes to have a preliminary set of minimal standards by December, Baker said. An endorsement period will take place next spring, with a set of standards ready for use by districts and states and at the federal level within a year. Use of the standards will be evaluated before a second phase is implemented.

The standards are being drawn from experience with previous assessment and accountability systems, both successes and those systems with unintended negative

effects, explained Robert Linn, co-director of CRESST, who shared the panel discussion with Baker. They also rely on the *Standards for Educational and Psychological Testing* and on CRESST's model criteria of validity, reliability, and utility.

Linn discussed several issues that have developed from test-based accountability systems. One has to do with overconfidence in the precision and accuracy of numerical indices. "We have to get the notion out that there is a margin of error to be considered," he said. Usually, the treatment of precision/accuracy issues stops with reports of reliability of student test scores, but even the new test standards provide only limited, indirect guidance relevant to accountability systems.

Discussing specific aspects of the standards, Linn noted that standard errors are more relevant than reliability coefficients. For example, when a test is used to make categorical decisions, estimates for the test taker or group of test takers should be provided that demonstrate "the likelihood that a student tested today would have the same classification on the test tomorrow." Standard errors should be used to obtain precision of composite indices for schools, Linn said, and include sources of error due to measurement of the component parts, the interrelationships of those parts, and the sampling variability due to students.

Another technical standard refers to classifying schools. When summaries of student assessment results or other school composite indices are used, estimates of the consistency of the classifications for replications with different samples of students and different forms of the assessment should be provided. Linn also covered technical standards for indices of progress; validity of accountability system results; school accountability indices; and use of tests, including intended and unintended consequences resulting from the use of the tests. Evidence regarding such effects might include changes in instruction, such as time devoted to different content areas, teacher awareness and use of content standards, time and materials used for test preparation, and changes in student results. Those changes would be reflected in assessment results, on other indicators of achievement, and in dropout rates.

The 1999 testing standards, Linn pointed out, specifically say that "the integrity of test results should be maintained by eliminating practices designed to raise test

scores without improving performance on the construct or domain measured by the test."

Linn said the performance levels and gains need to be evaluated as to their generalizability with comparisons to NAEP, college admissions tests, and school district norm-referenced test results. The "Lake Wobegon effect," when almost all students are above average, often is caused, he said, by repeated use of assessments, test familiarity, and a more restricted domain on the test than in the content standards. Studies of trends in percentile rank of state means found the Lake Wobegon effect as well as the sawtooth effect, which is when test results go up and down in relation to changes in the tests. The reasons for these effects include the use of old norms, exclusion of certain students from tests, test reuse, teaching to the test, and cheating, he said.

Target test preparation can lead to bizarre claims in commercial advertisements, such as "Raise Your Test Scores 20 to 200%." This is one of the realities that standards have to deal with, Linn said, although he admitted that the line between test preparation and instruction can be fuzzy. "Ten years from now we will have to ask if learning has improved, or are we just seeing test scores improve," he said. Misinterpretations will occur over setting such standards as proficient or below proficient, he warned. The same warning appears in the 1999 testing standards: "If there is a sound reason to believe that specific misinterpretations of a score scale are likely, test users should be explicitly forewarned."

Asked about the level of expertise needed to understand the standards and models, Linn admitted that policymakers and other laypersons generally lacked the technical knowledge necessary to understand the implications of test-based accountability. "To explain the standards in language that is not off-putting is a challenge," he said, "and term limits in legislatures are eliminating any long-term memory. . . . We somehow have to find a way to communicate more effectively."

CRESST's efforts to develop standards for accountability systems fit with new priorities of the Education Commission of the States (ECS), Jane Armstrong, director of policy studies for ECS and a panel member, told the conference. Policymakers do not have much time to make things work on their watch, she pointed out, so ECS must focus on a few issues and package information so that it is easily accessible. It has established a National Center on Education Accountability that offers policy research and action research, engages policymakers in the issues on accountability,

creates partnerships with other organizations to translate and disseminate research for policymakers, and serves policymakers through a clearinghouse and a redesigned Web site.

ECS plans to engage policymakers in building better accountability plans through in-state seminars, ongoing design and policy workshops, and linking of the designs to evidence of what works, Armstrong said. A partnership with CRESST could offer states multiple accountability models incorporating CRESST standards that would be part of ECS' outreach efforts. "If a state picks a norm-referenced test, for example, we can show strengths and weaknesses on the Web site using CRESST research," she said, "or if the issue is misclassification of students, we can give research information on it."

If an effort is made through ECS channels to discuss the trade-offs that politicians care about, Armstrong said, "then I think we can get the standards into the hearts and minds of policymakers."

Armstrong admitted that the political process is not always rational, but "as we see some of the backlash and get more research and policy options to states, we can encourage them to rethink their accountability systems." The most effective model for working with policymakers, she said, "is to get people from a few states working on the same issues in the same room several times a year."

What Do We Know About Assessing Quality Teaching?

There is a difference between teacher quality and teaching quality, as described by Michael Knapp, director of the Center for the Study of Teaching and Policy at the University of Washington. The former "is often taken as the ultimate policy target," while the latter should be the ultimate concern of the profession and the public. If policymakers are off target, it is because the issue of assessing the quality of teachers and teaching was not high on the education policy agenda of states for many years. States left major decisions to others. Parents often made marketplace decisions about quality issues, "voting with their feet." Quality issues supposedly were settled by the certification process, which would screen out less-competent teachers, or by aggregate test scores, which would reveal school performance and "tell all one needed to know about teachers' capabilities."

When state policymakers changed their assumptions and began looking at quality issues, they discovered four interlinked problems. One is technical and

unresolved—how to get good, valid assessments of what teachers do. Another is the political force field of formidable players who don't necessarily agree with each other, including teachers' unions, business leaders, teacher educators, school districts, and the news media. State policy-making has to work through this force field, Knapp said.

The third problem is the policy strategy field, where supports for teaching need to converge. "What if we have a wonderful assessment but terrible teaching environments?" Knapp asked. Finally, there is the problem of public trust, which is impatient and wants to see results in a year or two.

In this context, the time-honored approaches such as the marketplace, traditional certification, principals' annual classroom observations, or the National Teachers Exam are insufficient. States are heading in a number of new directions, including

- articulating standards for teachers and teaching itself, making them very explicit (half of the states are doing this);
- strengthening the assessment of teachers as they enter and exit preparation programs;
- generating intermediate, tiered certification systems with performance-based demonstration of teaching as a requirement (a few states are doing this);
- tracking down uncertified teachers or weeding out incompetent teachers through a loosening up of tenure;
- beefing up on-the-job supervision or allowing it to be done through such strategies as peer review and real engagement of principals in the processes of improving teaching;
- promoting certification for accomplished practice through the National Board for Professional Teaching Standards (more than half of the states have policies regarding National Board certification; the number of certified teachers has reached 5,000).

"The jury is very much out on these efforts," Knapp said. "We don't know yet if these things or a combination of them will do the trick." The movement toward sound teacher assessment policy presumes further progress in four areas, Knapp added. They are (a) the development of rich, explicitly articulated conceptions of 'good teaching'; (b) further evidence about the connections between the attributes of teachers or teaching and classroom performance, as well as student learning;

(c) enactment of teacher assessment policies as part of a broader teacher policy strategy; and (d) better understanding of the links between state-level policy and classroom practice.

If these issues are puzzling to policymakers, as Knapp claimed, then long-term research by Megan Franke of UCLA/CRESST might make their task seem easier. Franke's research on measuring the quality of math teaching has challenged many assumptions about quality and about professional development. Researchers at UCLA and the University of Wisconsin provided professional development to math teachers over time, then returned four years later to see what happened to classroom practice and whether teachers continued to learn.

The researchers found out that those teachers who were most successful were those who had become good learners themselves, not those who performed best during the time of the project "Consistent across all of the teachers who continued to grow was their detailed knowledge about what their children knew about math, which forced us to go back and see how carefully teachers understood children," Franke said. Also, knowledge was well structured in a way that made sense to them, "not how we had structured it." And finally, "all of the successful teachers had developed an ability to use organizational structures that evolved and changed. . . . Their job was to build on that knowledge."

Researchers don't often consider "the identity issue," Franke said, or how teachers view themselves as learners. The findings from the follow-up research led those involved in the project to create new principles for professional development, evaluation of teachers, and student outcomes. "It forced us to think of a much more layered approach," Franke explained, "and to look at teachers learning on their own, in workgroups, and as part of a school community."

She cautioned about drawing links between quality teaching and student achievement, noting that quality is affected by content debates, school cultures, the sorting of teachers, and the time and opportunities for teacher learning.

One of the most popular methods for evaluating teaching in current policymaking circles is the value-added assessment system, first developed by a statistician at the University of Tennessee. It is being considered in CRESST's work on developing models for accountability, according to Lorrie Shepard of CRESST/University of Colorado at Boulder. The Tennessee Value-Added Assessment System of William Sanders has two key features:

- Statistical techniques are used to calculate an estimated mean scale for each teacher and school. Rather than using student background factors, the system uses students' prior achievement scores as controlling variables. It then calculates how much each teacher and school contributes to Tennessee's state test score results.

- Gains in estimated mean scale scores, calculated from one grade to the next, are compared to national norm gains. Results are reported as a percentage of the normal rate of gain.

Shepard, however, pointed out some potential problems with the system. It requires a trade-off, providing technically sophisticated analysis but limiting the quality of the assessments used (multiple-choice tests or commercially available tests with a limited range of test formats). Also, the analysis adjusts for initial differences in student achievement, but it does so imperfectly, leaving considerable potential for systematic bias in the estimation and reporting of teacher and school effects. In particular, because there is a tendency to confound teacher effectiveness with student ability and background characteristics, "good teachers who teach high-ability students will have a better chance of being identified as good than equally good teachers who teach low-ability students," she said.

There are other issues of fairness with the Tennessee system, as well as year-to-year instability of gain effects, Shepard said. Yet, "we say these are problems, but until we have further research we will not be able to say if they are small or large problems." Simulations and real data sets are needed, but enough doubts exist so far about these problems and issues that they should be taken seriously, she added.

Ten years of experience with assessments by the National Board for Professional Teaching Standards has yielded some important lessons, according to Lloyd Bond of the University of North Carolina, Greensboro, and a senior advisor on assessment to the National Board. The lessons include the following:

- The information in a powerful, evocative assignment by a teacher "is the best single indicator of good teaching we have."

- While a guiding theory of teaching is certainly important and even primary, a comprehensive assessment of teaching must be firmly grounded in craft knowledge and in the wisdom of practice.

- Teachers should be "models of literate adults."

- Assessing teachers well and comprehensively is expensive.

- Assessing teachers well is hard.
- Teaching well is even harder.

Bond also discussed some fundamental validity issues in National Board certification, such as the hazards of having the freedom to choose what is presented; the importance of context, scorer selection, and training; scorer bias; and "glibness" of applicants. The issue of bias, he said, "had caused the National Board more pain than any of the others." It is an enormous challenge, he said, to have only three days to train teachers from Iowa to assess teachers from East St. Louis, but culturally responsive pedagogy is important to recognize.

In the discussion that followed, Shepard noted that the value-added system may be helping principals (in Tennessee, the information is shared with principals but they may not use it for formal evaluations). There are a lot of protections built in, she pointed out, but the information from the value-added system becomes more important if it jibes with anecdotal evidence gathered by the principal. However, "what I would like to see is much more nuanced information on whether teachers know how to intervene when students are not doing well, the deep information about kids that Megan Franke discussed," she said.

Asked about the possibilities of a more teacher-friendly accountability system, Bond replied: "As long as accountability is tied to student achievement and achievement is measured the way it is, I can't think of a teacher-friendly assessment system." Knapp said that accountability systems are not designed to be teacher friendly, although the National Board certification does encourage candidates "to internalize their work."

Franke added that accountability should be tied to teachers' work and that "we have to find a way to recognize what teachers are doing and what they are capable of and build from there."

Applying Research on Assessing Reading Skills

A number of studies and publications at the federal level preceded or give substance to activities under the Reading Excellence Act, said Joseph Conaty, director of the Special Initiatives Unit in the U.S. Department of Education's Office of Elementary and Secondary Education. His statement opened a panel discussion on assessment of reading skills. So far, 27 states have received funds in a competitive

process under the Act. Within the states, most of the funding is going to low-income districts where reading performance is a problem.

Conaty re-emphasized the uniqueness of the Reading Excellence Act. It defines basic reading skills such as phonemic awareness, decoding, fluency, vocabulary, comprehension, and motivation. Grantees must use research-based methods to teach reading. "The heart of the legislation," Conaty said, "is to change early reading instruction by changing teacher practices." The Act also emphasizes family literacy, development of resources, and investment in assessment.

The ideal assessment of early reading skills, according to Conaty, would be one that is conducted frequently, both informally and formally; it would identify children early who are at risk of reading failure, it would guide instruction, and it would enhance teachers' knowledge of best instructional practice. The accountability for teaching reading is on adults, he pointed out, because "you can't penalize young children."

To provide a research agenda that will support such an ideal assessment system, the RAND Reading Study Group was organized and is working on a framework for the development of reading across the whole range of learning, Catherine Snow, of Harvard University, reported.

At its first meeting, the group argued about and came to a consensus on the definition of reading comprehension. The group members decided comprehension has three major components:

- **Knowledge, or getting the gist.** This involves recalling what was read, gaining content knowledge, drawing appropriate inferences, learning new vocabulary items, and evaluating the relation of new to previous knowledge.

- **Application, or using newly acquired knowledge in further comprehension.** This is demonstrated by performing a task, solving a problem, building connections, and evaluating the utility of knowledge.

- **Involvement, or aesthetic reaction.** This involves motivation and efficacy, topic interest, and critical evaluation—when a reader "gets lost in a story and is feeling like a good reader," Snow explained.

This definition of reading comprehension has several implications for the kinds of assessments needed, Snow said. The assessments ought to

- have the capacity to reflect authentic outcomes;
- be congruent with the component processes of reading comprehension;
- be developmentally sensitive;
- have the capacity to provide for the identification of individual children whose comprehension is poor;
- have the capacity to identify subtypes of poor comprehension;
- be instructionally sensitive;
- be open to intra-individual differences;
- be useful for instructional decision making;
- be valid across social, linguistic, and cultural variation.

Once assessments determine students' reading levels, teachers can figure out interventions, Snow said. The assessments might show, for example, weaknesses in vocabulary, word knowledge, engagement and motivation, linguistic knowledge, fluency, or integrating nonprint information with text such as using charts in science classes. Other areas for research should include variation in performance across text types, the impact of accommodations on the test performance of second-language learners, the impact on performance of specifying different purposes for reading, the capacity to differentiate domain-specific and reading-general components, issues related to electronic reading, and the development of automated reading assessment systems.

Currently used vocabulary tests are an example of treating reading skills too simplistically, Snow said. They provide only a piece of the picture of comprehension. More helpful are assessments of the breadth of a student's vocabulary, such as: "Tell me all you know about the nose."

Snow emphasized that language background differences are important because learning to read in other languages can be a different experience. English has deep orthography while Spanish does not; that is, Spanish tends to read the way it is spelled. "We must take into consideration what second-language learners know about reading in their own language because that represents a set of skills," she said. "Do we care mostly about reading in English or mostly about reading?" she asked. Also, she added, "we do not have any good test to assess knowledge distributed over two languages."

Texas' emphasis upon early reading resulted in the Texas Primary Reading Inventory (TPRI), a diagnostic assessment adopted for statewide use. Developed by psychologists and other researchers at the University of Texas, Houston, and the University of Houston, and tried out in the Houston area, it was formally supported by 1997 legislation that called for a diagnostic system linked to the state curriculum standards. The goal was to identify reading problems early in order to have all children reading by Grade 3.

Reviewing the research behind the inventory, David Francis, director of the Texas Institute for Measurement, Evaluation, and Statistics (TIMES) at the University of Houston, said that reading outcomes vary across children. About 5% learn to read "almost magically," another 20%–30% learn easily no matter what method of instruction is used, but about 30% of children struggle to learn to read (equally true for boys and girls).

Most reading problems, Francis said, occur at the level of the single word, where children's reading is characterized by slow and labored decoding, which inhibits comprehension. Phonemic awareness is important to developing word recognition skills, he said.

Early assessments are needed because "children do not simply outgrow reading problems," according to Francis. Research shows that early intervention is clearly effective, but not all children respond equally well to all interventions. By combining early identification with targeted intervention, he said, "the effectiveness of early interventions may be enhanced."

The TPRI is administered by teachers in Grades K–2. At each grade, the test has two major sections—screening to identify children not at risk and an inventory to inform instruction. In kindergarten, the test is administered at the middle and end of the year using letter names and sounds and phonemic awareness for screening. The inventory part of the test includes book and print awareness, phonemic awareness such as rhyming and blending word parts, phonemes, graphophonemic knowledge, letter-to-sound linking, and listening comprehension. Some of these same components are repeated in the first- and second-grade tests.

Describing the process used in the preliminary studies, Francis said the researchers used an IRT model because it allowed construction of a test that would require a minimal amount of time for screening (5–8 minutes) yet still provide accurate discrimination of skills around the cut-point for abilities. It also allowed

new screening and inventory items through appropriate linking studies and the development of word lists linked to the stories that are used to determine students' comprehension. The researchers also wanted to minimize the chance of not identifying children who were at risk, yet keep the false negative rate below 10%.

Currently being used in more than 90% of Texas' public elementary schools, the TPRI has been well received by teachers, who are requesting more training. Regional service centers provide most of the training for teachers. Training is provided through workshops with supporting videos and CD-ROM materials. Teachers report that the screening now allows them time to focus on students who need the most help while reducing their time used for assessment.

The TPRI provides opportunities for further research, Francis said, including linking it to outcome assessments used in the accountability system to improve identification of at-risk students, evaluation of instructional decision making and intervention strategies, improved decision making through computer-aided administration and student profiling, and development of comparable instruments for literacy instruction in Spanish.

In the discussion that followed, panelists agreed that few resources exist to help with diagnosis of reading skills and intervention for students in the fourth grade and up.

Asked about a highly structured reading program adopted in the Los Angeles schools, Snow said that "the conditions of a school district determine the approaches to early reading instruction, and where a district has many teachers not prepared to teach early reading, a structured approach might be needed."

In discussing comprehension in the early grades, Francis said that the TPRI focuses on screening for students with problems, but that all students are tested for comprehension as part of the inventory. Snow noted that good reading instruction in Grades 1 and 2 generally has been defined as producing students who are reading words, while comprehension is saved for Grades 4 and 5. "That obviously doesn't work," she said. "You can't read words without the context, and we have effectively hidden a secondary source of reading difficulty by not going deeper."

Increasing Accountability at the National Level

Drawing from elements of the Clinton administration policy and platforms of the Republican and Democratic presidential candidates, George Bohrnstedt, of the

American Institutes for Research, asserted that all evidence points toward a growing demand from the federal government for educational accountability.

A major issue will be the role of the National Assessment of Educational Progress, he said. There are good reasons to expand its reach because it is based on a national consensus of content frameworks, its items are closely aligned with the frameworks, it uses a mix of item types, and it represents a state-of-the-art assessment in terms of psychometric criteria.

On the other hand, as a National Research Council study pointed out, if NAEP were to become the assessment of choice, its data could become corrupted because teachers might teach to it, the increased testing being discussed would place a lot of burden on it, and it might increase problems in recruiting schools for non-high-stakes NAEP subjects, burden item development, and increase exclusion rates. Bohrnstedt recommended instead that states develop their own assessments aligned with their content standards and then use NAEP as a low-stakes assessment to independently monitor state assessment results.

Building good accountability systems is not easy and must be composed of more than tests, he said. States and school districts need to provide adequate financial resources for facilities, equipment, and instructional materials; hire qualified teachers; provide professional development tailored to teaching placements and needs; develop standards-based assessments to track progress toward goals; and provide feedback to students, parents, and teachers.

Drawing from criteria proposed by CRESST Co-Director Robert Linn, Bohrnstedt said there must be safeguards against selective exclusions, and the assessments should be equated from year to year, include multiple indicators of school success, and emphasize year-to-year growth. Accountability systems also should be evaluated for both intended and unintended consequences.

Introducing more radical thoughts about accountability, Edmund Gordon, professor emeritus of psychology at Yale University, proposed that “accountability and standards may be the wrong road if we are talking about significantly improving education” unless, he added, “you are willing to also include responsibility as a part of accountability.” Current policies regarding the centrality of the accountability movement and the federal government’s role in it “simply won’t get us where we want to go,” Gordon said.

His approach is to support what he called "affirmative development of academic ability." So much of schooling, Gordon said, depends only on the availability of a variety of education-related capitals. Neglected are other factors necessary for academic success, including human capital (health and "people in your lives who have enough development to help you develop"), social capital, and polity capital ("a sense of belonging to society"). If these are not present, "it may be immoral for us to put accountability on school outcomes."

The first component of a national system of accountability, Gordon said, would have to be responsibility for the universal access to these essentials for academic development, even though this approach may seem to let those who teach "off the hook." He added, however, "I don't think middle class Blacks and Hispanics are underachieving because they don't know what the standards are. Accountability means being responsible for delivery of universal access to resources/capital we know are essential."

Another kind of responsibility is to become much more accountable for the minutiae of accountability systems. Intellectual behavior is in danger of being "Balkanized," he said, when, instead, it should be seen as the capacity to solve problems using trans-knowledge, or bringing various sources of knowledge together to solve problems. "Even if we stick with the idea of accountability," he said, "we have to ask: accountability for what?"

Trends in national policy regarding accountability illuminate another issue—that of assessment for learning versus assessment for accountability. The two approaches are sometimes perceived as diametrically opposed, and they do require different configurations to meet different purposes under different constraints, said Robert Mislevy of CRESST/Educational Testing Service.

Discussing assessment as evidentiary reasoning, Mislevy said what is really wanted is knowledge about what students know, can do, and should work on next. However, "what we get to work with is quite different," he said. "We see them say, do, or make something in a handful of particular situations." An attempt to get at more was the California Learning Assessment System (CLAS), which consisted of on-demand assessment, embedded assessment, and organic portfolios. Another example of going beyond the snapshot is the Advanced Placement Art Portfolio Assessment, which includes work created throughout the year, usually as part of a

class; is standards-based; and uses statistical models to monitor and improve communication.

Mislevy discussed two principles from evidentiary reasoning:

- Evidence forming the basis for probabilistic conclusions has three major properties that must be established—relevance, credibility, and inferential weight.

- The evidentiary value of data depends on the question you are asking and what else you know to condition your interpretation of the data. Examples of the latter might include the ACTFL reading proficiency scale, the AP Studio Art concentration section which makes students define the problem they were addressing and explain their context, and CLAS' on-demand assessments compared to embedded assessments. Also, a further consideration is "what you can ask of students," such as expectations and standards of evaluation. The context will be different, he noted, between a multiple-choice assessment and studio art where it may take a whole year to understand the standards.

The challenge, Mislevy said, is whether "we can retain the relevance and connectedness associated with classroom assessment and, at the same time, serve the communication and credibility functions associated with low-context, drop-from-the-sky assessment." His answer was "yes," but not by using the same data. Classroom assessments maximize information about the context for learning. Accountability assessments, on the other hand, are much more restrained and use context only insofar as it applies to everyone who takes the tests or interprets the results. Developments in educational measurement and in technology can help break down the dichotomy between classroom assessments and accountability assessments by making it possible for people to share what they are learning about the different contexts for assessments.

Rather than start with an assessment and then try to figure out its purposes, Mislevy advised starting with purposes—"what kind of learning do we want to know about." Data with different properties means data with different relationships to what students are working on, he said. There are different degrees of contextualized and conditional interpretation of performance and different relationships between data and the amount of context to be agreed upon and shared. This, he said, is what the now-defunct CLAS was trying to do. These issues, he said, will be with us continuously.

The third panelist addressing national policy, Martin Orland, of the Office of Educational Research and Improvement at the U.S. Department of Education, pointed out that accountability pervades federal policy in general. The Government Performance and Results Act (GPRA) has changed how agencies work, he said, requiring them to set annual goals, report on successes, and describe future strategies.

In a review of 100 programs under GPRA, the performance plans indicate two types of problems, Orland said. They undershoot the mark, such as an indicator adopted by the Eisenhower professional development grant program that only measured how many teachers trained in the program taught in high-poverty schools (no quality or outcome measures). Or they overshoot the mark, such as an indicator that there will be lower incidences of drugs and violence in schools, even though drug use depends on nonschool factors.

The future of GPRA is unsure, Orland said, but it is one piece of evidence proving the high interest in performance indicators. He discussed two other indications of interest:

- **Return of randomized experiments.** There is expanded interest in holding intervention programs accountable by using experimental models, he said. The National Reading Panel report, for example, was a mega-study of experimental studies. Just as developing good indicators for GPRA is an issue, so does this trend beg the question about having the right kind of outcomes and indicators. Research on class size and results from TIMSS raise concerns about how long interventions need to be studied.

- **Democratic and Republican platforms' emphasis on accountability.** They reflect policies that are salient to the American public, Orland said, and do not hesitate to set high stakes for institutions rather than for students directly.

The challenge, he said, "is to see if we can work with policymakers and the public to develop systems that get what the public wants and also are credible." He said he was "moderately optimistic" that systems could be built in more thoughtful ways, and was sure that the issue at the federal level was not whether there would be a larger role, "but how it will get done."

During the discussion period, conference participants asked how individual results could be better tracked, such as using social security numbers. That would be advisable but is not likely, Orland replied, even though the National Education Goals Panel had discussed it as a strategy to track high school completion. Some

states, such as Texas, have developed ways to identify individual student results, he said. Mislevy added that even with a very good tracking system, "the information you would like to know about students cannot be fully captured. . . . There still would be things going on in a school that you would not know about, no matter how big the data file."

One participant commented that there is a perverse relationship between a national accountability system and the effectiveness of schools, which depends on "a nurturing atmosphere." If a national accountability system identifies ineffective schools, more than likely teaching freedoms will be restricted at such schools, and then they cannot attract teachers who would help them. Bohrnstedt replied that the reason accountability is such an issue is because of the persistent gap between students in low-performing and high-performing schools, but "we need to be sure we use creative solutions for kids in low-performing schools."

Orland said a "lot of work needs to be done on validity because we are not sure schools identified as low-performing really are." In Gordon's opinion, the problem of low-performing schools may not be solved in the schools themselves. Benjamin Bloom's mastery learning, a program used in low-performing schools, was at best able to move only 63% of the students up to grade-level performance, he said, "but the most serious education problem for the nation is what to do about the other 35%. . . . We cannot function as a democracy if we let them fail." Sending them to another school will not solve the problem, he said, re-emphasizing his call for much more serious engagement with opportunities to learn.

Final Comments

The 2000 CRESST national conference was about making valid decisions about the education system, CRESST Co-Director Eva Baker said at the concluding session. The discussions informed work at CRESST, which will include analysis of different modes of accountability, computerized assessment models for teachers, and developing good indicators of practice. CRESST also will explore teacher assessment issues.

The conference, CRESST Co-Director Robert Linn commented, "brought together different perspectives and affirmed that much was to be gained from this community about accountability." What needs to be communicated, he said, is the

responsibility for informed discussion and decision making "before accountability is written into law."

Joan Herman, CRESST associate director, reminded the participants that the stakes in accountability systems are "enormous," but the conference discussions showed "that we know a heck of a lot and need to move ahead and do something useful." That is why CRESST is seeking as broad a constituency as possible for involvement in the development of standards for accountability systems, she said.

Small Groups, Large Questions, and Reports on Ongoing Research

The CRESST conference provided several opportunities for participants to discuss progress on research projects in small groups and to air their concerns at informal forums.

Benchmarking and Alignment of Standards and Testing

Following a morning of panels on state and district accountability systems, researchers affiliated with Achieve, Inc., discussed that group's work in helping states design appropriate standards and align assessments with them. Achieve, Inc., established by state and business leaders, helps states benchmark their standards and assessments, builds partnerships, and serves as a national clearinghouse for policymakers, according to Robert Rothman, senior project associate for the organization.

Benchmarking to quality standards and assessments is a practical as well as desirable strategy for states, he said, noting that state investment in assessments has more than doubled in the last four years, from $160 million to $360 million. With all this activity underway, "states needed some guidance on how they were doing." States undertook standards-based reforms without external references as to whether their standards were as high as they should be, or whether their assessments were measuring the right things, and whether their accountability systems were fair. Achieve, Inc., came into being as an external agent to assist states in informing and evaluating their work

A guiding principle for the work, Rothman said, is that quality matters. Judging quality requires looking outside a state's own borders. Also, sustaining standards-based reform requires a willingness to make mid-course corrections and it

requires leadership and commitment. Most of the states, Rothman reported, have acted on Achieve, Inc.'s recommendations for improvement.

Describing the process for reviewing state efforts, Achieve, Inc.'s Jean Slattery said two to three experts examine standards using several criteria: Are the standards rigorous? Do they have clear progression of knowledge from kindergarten through high school? Are they measurable? Fifteen states are in the process of revising their standards through such reviews, she said. The review teams, which always are from out of state, are balanced carefully and include an academician, an experienced teacher, and an assessment expert.

The process also evaluates the assessment-to-standards alignment. This includes five analyses:

- The individual items are examined to confirm the test developer's blueprint. The analysis seeks to determine how well items match the content and performances described in the standards, and to verify that the source of challenge stems from the knowledge or skill required and not from a flaw in the way the item was constructed.

- Each set of items relating to a standard (for example, all items mapped to algebra) is then evaluated for the level of challenge the items pose. Reviewers look for a span of demand and track individual items as being easy, medium, or hard.

- Each item set is further evaluated for balance to ensure that items mapping to a standard are representative of the emphasis that the standard places on content and performances.

- Range is calculated for each standard.

- Finally, reviewers comment on the strength and weakness of the test as a whole, and when evaluating several tests at different grade levels, they comment on progression on a standard-by-standard basis and across the tests taken as a whole.

Reporting on findings from nine states so far, Achieve, Inc.'s Jennifer Vranek said the good news is that standards, for the most part, are now written in clear and jargon-free language, compared with early attempts, and standards often emphasize both the basics and higher level skills.

When state standards were compared with the benchmark standards, however, the reviewers found that state standards sometimes were not specific enough to provide clear expectations and guidance for schools and students. They often omit

important content, particularly in early literacy, and the standards often have not been focused enough. “Developers have not made tough choices,” she said.

As for assessments, the tests in nearly all states measure content found in the standards. In math, an attempt has been made to emphasize both basics and higher level applications. However, in many states the test blueprints do not typically provide enough information about the match of items to standards. Where standards include a range of performance expectations, the test items tend to measure the least cognitively complex skills. This is an important issue to solve because “it sends very different messages to teachers,” she said.

Furthermore, the tests do not always use multiple-choice or open-ended item formats as effectively as they could. “It is not clear to us that test makers know how to use the different formats,” Vranek reported. “For example, there are times when an open-ended item could have been done better as a multiple-choice item, and vice versa.” Some state tests focus on certain standards to the exclusion of others, thus omitting important content included in the standards. “What is the message to educators if we tell them to focus on high-level skills, and then we don’t measure them?” she asked.

Reports to states on their reviews also comment on the quality of the standards, providing detailed information to the states on how their standards compare with those in high-performing countries, TIMSS data, and other states with good standards.

Commenting on the reports from Achieve, Inc., Lauren Resnick, of CRESST / University of Pittsburgh, said the information contradicts claims that state tests are aligned with standards, even when the tests are bought directly off the shelf. “Many have felt a discomfort with this claim, but we haven’t had a methodology to check out the statement, and now we have one,” she said, noting that the process so far has taken three years and is very complicated and difficult—“not like going to the flicks.”

An important component of the process is that it recognizes there are social judgments made in selecting standards and assessments and is willing to participate in debating them and making the process public, according to Resnick. People often do not realize, for example, that there are human judgments behind multiple-choice items. Yet she predicted that a “whole new revolution in psychometrics” will enable multiple-choice items to be very similar to performance tests.

Resnick also noted that genuine alignment of standards and assessments "is a much more complicated process than we have been allowing ourselves to believe, but an unaligned system is not fair to students or teachers and is not very good." Efforts at test alignment are providing a better way of looking at standards because one can circle back and ask: "Are these standards even worth aligning to?" Within a year, she predicted, the process will provide much more explicit understanding of what makes good standards.

Making sure those teaching or taking the test know the standards is a fairness issue, Resnick said. "People really do care. They are not trying to fail a lot of kids, but some practices have that result."

Using Technology

In another small-group session, Joan Herman, associate director of CRESST, reported that using data to move reforms along, which is a basic strategy of current efforts, has different effects, depending on the school context. A vision of data use in standards-based reforms, she said, would see them as helping develop a consensus on standards for all students. Data also would help researchers assess status and progress, explain why things were happening, plan improvement strategies, monitor and assure progress, help continuous improvement, and be grounded in the action and evaluation made by local schools.

The reality, however, is that data use runs into "technological, cultural, and political" issues. Herman offered descriptions of the strategies used by two different schools to illustrate this point. One high school, relatively high achieving in a suburban area with minority students bused in, had a well-organized school planning team of teachers and administrators. It used data to find sources of poor performance among the students bused to the school, resulting in new programs and additional support for these students and a commitment to monitor their progress. The other school, in an urban, economically poor area with high proportions of English-language learners, was under pressure to improve test results. A teacher committee focused on improving language arts and decided to use data for school and classroom planning, integrating mandated and classroom evidence. The contrast brings up issues of school assessment capacity and using data fairly and honestly, she said.

The studies on data use, conducted in California, Illinois, and Wisconsin, point to additional challenges, Herman said, including leadership and real reform

processes, alignment of classroom assessments, the infrastructure to support access to good data and its wise use, and continuing tensions between top-down accountability and authentic bottom-up reform.

Donald Morrison of Co-nect Schools, a reform model, said experience with technology and its use to support school and classroom assessment reveals several issues. Co-nect, adopted in more than 150 schools, uses community accountability for results, project-based learning, authentic assessments, team-based school organization, and sensible implementation of modern technologies.

The project has learned, Morrison said, that more reliable, valid, and valued ways of measuring the results of technology use in schools are needed. Also, technology-based assessment tools need to expand, rather than limit, visions of what kinds of results really matter—"and lead to a truly deep understanding of what children know and what they are thinking." Finally, he said, school-level people need help in managing the great amount of assessment data that are being generated so they can turn the information into widely shared knowledge and understanding.

An answer to the last problem could be the CRESST Quality School Portfolio (QSP), reported Derek Mitchell, chief architect of the QSP data management technology. It was developed, he said, because schools rarely collect relevant information on progress and efforts within school buildings. They collect data for other agencies and for the news media, but not to further their own goals.

The portfolio software allows schools to collect and analyze data relevant to their preset goals, make corrections, and stimulate school discussions around their reform efforts, Mitchell said. Pilot programs were conducted in Los Angeles and Chicago, and training is underway in collaboration with other groups such as the American Association of School Administrators, the National School Boards Association, the Wisconsin Center for Educational Research, and the Illinois State Board of Education.

Another source for school and classroom assessments is the Performance Assessments Links in Science (PALS) funded by the National Science Foundation. It shares more than 150 K–12 exemplary assessment resources online and provides Web-based professional development support. Users can match the assessments to their state science standards and curriculum frameworks, explained Edys Quellmalz, of the Center for Technology in Learning of SRI International.

Effect of Accommodations

Another small-group session focused on continuing research related to accommodations made for student test takers with disabilities or who are English-language learners. Participants learned that the research continues to support the conclusion that language is a major background variable in NAEP math performance.

English language learners, given several kinds of accommodation, overwhelming preferred linguistically simplified items and performed better with this accommodation than with items in the original English version or in a Spanish-translated version, reported Jamal Abedi, of CRESST/UCLA. In another study, three forms of accommodation were used in addition to the standard NAEP condition—extra time, glossary, and glossary plus extra time. Using a glossary plus extra time had a large impact on the performance of both English language learners and regular learners, but it was more evident with the latter group, "which raises concern over the validity of the accommodations," he said.

The research confirms that the greater the language demand of a test item, the more difficult it is for limited English proficient learners, Abedi said. "We hope to pinpoint how to reduce the difficulty of the language but in a way that does not affect an understanding of the content," he said.

Another study examined New York State's efforts to include students with disabilities in a field test of its revised Regents Comprehensive Examination in English. Data analyzed by Laura Hamilton (RAND) and Dan Koretz (RAND/CRESST) showed that accommodations were being used with needier students, and extending the time was more beneficial on open-response questions than on multiple-choice items. Test forms with more multiple-choice items were more problematic for children with disabilities even though the researchers did not find very many multiple-choice items that were excessively difficult for disabled children.

Compared with findings from similar research in Kentucky, New York had a higher rate of exclusion from testing and a lower rate of accommodation. The research base on accommodations is not strong, Hamilton said. "We don't have real criteria with which to compare performance or know why accommodations were or were not given and how they were given," she added. Generally, there are very few experimental studies on assessments of students with disabilities.

State data presented by Judy Elliott affirmed the different findings in many states. Formerly with the National Center on Education Outcomes and now assistant superintendent for special education with the Long Beach (CA) Unified School District, Elliott said the center began collecting accommodations data from the states in 1993. At that time, only 21 states had guidelines or legislation regarding accommodations, and the accommodations were very limited. Now, 48 states have guidelines, but there is little consistency across the states. "States may be using the same test, like the SAT-9, but have totally different allowances for accommodations," she said.

The most consistent trend over the years has been to leave decisions about accommodations up to the Individualized Education Program (IEP) team, Elliott said, "even though most teams have not been trained to make such decisions."

A major issue centers on standard versus nonstandard accommodations, or those allowed or not allowed by test publishers and other technical groups. Braille, for example, is standard in half of the states but nonstandard in the other half. In Oregon and California, accommodations can be written into IEPs, but if a nonstandard accommodation is used in SAT-9 testing in California, the student's test results are kicked out of the accountability system, known as the Academic Performance Index (API), she said. The greatest change since data were first collected is the approval of spell checkers (not allowed in 19 states). Almost all states allow extended time, but not necessarily as a standard accommodation.

Now that she's able to look at data from the school district perspective, Elliott said, she's discovered that some teachers have been excluding students with disabilities because they didn't believe in testing them. Other teachers do not fully understand the purpose of the assessment program. The district's response has been to revamp the IEP document, adding categories of accommodations for instruction and for assessments, both classroom and district. However, understanding accommodation use, standard and nonstandard, for the SAT-9 "is a different story," she said.

Practical Aspects of Assessment-Driven Reforms

Drawing on research in Washington state and Kentucky on schools making exemplary progress on reforms, Hilda Borko, of CRESST/University of Colorado at Boulder, found that state-mandated reforms were making a large impact. From four years of study, she found that the most successful schools

- had curriculum alignment, working on one subject each summer; professional time and money were spent on aligning the curriculum;
- did not automatically accept reforms but, rather, had teachers who spent time studying them;
- tailored professional development to the unique needs of their staff, except for common training on portfolios—"across all the schools there was a real commitment to working together to do what was needed and to use state resources to get it done";
- in math, emphasized both basic and higher order skills, and integrated portfolio tasks into the ongoing instructional program; teachers started the year with a lot of scaffolding and modeling for portfolios, then allowed students to do the work themselves;
- made writing instruction very explicit, emphasizing different genres and audiences, and using scoring rubrics so students could evaluate their own writing.

Overall, Borko said, "there was a wonderful commitment to students. Teachers were not going to lose any of them, an attitude which goes against stereotypes that reforms are all about content." Also, she said, reform became a schoolwide effort, even though state testing was only at two grades.

In Washington state, the exemplary schools focused on process. Although there was attention to curriculum alignment and resources, most of the energy went into including all teachers in the reforms, not just those in the grade being tested for accountability measures. Again, she said, "there was a commitment to doing what was best for kids."

Studying teachers' written feedback on student writing in Los Angeles schools, Lindsay Clare, of CRESST/UCLA, found that half of the elementary and middle school students received only surface feedback on their compositions, and some received no feedback at all. The type of feedback did not improve the quality of writing; the quantity of feedback related more to improving surface qualities such as spelling and grammar.

"We had expected that content-level feedback would produce changes in content, but when we restudied the feedback, in half of the cases it was not very substantive," she said. There were changes in the mechanics of the compositions, but not in the content. "It is important," she noted, "for teachers to know how to give

feedback that will push students to expand their ideas. Otherwise, their revisions look like a copying process."

Addressing the issue of state reforms in a large district, Geno Flores, of the Long Beach (CA) Unified School District, said the process often involves "a lot of mundane things." Describing the constant change in assessment policy in California, he said teachers become confused, often disillusioned. In the current situation, for example, the law prevents districts from using parallel test preparation systems. However, a textbook company whose texts were approved by the state produces a student handbook for taking the Stanford 9, "and the state gave us money to buy the handbooks," said Flores.

The district confronted test distribution problems, trouble meeting data requirements regarding background information from parents that was to be obtained on the test day, and data requests from the state that may change from April to October.

"Some really good ideas got lost in implementation," Flores said, "or those designing the system lost sight of what actually happens at the school site."

Issues in High-Stakes Accountability Systems

California's accountability system is high stakes for schools and for some teachers, who stand to gain significant increases in income. As a result, it was the total focus of a small-group session on high-stakes accountability systems.

William Padia, director of the Office of Policy and Evaluation for the California Department of Education, explained the heart of the accountability system—the Academic Performance Index. It applies to 7,000 schools in the state (another 1,000 schools are either alternative or small schools and will have a different system). An interim statewide performance target score of 800 was set at the recommendation of a technical design group (co-chaired by Eva Baker [CRESST/UCLA] and Ed Haertel [Stanford]). The API is meant to capture growth not status, he said. Growth targets are set for both schools and subgroups of students; the latter are less than the former. Annual targets generally are 5% toward the 800 figure, which is for the interim because it uses the SAT-9. The API is calculated by using a school's base as set in 1999, growth, and distance to the target. The participation rate is an important piece of the policy, Padia said. At the elementary level, 95% of students must participate in the testing; at high schools, 90% of students must participate (this will

increase to 95% in a few years). If participation is lower than this, the school would not be eligible for rewards.

Because of "a huge problem with data," the first release of the API rating of schools in October excluded about 1,000 schools. A second release in December included most eligible schools.

When the accountability legislation was passed, legislators adopted one award program, then added two more, and the three now total $670 million extra for schools and individuals, all dependent on the API, Linda Carstens, of the California Department of Education, explained.

Schools will receive funds either for meeting their targets or if they don't meet the targets. About 67% of the schools met their targets in the first use of the API for this purpose, she said, but schools that missed their targets can voluntarily enter an improvement process and receive additional resources.

The first award program, the Governor's Performance Award, basically provides schools with up to $150 per student. Some questioned why schools could receive awards after just one year of the accountability plan, but Carstens said it was meant to be "a boost for schools, to keep their motivation going."

The School Site Employee Bonus program, funded at $350 million, is a one-time-only program, with half of the money going to the school for schoolwide purposes and the other half going to school-site personnel.

The third award program is competitive and only open to schools placed in the lower five deciles in 1999. If they more than doubled their school and subgroup targets, they are eligible to compete for individual awards. Schools that did the best will get $25,000 for each certificated staff member; two other tiers provide $10,000 each and $5,000 each for staff members at underperforming schools that did exceptionally well.

A cross-interest working group has been trying to work through all the implications of this third award program, Carstens said, such as defining certificated personnel.

California is at a point of offering schools a chance to use some badly needed and well-deserved resources that are important to teaching and learning, Carstens said, "but the answer to getting scores up is standards, not test preparation."

Improving Methodologies for Evaluation and Assessment

As states get more involved in accountability, some of their policies need to be better informed, according to presentations at a small-group discussion on improving methodologies for assessment. The devil is in the details, the panelists indicated.

Katherine Masyn, of CRESST/UCLA, presented a general growth mixture model to assess intervention effects in randomized trials. Using data from a school-based intervention program for aggressive boys in Baltimore, she showed how researchers could determine what interventions were most successful with which group of youth by classifying the students as low, middle, or high aggression and using repeated measures of individuals over time. This information, she said, might "inspire future study designs to screen children according to the level of their negative trajectories, then vary interventions according to their estimated class membership."

The use of longitudinal cohorts has been ignored in practice and in the literature addressing the progress made by schools, according to Dale Carlson, a consultant to the U. S. Department of Education and the Council of Chief State School Officers. "Everyone focuses either on status, whether a school is high or low scoring, or they focus on change using successive groups of students," he said.

Carlson said that his research shows that different approaches give strikingly different results. "Many schools that have declining scores using the successive-groups method and might be identified as schools in need of improvement, for example, turn out to be doing quite well with the students that have the opportunity to benefit from the school's program using the longitudinal cohort approach."

Aggregated data of ethnic or socio-economically disadvantaged subgroups can miss progress of these groups and "lead to a lack of clarity for the API," Maria Castro, of CRESST/UCLA, noted. She recommended a much more complex system for arriving at API estimates for schools, which would include a multivariate, multilevel analysis of scores for every student, student variables, and school situation. She gave examples showing a different picture of improvement among ethnic/SES groups using this kind of analysis.

Michael Seltzer, of CRESST/UCLA, presented research on using data about where students start academically and how rapidly they progress in order "to

generate insights on effects of schooling and on interventions and especially on who benefits from the interventions and who doesn't."

Seltzer undertook this research partly because current strategies may find that two schools are similar in overall rates of progress, "but very different things are going on under the surface." Analyzing data from a group of 72 students in math classes over several years of secondary school, he found the data to be "noisy" with somewhat of a linear trend. Using quadratic models and other strategies, he looked at gender differences, then further at gender differences within groups—low, middle, and high initial status. At this specificity, he said, "there is significant interaction" regarding the growth rates. If the initial status was low, girls grow at slightly faster rates than boys. Among those who start at a high level, boys progress academically at a substantially faster rate than girls.

"This kind of information can be hidden when you are only looking at the overall gap," Seltzer said. The analysis can be done fairly easily with existing software, and can be extended beyond gender to look at other demographic categories.

"Focusing only on differences in growth rates can be misleading," Seltzer said. "The magnitude and direction may depend on where students start from."