“Closing the gap” is a metaphor for equity. It suggests that achievement test differences among groups of students differing in first language, socioeconomic status, or cultural background will be reduced by the educational efforts of schools. Another metaphor, “leaving no child behind,” similarly signals the intention of the educational system to address the specific needs of each student. From time to time we hear that the gap between two groups has been reduced. Others contend that performance differences remain. Accountability systems such as those in Texas and California explicitly set expectations for comparable progress among subgroups within schools. In California for example, schools are not eligible for substantial monetary awards unless subgroups make progress as well as the school as a whole.

The glory of metaphors, and the problem with them, of course, is that they are subject to interpretation. “Closing the gap” interpretations will vary depending on one’s belief in:

1. what content or skills should be measured;
2. the degree to which schools (as separate from other family, social or cultural experiences) can or cannot reduce achievement differences;
3. how to interpret gap reductions on one test and simultaneous gap increases on another.

The other night in Los Angeles, a group of researchers and commercial test publishers earnestly debated all three points. We were split between those who saw themselves as measurement experts and those who identified themselves as learning specialists. The lines of our disagreement were these:

**Position I: Overwhelming Effect of Background Variables**

Schools can only marginally reduce academic differences in the achievement of subgroups. Because of the pervasive impact of background and socioeconomic experience, schools can only do so much. There will always be variation within and between groups. Reduction in achievement differences can be explained by:

1. a ceiling effect. Low-achieving subgroups increase their performance substantially because they have plenty of room to improve even on tests that are not challenging. High-achieving subgroups cannot move higher on tests of limited difficulty.

**Call for Papers**

Joan L. Herman and Eva L. Baker, co-editors, invite submissions to *Educational Assessment*, a journal of original research and scholarship on the assessment of individuals, groups, and programs in educational settings. This journal covers broad issues related to assessment theory, empirical research, and use.

Requirements are posted on the CRESST Web site, www.cse.ucla.edu. For further information, please call Chris Company at 310-206-1532, ext. 71232, or company@cse.ucla.edu.

We look forward to your submissions.
2. Tests in some accountability systems cover only a relatively narrow range of content and skills. They often reuse subsets of items in administrations from one year to the next and are therefore easy to teach to. If the test covered a broader curriculum, it would be difficult to show similar gap reductions.

POSITION II: SCHOOLS CAN REDUCE THE GAP

Quality education absolutely can and will reduce gaps. We can reduce the achievement gap even when assessments measure broad, challenging knowledge and skills. We should avoid drawing conclusions about the reduction of differences when they are due to selection (who chooses or gets chosen for a program), test preparation focusing on format and tricks, or deliberate, narrow preparation for a small set of test questions.

DIFFERENT UNDERSTANDINGS OF LEARNING

How did our debate end, at least at dinner? First, proponents of both positions ultimately recognized that our arguments had been based on two very different understandings of learning. Part of the core of our disagreement about interpretation related to the academic training of the contenders. One side felt that gap reduction was a result of teaching to a narrow test and ignored transfer, that is, students’ ability to use their knowledge or skills in new situations or to solve academic or practical problems.

One side felt that gap reduction was a result of teaching to a narrow test and ignored transfer, that is, students’ ability to use their knowledge or skills in new situations or to solve academic or practical problems.

Second, we agreed that reducing the gap around a particular set of performances does not necessarily mean that the tasks are easy, low-level, or not really worth our time. Performance depends upon the quality of instruction, the clarity of outcomes, and the motivation of students. Consider, for instance, the substantial progress United States students made in writing when actual compositions became major components on state assessments.

Third, even gaps on challenging, explicit problems will be reduced rather than closed unless we can provide strong support for children who come to school with less language ability and with less preparation for learning (think about the “ready to learn” goal of the Education Goals Panel).

There is a relationship between who takes tests and level of performance. States, districts, and the federal government all have adopted policies to assure the inclusion of children on tests and to improve both achievement and graduation rates. The implication is that both performance and participation count. When considering the effectiveness of schools, it is clear that on general achievement measures there is a persistent relationship between socioeconomic backgrounds of the children attending a school and the performance of students on achievement tests. But there are also a host of studies that show major exceptions to the rule. We have schools that beat the odds, where students demonstrate achievement patterns that defy the predictions of low achievement based on their family circumstances.

It stands to reason that performance can be most directly affected by that which can be most directly taught. We need to clarify what is a legitimate focus and what is inappropriate. Casual charges about “teaching” the test or asserting big gains based on very small improvements will not help education in the long run.

Casual charges about “teaching” the test or asserting big gains based on very small improvements will not help education in the long run.
The September 14-15, 2000, CRESST annual conference, “Educational Accountability in the 21st Century,” focused on many key assessment issues confronting policymakers, researchers, and educators. Nearly 275 participants shared in more than 20 sessions with topics ranging from state assessment programs to current research on effectively assessing literacy skills. “The purpose of the conference is to present different perspectives,” observed CRESST Co-director Robert Linn, “while acknowledging that everyone is operating under a universal demand for accountability.” The challenge, he said, “is to communicate responsibility” for appropriate systems before accountability is written into concrete.

State Accountability Systems

The Texas accountability system—much in the news these days—provided a springboard for the opening session on state accountability systems. The Texas Assessment of Academic Skills (TAAS) is the result of a major campaign, according to John Stevens of the Texas Business and Education Coalition, gradually building up for more than a decade from minimal standards and testing to a system based on statewide data collection, improved standards, aligned assessments, and “consequences for persistent failure.”

The effort has remained “practical-minded,” said Stevens. He listed the essential characteristics of TAAS including annual testing, disaggregation of data, consistency over time, open access to data, and increasingly informed decision making “from statehouse to the school.” As indicators of success, Stevens noted that (a) TAAS scores have improved considerably despite higher performance levels; (b) ninth-grade dropout rates have improved over the past three years; and (c) school districts are using the results to select teachers for instructional leadership. The system is not perfect, Stevens admitted, “but I really don’t think another state can document these kinds of results for kids.” Furthermore, it is still evolving.

Two researchers presented the “however” of accountability systems such as that in Texas. Analyses of current accountability systems generally find things not as rosy as pictured,” said Daniel Koretz of CRESST/RAND. Accountability systems “have made a real retreat” from an emphasis on the quality of schooling. Instead, they are done “on the cheap,” lead to score inflation because of the emphasis on scores, produce incorrect attribution of causes for score increases, and often involve very high misclassification rates of students.

The new agenda for research on accountability at CRESST, Koretz explained, is to develop systems “that capture more of the benefits and create few unintended negative effects.” He presented a CRESST goal, later discussed by Co-directors Baker and Linn, to develop criteria for accountability systems. Koretz explained that the research should include technical work on aligning assessments with standards, better statistical methods for evaluating score gains, more work on value-added modeling, investigation of score inflation, development of better evaluations of best practice, and designs for better incentive systems.

Brian Stecher of CRESST/RAND also addressed the improvement of district and state accountability systems on the same panel discussion. He said that standards-based accountability has led to positive changes in teaching, such as Vermont teachers who gave more attention to problem solving when it became a core of student portfolios. The research, however, also suggests many shortcomings. For example, teachers tend to focus on the content of the tests more than the standards. This leads to undesirable changes in practice, such as the neglect of subjects not tested.

Stecher offered a number of recommendations to reduce these shortcomings: assess broader content, increase the variety of subjects tested, use more matrix sampling, and change grade levels tested over time. He agreed with Koretz that ac-
accountability systems should move toward “measuring appropriate practice and creating more elaborate standards about what ought to be seen in classrooms.”

“Policymakers lack adequate technical knowledge about accountability systems,” Linn added, and the challenge will be “to explain it in language that does not put them off.”

Jane Armstrong from the Education Commission of the States (ECS) described how CRESST’s work on developing criteria for accountability systems will be integrated with efforts at ECS to engage policymakers in accountability standards. She noted, for example, that the commission’s clearinghouse and new Web site will link research to policymakers. “Over time,” she said, “we could develop multiple models for states using CRESST accountability standards.”

Other conference presentations covered a diverse range of topics, including:

- the use of technology to assess and report student progress;
- inclusion into assessments of children with disabilities and children who are English language learners;
- benchmarking and alignment of standards and assessment;
- collaborative problem solving;
- rewards in high-stakes accountability systems;
- improved methodology for assessment and evaluation; and
- practical aspects of assessment-driven school reform.

Remaining plenary sessions focused on evaluating teaching quality, assessment of reading proficiency, and assessment at the national level.

**Assessing Teacher Quality**

State policy agendas only recently tackled the thorny problem of rigorously assessing teacher quality and teaching, according to Michael Knapp, director of the Center for the Study of Teaching and Policy at the University of Washington. In the past, teacher evaluation relied on time-honored approaches such as certification, aggregate test scores of teacher candidates, scores on teacher exams, principals’ annual visits, or the “marketplace,” in which parents dissatisfied with the teaching “vote with their feet.” States now use a number of other strategies, including:

- articulating explicit standards for teachers and teaching itself (in about half of the states).
• strengthening assessment of teachers as they enter and exit preparation programs;
• generating intermediate, tiered certification systems with performance-based demonstration of teaching (in a few states);
• tracking down uncertified teachers or weeding out incompetent teachers through a loosening up of tenure policies;
• beefing up on-the-job supervision such as peer review and real principal engagement in making teaching better; and
• promoting certification for accomplished practice (more than half of the states).

“We don’t know if these processes separately or in combination will do the trick,” Knapp said. “They represent both support and punitive actions, but the tendency among states right now is toward the punitive.” The field needs “a better understanding of the link between teachers’ instruction and student achievement, and between state policy action and what happens locally,” he said.

CRESST/University of Colorado at Boulder partner Lorrie Shepard addressed one approach to evaluating teacher quality that is growing in popularity, the value-added accountability model. Created first in Tennessee by William Sanders, the value-added model uses gains in student test scores between grades to evaluate teacher quality. However, Shepard said that even this popular model has some limitations whose importance will not be known until there is more research with simulations and real data sets. Among the concerns so far are these:

• The assessments are generally limited to multiple-choice format or to a commercially available test with only a limited range of test formats. Such tests frequently lead to test score inflation and narrowing of the curriculum.
• The analysis adjusts for initial differences in student achievement but does so imperfectly, leaving considerable potential for bias in estimating and reporting teacher and school effects. Missing data, for example, are estimated, thereby contributing to the error in the system.
• Anecdotal evidence suggests that teacher results can change dramatically when teachers change districts, suggesting that it matters who their students are and with whom teachers are compared.

...“information from a powerful and provocative assignment is the most powerful indicator” of teacher quality.

Professor Lloyd Bond of the University of North Carolina, Greensboro, contended that “information from a powerful and provocative assignment is the most powerful indicator” of teacher quality. He based his judgement on the lessons learned from the teacher assessments used by the National Board for Professional Teaching Standards. “Assessments must be grounded in what teachers actually do,” he said.

Recognizing the high costs involved in quality teacher assessment and certification, Bond noted that good teaching is hard. Consequently, we should not expect that evaluating teachers is going to be an easy job or inexpensive.

Megan Franke of the UCLA Graduate School of Education & Information Studies reported on a teacher research project conducted in collaboration with the University of Wisconsin, Madison. The researchers collected longitudinal data to investigate the effects of a project designed to improve mathematics instruction.

“Teachers who were most successful were those who had learned more since we left, not those who did best while in the project,” reported Franke. “Those teachers who continued to grow consistently demonstrated detailed knowledge about what their children knew about math.” Secondly, the findings persuaded the researchers to consider the different experiences teachers have as learners themselves—individually, in work groups, and as part of a school community.

The teacher quality panelists were asked whether an accountability system could be more teacher friendly. “As long as accountability is tied to student achievement and achievement is mea-
sured the way it is, I can’t think of a teacher-friendly assessment system,” Bond said.

Knapp pointed out that accountability systems are not designed to be teacher friendly, although the National Board certification process can help teachers internalize a set of norms about their work.

Franke said there needs to be a way to recognize what teachers are doing and are capable of, “and build from there.”

**Assessing Reading Achievement**

As with teacher evaluation, the teaching and assessment of reading has drawn national attention. In response, the Federal Reading Excellence Act marks a new approach to federal policy by defining what is meant by reading and establishing guidelines for assessment. The Act sets out, basically, “to change early reading instruction,” explained Joseph Conaty, who is responsible for the Reading Excellence Program at the U.S. Department of Education. The program has efforts underway in 27 states, primarily supporting low-income schools, and is guided by a basic set of principles that Conaty discussed.

- Young children, especially those at risk, should be assessed often, both formally and informally.
- The purpose of the assessment should be to identify children in need of improvement and at risk of failure.
- The assessment should not penalize young children.
- Assessment information should be used to enhance teachers’ knowledge of the best instructional practices for each child.
- The assessments should be part of an accountability system that has consequences for adults.

Unfortunately, accurate identification of young children with reading needs has often been hampered by a lack of dependable measurements. Another problem is that many early reading assessments require time-consuming evaluation of all children. Conference presenter David Francis, director of the Texas Institute for Measurement, Evaluation, and Statistics at the University of Houston, described the Texas Primary Reading Inventory (TPRI). Based on data from a longitudinal study of early readers, the inventory is efficient because it screens out those children who are clearly on target as good readers. Teachers can then assess in depth the remaining children, who are likely to be at risk of early reading difficulties. Now used in more than 80% of Texas’ elementary schools, the TPRI has led to an elaborate system of professional development and given teachers more time to focus on students in need of extra help.

Catherine Snow, Harvard University, presented findings from research conducted by the RAND Reading Study Group. The members, said Snow, are developing a framework for “a top-of-the-line reading assessment system and a research agenda for it.”

The group approached the problem by first defining reading comprehension. The key components include knowledge (getting the gist), application (using newly acquired knowledge in further comprehension), and involvement (aesthetic reaction such as getting lost in a story or feeling like a good reader). The assessment system must show the level of students’ reading skills linked with appropriate interventions such as vocabulary development, engagement and motivation, discourse knowledge, or fluency.

Snow pointed out that an effective assessment system must be able to measure reading levels in a student’s native language. Knowing how well a child can read in any language will help us build on his or her existing skills.

**Assessment at the National Level**

A final conference plenary session addressed the role of increased accountability in federal policymaking. Presenter George Bohrnstedt of the American Institutes for Research mentioned that both presidential candidates wanted to expand the use of the National Assessment of Educational Progress (NAEP). Governor Bush’s accountability plan would require any state receiving federal education funding to participate in NAEP, and Vice President Gore would provide financial rewards to states that show progress on the NAEP.
tests. Bohrnstedt discussed possible misuses of NAEP if it were used for high-stakes purposes including increased testing burden, potential corruption of the test, and possible efforts to exclude special needs students in order to increase state NAEP scores. He also discussed reasons why NAEP is an effective national assessment in its current form, including its high level of validity, state-of-the-art psychometrics, and usefulness as a state-by-state monitoring tool.

The first component of a national system of accountability, according to Edmund Gordon, professor of psychology emeritus of Yale University, “would have to be responsibility for universal access to the essentials for academic development.” In his view, the essentials would include good health, human capital development, social capital, and polity capital (a sense of belonging to society). Without these, “it may be immoral for us to put accountability on school outcomes.” Gordon noted that middle-class Black and Hispanic students are not underachieving because they do not know what the standards are. Rather, they are trying to achieve in an environment that is not responsible for the delivery of universal access to the resources and capital they need to succeed.

Addressing the issue of whether assessment should focus on learning or on accountability, Robert Mislevy (CRESST/Educational Testing Service) came down on the side of the former. “What we really want to know about is what students know, can do, and should work on next.” Rather than start with an assessment and figure out its purposes, he said, “we should start with the purposes and the kind of learning we want to know about.”

Mislevy also pointed out that even if a good system for tracking student progress is available, “it cannot fully capture all the things you would like to know about students. No matter how big the data file, there are things you will not be able to know.” Mislevy used evidence from the Advanced Placement Studio Art portfolio assessment to illustrate his point of view.

Martin Orland (U.S. Department of Education, Office of Educational Research and Improvement) agreed that federal pressure would continue to increase accountability. His analyses suggest an increasing reliance on outcome-based performance measures, tests, and assessment systems as central policy levers but much less certainty about how these instruments will be constructed and used to foster educational improvement. He noted as well the increased pressure to ratchet up the scientific quality of educational research in terms of experimental or quasi-experimental designs.

CONCLUSIONS

In summary, Baker observed that the whole conference had been about validity—Are we likely to make valid decisions about the education system because of assessments and accountability? “Among the several priorities for CRESST in the immediate future is to find more descriptive ways of mapping school learning,” she said, explaining that “if we could get a small handle on that, it could help us understand” the role of assessment and accountability.

The weight of high-stakes assessments and accountability is “enormous,” Joan Herman, newly named co-director of CRESST, said in conclusion, “but over the last few days we have learned that we know a great deal about the issues. I am hopeful that we can move ahead and do work that is useful.”

Overheads from many presentations made at the 2000 CRESST Conference may be found on the CRESST Web site at www.cse.ucla.edu.
RIGHT ASSESSMENTS AT THE RIGHT TIME

By implication, we might reconsider whether using broad-based, transfer-oriented tests at the start of an accountability program is our best choice. We might consider in the early stages of such systems the use of high-stakes measures on tasks that can be well explained, well sampled, and effectively and legitimately taught. Only later would we move to broader tests that measure transfer. The validity of conclusions we draw about the power of educational reform or the effectiveness of specific efforts to “close the gap” may depend upon the right emphasis on the right assessments at the right time.

IN MEMORIAM—RICHARD JAEGER

The measurement and education community mourns the loss of Richard Jaeger, 62, who passed away on October 21 after complications from a lung transplant. A well-known researcher and professor emeritus at the University of North Carolina, Greensboro, Professor Jaeger was an international scholar who made important contributions in survey sampling and techniques for setting educational standards. Prior to his retirement, he was the director of the Center for Educational Research and Evaluation at UNC and the director of the technical analysis group for the National Board for Professional Teaching Standards.

“Dick was expert both in technical testing topics and in the important realm of genuine compassion. He was a good friend to everyone interested in education,” said CRESST co-director, Eva Baker.

“Dick Jaeger was a giant in the field of educational measurement and a wonderful colleague who inspired better work and actions from colleagues and students,” said Robert Linn, CRESST co-director. “He was committed to educational equity and contributed much to that cause,” Linn added.

Professor Jaeger’s family has requested that contributions be made to the American Lung Association, 3409-D West Wendover Avenue, Greensboro, NC 27407.