California’s Student Testing System: Hard choices and new directions

California is struggling to create and implement an assessment system that would support standards-based reform and would be acceptable to all stakeholders—students, parents, educators, policymakers, education researchers, and business and community members. At EdSource’s 24th Annual Forum in April 2001, “Tests and More Tests: The Road Ahead for Student Assessment,” speakers focused on the issues central to that challenge.

Gerald Hayward, director of Policy Analysis for California Education (PACE), who moderated the EdSource event, discussed the differing perspectives. “The researchers and testing people are really pushing hard for accuracy, for limiting consequences to what they feel the tests can accurately measure. That usually means going slowly and cautiously, and often having long and expensive and multiple tests,” he said. Business leaders, on the other hand, are “willing to trade off some accuracy for speed. They would argue that we’ve dithered long enough.”

Educators, Hayward added, want the testing system to be fair. They want clear signals, lead time, curricula and textbooks aligned with the state’s academic content standards, a competent teaching force, and adequate preparation.

“When you begin to see some of the issues that policymakers grapple with, and some of the very legitimate points of view often coming at them from different directions, you’ll appreciate the complexity of this issue,” Hayward said. Out of these competing demands “emerges an assessment system that probably pleases no one, but may survive if we are explicit about what we are doing and why.”

State Board director outlines new direction

At the April Forum John Mockler, executive director of the State Board of Education, took on the challenge of clarifying where state leaders are planning to go with California’s mandatory assessment system.

When Gov. Gray Davis, a new state Legislature, and a new State Board of Education took over in January 1999, they brought a fresh look to California’s education system, Mockler said. Twenty-seven months and $12 billion in additional programs and services later, the state is well on its way to helping its students meet its new, higher standards, he said. Teacher training, professional development, curriculum, and textbooks are all being aligned to the state’s academic content standards.

Mockler admitted that this fast pace plus a lack of communication has left many educators who must implement the new standards confused. But, he said, the state is clear in its direction and is working to improve its communication with educators. To measure the impact of this outpouring of time and money, the state needed to move quickly with its assessment system, Mockler said. In 1999, the system did not yet include tests based on the state’s academic content standards.

“We had no way of measuring the system against itself,” he said. “We had no system to recognize those schools that made extraordinary progress. We had no goals, no expectations. We always talk about schools having higher expectations for students; we also have to have higher expectations for schools. I think they work together.”
STAR assessment program faces reauthorization this year

A year ago, Mockler said, the State Board of Education and the Superintendent of Public Instruction agreed on recommendations for a long-term assessment plan. State law requires the reauthorization of the Standardized Testing and Reporting (STAR) program this calendar year, giving policymakers the opportunity to put that long-term plan into law. The goal is for state tests to conform to high technical standards and focus instruction on the mastery of the state content standards, Mockler said. This would mean moving away from the current reliance on a nationally normed test such as the Stanford-9 (a multiple-choice, “basic skills” test that compares California students to a national sampling of students) toward a system that relies primarily on standards-based tests.

Currently the STAR program includes the Stanford-9 plus some tests that are directly aligned with California’s content standards. However, for 2000–01, the results of the Stanford-9 tests alone determined a school’s Academic Performance Index (API) score. California’s new accountability system ranks and rewards schools based on their API score and sets goals for each school based on raising that score. The state also finances intervention programs for schools with low API scores.

Mockler predicted that the standards-based tests, rather than the Stanford-9, would become the core of the state accountability system in the future. The API will “rapidly become a standards-driven index,” he said.

Plan emphasizes standards-based assessments and streamlining of tests

Mockler gave a sneak preview of this new plan, which he hopes will eventually create a seamless, standards-based assessment and accountability system from second grade through university.

“We will eliminate redundant tests,” Mockler said, “and we will provide continuity over time. That’s the core direction we’re going.”

The new direction includes the standards-based High School Exit Exam; a nationally normed test of basic skills, such as the Stanford-9; the Golden State Exams; the California English Language Development Test; and some changes to the standards-based tests, which are in four core academic subjects (English/language arts, math, science, and history/social science). (See Figure 1.)

- In English/language arts, all students in grades 2–11 will take augmented standards-based and nationally normed tests as they are doing today, according to Mockler. The writing assessments in grades 4 and 7 are expected to continue, he added.

- In math, all students in grades 2–11 currently take a nationally normed test, the Stanford-9, as well as augmented standards-based tests. In the future, the nationally normed test would be shortened, Mockler said. Stanford-9 questions now included in the standards-based tests would be removed so the two tests would be entirely separate, he predicted. “This system will save us time and provide us a lot more room to make the tests as good as they can be,” he said.

For students taking Algebra I or above, the standards-based content test currently is linked to the course the students are taking. When a student takes Algebra I, for example, the standards-based test covers Algebra I. Thus standards-based tests act as end-of-course tests, Mockler said. This approach was implemented in 2000 and is part of the proposed STAR reauthorization plan, he said. In grade 11, students who have completed the sequence of math courses (Algebra I, Geometry, and Algebra II or the Integrated Math series) would continue to complete a cumulative standards-based test as well as a nationally normed test, as they do today, he said. Under the new plan, high school students not enrolled in those specific courses would probably take a standards-based test on material up to and including Algebra I, very similar to the range of the High School Exit Exam, Mockler said.

- In science, Mockler said, the future plan calls for a standards-based test in at least one upper elementary grade, which is not currently required. Grades 9–11 would take a shortened form of a nationally
normed test, he predicted. (Currently students in grades 9–11 take the full Stanford-9 science test.) High school students who are enrolled in specific science courses (biology, chemistry, physics, or earth science) would also continue to take a standards-based test tied to a particular course, so the tests would act as end-of-course tests.

In **history/social science**, students would continue to take standards-based tests in grades 9–11 as is currently done, Mockler said. The future plan, he noted, calls for at least one test in the upper elementary grades, which is not done today. “We will eliminate the nationally normed test for history/social science in high school,” he predicted, because that test is the least synchronized to what California does. “It tests course knowledge in grades that are out of sequence with our standards, and so it really has no connection,” he said.

In the future, some of the **Golden State Exams** would be merged with the standards-based tests to denote high achievement, Mockler predicted. (Golden State Exams are voluntary, end-of-course tests recognizing high student achievement.) “We will also investigate that merger and the use of those two tests [standards-based and Golden State] to serve as at least a post-secondary placement test,” he said, adding that eventually they might be used as admission criteria to post-secondary institutions. There is no reason, he said, why California cannot have its own tests that are equal to the Advanced Placement or International Baccalaureate tests.

The new **California English Language Development Test (CELDT)** would be used to track the progress of English learners toward proficiency in English, Mockler said. The SABE/2 (Spanish Assessment of Basic Education, Second Edition) will still be used to test Spanish-speaking students who have been in a school district for less

---

### Figure 1
**California’s Current and Future Testing Plans**

By law, STAR is scheduled to be reauthorized in 2001 and may, or may not, include the Stanford-9 as the nationally normed test. The items under “2002 and beyond” reflect recommendations from the State Board of Education for that reauthorization.

#### English/Language Arts

<table>
<thead>
<tr>
<th>Grades</th>
<th>2001</th>
<th>2002 and beyond (proposed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2–11</td>
<td>Full nationally normed test (Stanford-9)</td>
<td>No change</td>
</tr>
<tr>
<td></td>
<td>Augmented Standards-based test</td>
<td>No change</td>
</tr>
<tr>
<td>4 and 7</td>
<td>Writing assessment</td>
<td>No change</td>
</tr>
</tbody>
</table>

#### Math

<table>
<thead>
<tr>
<th>Grades</th>
<th>2001</th>
<th>2002 and beyond (proposed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2–11</td>
<td>Full nationally normed test (Stanford-9)</td>
<td>Shortened nationally normed test</td>
</tr>
<tr>
<td>2–7</td>
<td>Augmented Standards-based test</td>
<td>Standards-based test entirely separate from nationally normed test</td>
</tr>
<tr>
<td>8–11</td>
<td>End-of-course Standards-based tests (Algebra I/Geometry/Algebra II or Integrated Math I,II,III)</td>
<td>No change</td>
</tr>
<tr>
<td></td>
<td>For those who have completed only Algebra I, a Standards-based test through Algebra I</td>
<td></td>
</tr>
</tbody>
</table>

#### Science

<table>
<thead>
<tr>
<th>Grades</th>
<th>2001</th>
<th>2002 and beyond (proposed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>K–8</td>
<td>None required</td>
<td>Standards-based test at one grade between 5 and 8</td>
</tr>
<tr>
<td>9–11</td>
<td>Full nationally normed test (Stanford-9)</td>
<td>Shortened nationally normed test</td>
</tr>
<tr>
<td></td>
<td>End-of-course Standards-based tests (biology, chemistry, physics, earth science)</td>
<td>End-of-course Standards-based tests (biology, chemistry, physics, earth science)</td>
</tr>
</tbody>
</table>

#### History/Social Science

<table>
<thead>
<tr>
<th>Grades</th>
<th>2001</th>
<th>2002 and beyond (proposed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>K–8</td>
<td>None required</td>
<td>Standards-based test at one grade between 5 and 8</td>
</tr>
<tr>
<td>9–11</td>
<td>Full nationally normed test (Stanford-9)</td>
<td>Standards-based test only</td>
</tr>
<tr>
<td></td>
<td>Standards-based test</td>
<td>Standards-based test only</td>
</tr>
</tbody>
</table>
Speaker offers 10 “dilemmas” to consider

To put California’s testing effort in perspective, Moderator Gerald Hayward, director, Policy Analysis for California Education (PACE), offered a list of policy choices that must be considered and reconciled when creating an assessment system. His 10 “dilemmas” follow:

Race, primary language, and socioeconomic status: In an accountability system, how do educators balance the need to include children disadvantaged by race, language, or poverty with a need to be fair to schools with large numbers of those children?

Validity: Tests must be aligned with the standards to have meaning, but what about aspects of schooling that are not measurable by tests?

Reliability: Tests are snapshots. Will a snapshot on one day match the snapshot on another day? And what is the appropriate balance between simple but objective multiple-choice items and deeper but more subjective performance indicators such as essays?

Different audiences—different approaches: Policymakers care about the quality of education generally, parents care about their children’s education, and educators care about the child, the school, and information that will improve instruction. Can the tests satisfy all these concerns?

Criterion-referenced or normative, or both: What is the proper balance between criterion-referenced tests that measure performance against the state’s academic content standards and nationally normed tests that compare California students to other groups?

Nature of consequences and for whom: Tests have to matter to test-takers. But as consequences increase, the accuracy of the test becomes more important. If a school is closed down, a teacher fired, or a student kept from graduating, the burden on the test becomes huge. Can the tests carry the burden?

Outcome measures or process variables: An assessment system needs outcome measures such as test scores, but should not ignore process variables such as access to the course content being tested.

Stability or flexibility: Should a test’s content be relatively unchanging and include all the standards, sending a clear and unambiguous message? Or is it fairer, even if more confusing, to start easy and add difficulty later?

Cost: Costs can rise exponentially with the need for increasing accuracy. How much accuracy can California afford?

Time: What is the balance between increasing the number of items or tests in a search for accuracy and an overemphasis on testing?

Mockler said the state is working on establishing a passing (or cut) score on the new High School Exit Exam, which was taken for the first time by many high school freshmen in March 2001. (See a discussion of the exam on page 5.) “We won’t set [the cut score] on time because we don’t have the information to do a cut score in a proper fashion,” Mockler said. “It won’t be within the eight weeks allowed by law, but we’ll set that cut score.”

For standards-based tests to be used effectively, the state must set “performance levels” to evaluate student performance. The State Board of Education already has these levels for standards-based tests in English/language arts and plans to set them for math, science, and history/social science before the end of 2001, Mockler said. Students would be scored from a range of “far below basic” to “advanced.” The state has also approved using the English/language arts standards-based spring 2001 test results as part of the API, beginning in January 2002. Mockler predicted that math, science, and history/social science spring 2002 test results would be included in the API beginning in January 2003. The state would continue to have a nationally normed test to provide national comparisons, Mockler said, but this test would be reduced in size and importance in relation to the API.

New state standards are set very high

The state’s academic content standards are based on some of the toughest standards in the world, according to Scott Hill, California’s chief deputy superintendent of public instruction.

Hill, who was director of the Commission for the Establishment of Academic Content and Performance Standards, said the commission “took seriously the charge in the legislation (Assembly Bill 265, 1995) to develop standards comparable to this country’s global economic competitors.” The commission began the conversation for each content area, he said, by looking at:

Other states’ standards
National standards efforts
Earlier standards efforts in California
Other nations’ standards and curriculum

Regarding other countries, Hill said, “we looked far and wide, focusing especially on successful TIMSS [Third International Math and Science Study] nations, such as Japan and Singapore.” The group also reviewed the standards set by many European nations.

Mockler said that the public, including parents and educators, may be surprised at how poorly students initially
perform on standards-based assessments compared to their ranking on nationally normed tests, such as the Stanford-9.

He gave the example of the content standards for 10th grade. “If your student meets the content standards in the 10th grade, your student is [academically speaking] University of California eligible,” he said. “Essentially that means we’ve set our standards at a level in the 10th grade that fewer than 12% of our 12th graders now achieve. That’s really high.”

“Results of standards-based tests will show us a long way from our goal,” he said. “We have set our bar so high that it will take us a long time to get there.” And, he added, “explaining this to the public is going to be hard.”

High School Exit Exam reflects the state’s content standards

In March 2001, about 350,000 high school freshmen took the state’s new High School Exit Exam for the first time. Developed by the American Institutes for Research (AIR) for use in California alone, the exam is meant to measure knowledge of basic academic skills. It covers English/language arts through 10th grade and math through Algebra I. The test questions are based on the state’s academic content standards. It is a pass/fail test that students starting with the class of 2004 must pass to receive their high school diploma.

At the April EdSource Forum, George Bohrnstedt, senior vice president for research at AIR, described the rigorous review process involved in creating a test that meets the state’s academic content standards.

State creates test blueprints

AIR worked with the State Board of Education, which adopted a subset of the academic content standards for inclusion in what they called the “test blueprint.” They had separate blueprints for mathematics and English/language arts. (See box on this page.)

AIR developed, and the state review committees approved, content or item specifications for each of the standards that were in the state’s blueprint in the two subject areas. These content or item specifications were used to clarify and elaborate the standards, and then were put into a guideline for item writers. Writers had explicit instructions as to how to write items to those standards. This was a way to guarantee the content validity for the test.

These items were then reviewed and edited by the State Department of Education and by four state committees, including a community panel that read them for bias and cultural sensitivity.

“Each step of the way,” Bohrnstedt said, “we asked these groups to ensure that we were fairly representing the content and also that no subgroup was either advantaged or disadvantaged by a particular test item.”

In English/language arts, the High School Exit Exam test questions cover:

- word analysis,
- reading comprehension,
- literary response and analysis,
- writing strategies,
- writing conventions, and
- writing applications.

Math content strands cover:

- probability and statistics,
- number sense,
- algebra and functions,
- measurement and geometry,
- Algebra I, and
- mathematical reasoning.

AIR conducts field tests before the official March testing

Then AIR twice conducted field tests at high schools—one in the spring and once in the fall of 2000. AIR analyzed the data statistically to see how well the questions correlated with the specific content area they tested and whether the items operated in the same way for major subgroups.

“For example, we want to make sure that if an item is measuring statistics and probability for whites, that it’s measuring it for African Americans and for Hispanics as well,” he said.

In the field tests, multiple forms were used so that AIR could test a larger number of items. Eventually the test included 80 multiple-choice items for mathematics and 82 multiple-choice items plus two constructed response (essay) items for English/language arts.

Independent group evaluates exam

Human Resources Research Organization (HumRRO) conducted an independent evaluation of the exam, using surveys and questionnaires of teachers and administrators as well as direct observation of the scoring, Bohrnstedt said. They found there was a very close relationship between the items being used and the academic content standards that were being measured, he said. “The content validity of the test was high.” They also found that a high percentage of the items survived the field test—in other words, they were good enough to be used on the final test, he said.

However, HumRRO recommended postponing by a year or two which class would be the first required to pass the test in order to graduate. (Currently it is the class of 2004.) HumRRO made this recommendation, Bohrnstedt said, so that:
Students would have an opportunity to learn what is being tested, particularly Algebra I, which has never before been required in California for graduation. The report suggested that the state needed to be careful in this area to avoid potential lawsuits.

Teachers would have time to get the instructional materials and professional development they need to teach to the higher standards.

There would be time to develop a clear performance standard or cut score for passing, which has not yet been determined.

HumRRO also recommended showing moderation in setting the passing point. And the evaluators said the state should provide appropriate accommodations to special-needs students and to English language learners. Bohrnstedt said that the test administrators in March followed the Individual Education Programs (IEPs) of students who had them.

### Workshop grapples with passing scores

A workshop in mid-May tackled the issue of setting two passing scores, one for English/language arts and one for mathematics. The workshop included about 50 teachers, parents, educators, and business people on an English/language arts panel and another 50 from the same groups on a math panel, he added.

Based on the outcome of that standard-setting exercise, AIR is expected to make a recommendation to the California Department of Education, Bohrnstedt said. The plan is for the State Board of Education to determine the cut point for passing the exam at its meeting in June.

Bohrnstedt said that a concern will be the degree to which the cut point for the High School Exit Exam relates to the cut points on the STAR standards-based tests in English/language arts and math.

“This is something that’s going to have to be addressed over time to try to bring them into alignment,” he said, “so that we do move toward an integrated testing system in California.”

Bohrnstedt said the state plans to give the exam in spring 2002, and then multiple times in the following years. The CDE plans to administer the HSEE six times a year, three times for traditional schools and three times for year-round schools.

### Researcher explores criteria for a workable assessment system

In response to the descriptions of the state’s new student assessment system, Joan Herman, co-director, Center for Research on Evaluation, Standards, and Student Testing (CRESST), described what her organization believes makes a strong system and the extent to which California is meeting those criteria. She discussed the issue of alignment and other concepts—accuracy, fairness, inclusiveness, and consequences—at the EdSource Forum.

### Tests should reinforce standards

Ideally, Herman said, the state would start with the academic content standards—what the state expects all children to know—and then develop tests based on those standards. “We’ve reversed it a little bit in California, and now we’re going to have to catch up and transition, which is going to cause us a few problems,” she said. “But the idea is that the assessments become a primary vehicle for communicating what the standards really mean, because the tests make real...
what subject matter content counts and how well kids need to do. So the tests provide a signal to teachers and schools about what's important to teach and what students should be learning."

Besides providing information on how well students are doing, assessments should also provide diagnostic feedback, Herman said, so educators can determine better ways to teach.

**Accurate information is key**

Besides alignment to the state’s academic content standards, Herman said, sound and accurate information clearly is key. She reported on the work of CRESST colleague David Rogosa, who has examined the accuracy of Stanford-9 scores. Rogosa asks, “What is the likelihood that when a student is tested, the score that student receives will be considerably different from the student’s true capability?” In the extreme case, which was grade 8 mathematics, Rogosa found that there was only a 50% chance that a student whose true capability was at the 50th percentile would receive a score between the 40th and 60th percentiles. “It is essential,” Herman noted, “that we understand the margin of error in our scores before we make important decisions based on tests.”

Herman’s approach is reinforced by her colleague Robert Linn in CRESST Policy Brief 3 (Spring 2001). “To avoid improper use of test scores, states should report the probability that a student or school has been misclassified,” Linn wrote.

Test scores also tend to rise each year they are given as teachers and students become more familiar with the test. “Do the increasing scores represent true learning, or do they represent test preparation and gaming the test?” Herman asked. One way to tell, she said, is to track the scores every time a test form changes. Typically, she explained, scores drop when the new forms are introduced and then begin to rise in subsequent years as the test becomes familiar. (See Figure 2.)

“These bizarre relationships are one reason why most measurement people will tell you it’s not a good idea to give the same test form from year to year to year, or use exactly the same test items,” Herman said. “By changing test forms or changing the items, you prevent schools from over-focusing on the specific items that are on the test.” Changing test items also allows policymakers to broaden what is tested, giving better coverage of the state’s academic content standards, she said.

Herman also recommends validating any trends in test scores with results from other indicators, such as the National Assessment of Educational Progress (NAEP), other state or local tests, or results from college admissions and placement tests “that are intended to measure similar constructs.” That way the state can be more certain that any upward or downward trend reflects real changes in learning.

Another accuracy issue, Herman said, occurs when tests are asked to fulfill many different purposes, such as “assessing individual student progress; determining the quality of schools; evaluating curriculum; providing diagnostic feedback; assessing teachers or administrators; making promotion or placement decisions; etc. The type of test that is reliable and valid for one purpose may well not be reliable and valid for another purpose,” she said.

For example, a matrix-type test covers a wider range of knowledge because each child takes only one part of the test. Thus more items can be covered without causing test-taking fatigue. But matrix-type tests do not give individual scores, making it impossible to assess how well any particular child has learned the content.

“A single test simply cannot tell us what we need to know about students’ and schools’ accomplish-
ments and progress toward standards,” Herman warned. “Students and schools need multiple and diverse opportunities to show what is being learned.” Multiple-choice tests, for example, are not good indicators of whether a student has learned critical-thinking or problem-solving skills.

“California should not be using a single test for a ‘high stakes’ purpose such as graduation from high school or generously rewarding or severely punishing schools for performance,” Herman said. “The state should consider a two-tiered process, where the second tier independently confirms important decisions based on the first, because our assessment results simply are not as accurate as we give them credit for.”

**Fairness and inclusiveness are crucial**

Another important quality of a testing system is fairness, Herman said. Being fair includes developing assessments that are culturally sensitive, that avoid negative stereotypes, and that don’t show bias for various subgroups. But tests also need to be fair in the sense of enabling all children to show what they know, she said, which is a big issue for English language learners. “Testing them in English certainly places a few barriers on their ability to show their content knowledge in other subject areas,” she said.

Fairness also demands that all children have the opportunity to learn what is being assessed, Herman added. “Children don’t enter school on a level playing field,” she said. The socioeconomic status of a student accounts for 75% of the variance in school performance, she added. And poor kids are often in schools with the worst facilities, the fewest textbooks, and the least-trained teachers.

“Setting the bar so high that many underprivileged children cannot jump it will be counterproductive—certainly for the children whose life chances will be affected, but for the education system as well,” Herman said. “Many states are facing this problem and are considering slowing down the time when the test results ‘count’ to give teachers, parents, and students adequate time to prepare. Some are also planning to gradually increase the passing score over time. You don’t want to jump too quickly and leave too many children behind.”

However, students who face issues of poverty or language should not be excluded from the testing system, Herman said. “Because if they’re not included, they’re invisible; and if they’re invisible, their needs are ignored.” But, she added, it’s one thing to say everyone needs to be tested; it’s another thing to know how to do it in a valid and reliable way. For example, researchers do not know how to make tests with accommodations for students with disabilities or English language learners comparable, reliable, and valid, she said. “We don’t know what the results mean.”

One way to avoid negative consequences of testing such a group is to place more emphasis on school improvement than current performance, according to Linn in the CRESST policy brief. “This allows for differences in starting points while maintaining high standards and expectations for all.” California has used this approach with its accountability system.

**Consequences must be positive**

Finally, “we want the consequences of our assessment system to be positive,” Herman said. “We want it to help and not hurt children. If history is any guide, we better worry because we’ve had lots of testing systems with lots of good intentions.”

Assessment, Herman said, can start the education system moving in the right direction. “The good news is that teachers and schools indeed listen to the signal sent by large-scale assessment programs and try to respond accordingly. They modify their curriculum to focus on what they understand is important, and they use the assessment as a model for their own instruction.”

“The bad news, at least initially, is that they may overly focus on the test,” she said. “They don’t teach the subjects that aren’t tested, and they don’t teach [some] elements of tested subjects that don’t appear on the test. So it’s the test and not the standards that becomes the thing, which seriously narrows the curriculum and, by the way, kills the validity of the test. After all, we don’t really care how kids do on the 60 items or so in reading on the Stanford-9; we care because we think that test represents something larger.”

Another potentially negative consequence is the cost of tests—in time as well as money. “The better our assessment systems become, the more expensive they’re going to become, both in the amount of student time they take and in their costs,” she warned.

**System is becoming better aligned**

From Herman’s point of view, California’s glass is at least half full. “The system is more aligned than we give it credit for,” Herman said. “Teacher preparation is being brought in line with the standards. Professional development is being brought in line with the standards. Instructional resources are being brought in line with the standards. We have to find a way to help the schools understand that that indeed is the case.”

In the future, the state must make sure the tests “measure the depth and breadth of the standards,” include a number of assessments, and do not assign high stakes to inaccurate measurements, Herman said. Policymakers must also realize that the system is not perfect, she said.

“We need continuous scrutiny to make sure the assessment system doesn’t do damage and that it promotes the goals we want it to promote,” she concluded.
Panelists present multiple perspectives on testing

“Country people know that weighing the pig doesn’t make it fatter,” said John Mockler, executive director of the State Board of Education, at EdSource’s Forum. “But if you go out and buy a lot of food and the pig doesn’t get fatter, you get kind of upset.

“And so when 27 months ago we made a decision to invest an additional $12 billion in programs and services in public education over a three-year period, we had to know whether this food we were going to add was going to make that pig better for market.”

But the assessment system created by the state to measure this improvement effort has ignited controversy. What follows is a report on the wide variety of views on this issue, with comments from both the morning speakers and the afternoon panel of experts interspersed.

Is the system here to stay?

Peter Schrag, a columnist with the Sacramento Bee, noted the cyclical nature of education reform and warned that issues such as the energy crisis could derail current efforts. The school reform pendulum swings back and forth because of a basic disagreement about whether tougher standards and assessments work for all students, he said. “Our whole sensibility is divided between the desire for high standards and meritocratic demands and a democratic sense of giving people perpetual second chances and being open and democratic in access.”

But Scott Hill, the state’s chief deputy superintendent of public instruction, believes this pendulum has stopped—that school reforms are here to stay. He also said the assessment system is well on its way to reflecting the new academic content standards, the basis of the reform effort. “There is a broad consensus about where we think the system is going to go over the next few years,” he said, “the cornerstone of that being a movement to standards testing, shortening the amount of time we spend on norm-referenced testing, and then looking at a lot of consolidation efforts.”

Carl Cohn, superintendent, Long Beach Unified School District, is not so sure. “I just think Sacramento insiders talking to each other about becoming coherent in this is just Sacramento insiders talking. And it may not have any real impact on what goes on in the classroom.”

Suzanne Tachen, executive director, California Business for Education Excellence, disagreed with Cohn. She said the consensus reached beyond insiders, with lobbyists from school districts and teacher unions, for example, also being involved.

The current debate is not about whether to test or even what to test, she said. “That’s a fundamental fight. The conversations that are currently happening in Sacramento are around the little pieces of how to implement a policy we have all agreed to.”

Are tests fair to all students?

Marilyn Whirry, National and California Teacher of the Year, touched on a topic that stirred debate: What about children who enter the school system disadvantaged by race, language, poverty, or parents’ education level?

Whirry has been a member of the National Assessment Governing Board in Washington, D.C., which oversees the National Assessment of Educational Progress (NAEP) test that measures student progress across the United States.

“If you could have been in my position over these past eight years and have seen some of the low expectations we have for some of our children across this country, you would praise standards,” she said. “Everywhere people are crying for higher standards.”

But, she added, NAEP test results show that “the top kids keep getting better and the bottom kids keep getting lower. We are not fulfilling our need for higher standards for all children. Our expectations cannot be just for the top kids but for all kids.”

Joan Herman, co-director, Center for Research on Evaluation, Standards, and Student Testing (CRESST), warned that standards-based testing can backfire for underprivileged children. “If we don’t watch out, high-stakes, standards-based assessment is liable to add to the problem rather than help to ameliorate it by leading to a shadow curriculum for traditionally underperforming kids,” Herman said. When teachers worry about kids not being able to pass a high-stakes test, the educators tend to “drill them ad nauseum” until they think the kids can pass, she said.
Tacheny suggested investing money in the system to mitigate the impacts of socioeconomic issues. But, she said, assessment is key in determining how to invest.

“We haven’t been able to measure how kids are performing and the investment against that to see if it’s adequate,” she said. “We can use assessment to hold the system accountable and make sure that if you do come to school poor, tired, and hungry, you still have a shot.”

Mockler said the current school ranking system based on the Academic Performance Index (API)—which ranks schools from 1 (low performing) to 10 (high performing)—could give the state some clues.

“We know that poor kids and kids that don’t speak English tend to score low,” he said. “But we also know that there are about 20 schools in California that should be in the lowest API decile because they have poor kids, kids that don’t speak English, and immigrants. And yet those schools are up there in the five, six, seven, eight, nine, and even 10 API rank.

“So it informs us that it is possible to overcome the nature of our bias,” Mockler argued. “Our bias, fueled by long-time beliefs and results, is that there are certain kids who can’t learn. As I talk to people and hear the notion that poverty and linguistic circumstance impede instruction, I say to them that the parents are sending us the best kids they’ve got. They’re not holding the good ones back. And it’s our job, with the six or eight thousand bucks the taxpayers allow us to spend—perhaps not enough, but it is six or eight thousand bucks—to move those kids along.

“And so the assessment system informs us whether that is really possible, whether it’s a dream or some kind of a political statement, or it can be done. And when you look at the results of the assessment system, the answer is ‘yes.’ It’s not done enough, it’s not done everywhere, but it is done.”

Do students have a chance to learn?

Mockler touched on the school and teacher capacity issue—whether the support and resources are there to help students reach the new, more demanding goals. Other panelists also expressed their concerns about capacity.

Andrea Venezia, director, Stanford University K–16 Bridge Project, is skeptical that enough resources have been allocated. “It’s hard for me to get on board with strict accountability measures when there’s not enough capacity to meet what’s being asked,” she said. “I have concerns with access and equity issues for students. The concept of holding all students to high standards and ensuring that that’s happening at a local level is laudable, but we need to make sure that students of color aren’t being exempted by putting them into Special Education classes, that all students have access to Advanced Placement courses.

“There needs to be a focus on making sure that all students have access to high quality learning,” Venezia said. “How do you take the mechanisms that we have right now and put it together in a way that’s fair for teachers and other people at the school site to meet expectations?”

Whirry agreed that teachers need more support. “After traveling through the states for all these months, I’m more convinced than ever of the importance of teachers. I think our success and the success of any state depends on teachers. If I were in charge of professional development for this country, I would pour money into substantial retraining of all teachers, starting with early reading especially; and then I would work in other areas that teachers feel they need help in.”

Is the state communicating with educators about assessment?

Besides more support, teachers also need to be consulted more in the development of an assessment system, many panelists agreed. They criticized the state for keeping teachers too distant from the process.

Jeff Orlinsky, classroom biology teacher and chair of the California Teachers Association’s Assessment and Testing Committee, said teachers often see problems with tests—such as the amount of time allowed or children’s ability to focus long enough—but feel unable to communicate these issues because they sign an agreement not to discuss the test with anyone, including other teachers and administrators.

“There’s no avenue to send a message up,” he said.

Superintendent Cohn agreed. “There’s a huge disconnect right now between state policymakers and what is actually happening at the classroom level,” he said.

Columnist Schrag said there is nothing new about this disconnect. “There’s a rather substantial gap and always has been between the profession and the political system,” he said. “And the political system, and to some extent the business community, have driven school reform, and the profession has often been reluctantly (not always, and I don’t want to generalize) dragged along behind that.”

Whirry agreed there “is too much space between the teacher and the test,” but she encouraged teachers to speak up. “We have to stop being wimps as teachers,” she said. “If we are against this test, let’s find out what the test is, then let’s say this is why; because when we find out what it is, I don’t think we will be against it. There are problems, of course, but we must hit the base problem of teachers in the classroom afraid to speak out, afraid of what that accountability might be, afraid of what his or her students don’t know. I think we have to attack those questions.”

Are assessments hurting teaching?

Researcher Joan Herman is concerned that the current assessment atmosphere might discourage talented people from
entering the profession. “This high pressure environment may not attract and retain high quality professionals,” she said. “Who would want to live or work in such an environment, where you live and die based on a single test score?”

Mockler disagreed that anyone is living or dying by a single test score and pointed out that teacher retention is going up. “I think that people like to come to work where there is an agreed-upon focus, a common vision, where training is connected to the task, and where there are measurable results. If teaching those children more (as measured by a test) is somehow viewed as degrading, then that’s a real sad situation.”

“I don’t think it’s the case that teaching children more is degrading,” Herman countered. “I think it’s perhaps the limited measures that are used of one’s performance.”

Hill believes that accountability is “fundamentally a good thing. It motivates us; it disciplines us; it focuses us,” he said.

But Cohn said external pressures not perceived as meaningful or credible can sometimes make people dig in. “Can anybody think of something more empty than the threat of state takeover of a school?” he asked.

Tacheny argued for perspective, saying that although the teaching profession is going through rapid change, it is not alone. “High-tech manufacturers are experiencing the double whammy of the energy crisis and the bottom falling out of the stock market in Silicon Valley,” she said, referring to one person she knows whose stock dropped “pretty much overnight from $141 a share to $9 a share. That’s chaos. That’s high pressure. We forget sometimes that there are many things that are going on in the teaching profession is going through rapid change, it is not alone. “High-tech manufacturers are experiencing the double whammy of the energy crisis and the bottom falling out of the stock market in Silicon Valley,” she said, referring to one person she knows whose stock dropped “pretty much overnight from $141 a share to $9 a share. That’s chaos. That’s high pressure. We forget sometimes that there are many things that are going on

How many tests are too many?

Other panelists talked about the chaos they see in the current assessment system.

“You keep changing tests, you keep changing rules,” Orlinsky said. “You change rules within six months, even within days recently for the High School Exit Exam.”

Cohn said his district alone requires teachers to do 14 different assessments. “Because we don’t have confidence in the current state system, we think it’s very important to build these assessments that actually do, in fact, focus on the standards. But when you look at the burden on teachers—14 coming from us, all that’s coming from the state, and then whatever assessments they are doing as classroom teachers—you can see the real potential for meltdown, and why all of us who have positions of authority need to get at this issue.”

Both Mockler and Tacheny said the state is working to reduce the number of tests, particularly for the transition from high school to college. “There is agreement that there are way too many tests and that needs to be fixed,” Tacheny said. “There is fundamental consensus around that broad direction.”

Venezia said she was excited about University of California President Richard Atkinson’s proposal to eliminate the use of the SAT for admission and Mockler raising the possibility of looking at standards-based tests, Golden State Exams, or a combination to determine eligibility for college admissions.

“You add the 14 tests that Carl Cohn mentioned onto approximately 10 differing assessments in the transition between high school and college, and that doesn’t count the multiple Advanced Placement tests, the multiple SAT IIs—it’s just a tremendous testing burden,” Venezia said. “And there’s that balance between having only one indicator of readiness and having indicators that everyone wants to have for all their different personal reasons. So I’m excited about what California is thinking of doing.”

George Bohrnstedt—senior vice president for research, American Institutes for Research, which developed the High School Exit Exam—said the state needs to create more time for instruction by combining and eliminating exams so
that the testing system is the most efficient it can be. “I think we’re in a state of disarray at this point, but John Mockler gives us hope that maybe in three or four years we’ll be more efficiently testing, and testing the standards that we want to test.”

**Should tests have high stakes?**

How these tests are used was also brought up as a matter of concern by some panelists, particularly with “high-stakes” tests that have serious consequences, such as firing a teacher, closing a school, or preventing a student from graduating. Panelists stressed the frequent disconnect between the classroom and Sacramento.

Venezia said it is imperative that educators, students, and policymakers be on the same page. “It’s a little scary to me with high stakes now,” she said. “Everyone needs to be on the same page before the stakes get implemented, before people start drastically changing what they’re doing, and before students get terrified about what is going to happen to them when they’re trying to graduate.”

Herman said, based on Mockler’s speech, that there appears to be more coherency in the system than most people realize. “But it doesn’t feel that way when you get in schools,” she said. “I think that people are feeling like they’re getting jerked around, and the ground rules change from day to day. And people don’t feel good when they feel jerked around.”

Herman advocated a more gradual change in the testing program. “Our assessment results simply are not as accurate as we give them credit for,” she said. “Given that they’re so imperfect, we really ought to worry before we attach high stakes to the results of a single assessment. We need safety nets.”

Orlinsky also questioned the high-stakes use of tests, saying that the best use of a test is to inform classroom instruction.

“I learned a lesson in my second year of teaching a biology course when in class I would say: ‘You’re not doing as well as my first-period biology class.’ And finally one day a student stood up and said: ‘Why do you keep comparing us? I’m not in your first-period class. I am in your second-period class. I am learning what you’re teaching in second period.’

“There’s no need to compare,” Orlinsky said. “Give the test results to the teacher. They’ll use the results; they’ll do what they need to do to help those students. And we don’t have a police performance indicator. We don’t have a city performance indicator. It’s just teachers.”

Mockler agreed that the “issue of assessment is not just about how to blame. It’s also about how to inform. So it’s important that it’s accurate and focused, and we’re trying to make it that way.”

**Policymakers are facing high stakes too**

The state is clearly on its way toward implementing and aligning its assessments to its standards. But the high stakes involved for both schools and students mean that all eyes will be on how California’s leaders balance the many conflicting priorities as they craft the next edition of the state’s testing system—the STAR reauthorization in 2001.

As the speakers at the EdSource Forum made clear, policymakers face a variety of complex balancing acts:

- The need for accountability vs. fairness to disadvantaged students and the schools that serve them;
- The objectivity of a multiple-choice test vs. the depth of essay questions;
- The need to know how individual students are performing vs. how a school or the state as a whole performs;
- The cost in terms of time and money vs. the reliability of test results;
- The ability to compare California to the nation vs. holding students and educators accountable to the state’s own high standards.

But perhaps the most difficult balancing act of all will be determining the appropriate consequences of the assessments. How do policymakers convince educators and students to take the tests seriously without making the stakes too high in terms of fairness and reliability? Will the public support the idea that test results can determine whether a student graduates, a teacher is fired, or a school is shut down? As Forum Moderator Jerry Hayward asked: “Can the tests carry this burden?”