CPRE POLICY BRIEFS
Reporting on issues and research in education policy

Tracking Student Achievement in Science and Math: The Promise of State Assessment Programs

The National Science Foundation is investing millions of dollars in efforts to improve math, science, and technology education through its Statewide Systemic Initiatives Program. However, states are facing difficult challenges in accurately assessing the effects of these efforts on student learning.

Launched by NSF in 1990, the SSI program aims to address widespread concerns about student performance in science and mathematics in the United States, and related concerns about inequities in student performance associated with race, ethnicity, and gender. NSF solicited proposals from states willing to create broad-based coalitions to undertake comprehensive and coordinated reforms of science, math, and technology education to address these problems. NSF made 10 awards in 1991, 11 in 1992 and 5 in 1993.[1] The typical award was for as much as $10 million over a period of five years. The states have provided substantial matching funds to support their initiatives. The resulting state initiatives are often referred to collectively as the SSIs.

Naturally, given the magnitude of this investment in education reform, state and federal policymakers are anxious to know whether the SSIs are having the desired effects on student performance in mathematics and science. Are students learning more because of these state initiatives? Are they better equipped to apply what they are learning to everyday problems in the community and in the workplace? Are the inequities in performance among different groups of students being reduced? Answering these questions requires good methods of assessment.

This issue of *CPRE Policy Briefs* examines the capacity of state assessment systems to track the effects of the SSIs on student performance in mathematics and science. It also identifies some of the major issues state policymakers are facing as they attempt to re-align their state assessment systems to meet the changing goals for public education.[2]

The Importance of Appropriate Assessment

The SSIs are founded on ideas about good practice in mathematics and science education that are widely shared among education reformers and have gained considerable currency among teachers. These ideas about what constitutes good curriculum and instruction and what students should know and be able to do in the two disciplines have been promoted and disseminated by national and state professional groups. A "national vision" of good practice in mathematics seems to be emerging from the National Council of Teachers of Mathematics' *Standards*, and in science from the American Association for the Advancement of Science's *Benchmarks for Scientific Literacy*, the National Science Teachers' Associations' *The Content Core: A Guide for Curriculum Designers*, the National Research Council's *Draft National Science Education Standards* and related state documents.[3]

As described in these sources, good practice is characterized by challenging academic standards; a

RB-17-July 1995

CONSORTIUM FOR POLICY RESEARCH IN EDUCATION
CPRE
Rutgers, The State University of New Jersey
University of Wisconsin-Madison · Harvard University
Stanford University · University of Michigan

hands-on approach to instruction; active learning by students; the use of curricula and materials relevant to students' lives; an emphasis on thinking, problem-solving, and the application of knowledge; the integration of mathematics, science, and the use of technology in the classroom; and the use of assessment that is consistent with the more rigorous content and reinforces these instructional practices.

For example, in Montana, where the SSI has concentrated on developing and implementing new high school curricula, an observer saw the following mathematics lesson demonstrating the kind of instruction reformers are trying to promote:[4]

Michael is teaching students about the mathematics of exponential growth, using populations of animals as an example...[S]tudents working in small groups are asked to make up rules for the life and death of two populations of animals, represented by two types of candies shaken in a box.... A rule relating to proximity of candies might be a model for predation—for example if a large and small candy are less than an inch apart, take the smaller one out of the box (it dies) and put an additional larger one into the box.

Using the set of rules they create, they set the populations going for about 6 life-cycles, or 6 shakes of the box. The student groups then move to one of the eight computers in the room, recording their data in a computer spreadsheet, and producing a graph of the two populations over time, which they print out. Students must be able to produce mathematical expressions using exponential notation that match the growth or decline of these populations.

Assessing the effects of this kind of classroom practice on students' knowledge of mathematics and science requires appropriate assessment instruments. Conventional multiple-choice tests probably cannot adequately capture what students are learning in this Montana classroom. What are needed are assessment procedures that examine students' ability to explain important concepts, apply knowledge to new problems, and integrate information across fields. Appropriate assessment procedures might include essays or open-response questions, complex multiple-choice questions, performance tasks completed by individuals or groups of students, or portfolios of student work. Assessments such as these could generate the kind of student performance information needed to answer fundamental questions about the effectiveness of the changes in curriculum and classroom practice being encouraged by the SSIs.

The Policy Studies Associates Survey

Will state assessment programs provide the information about student performance needed to answer policymakers' questions about the effectiveness of the SSIs? To answer this question, NSF asked its national evaluation team in the spring of 1994 to examine the capacity of the SSI states to assess student performance in mathematics and science and provide good information about the effects of the SSIs.[5] One of the cooperating organizations, Policy Studies Associates (PSA) undertook a survey of the current assessment capacity in science and mathematics in the SSI states.

In the spring and summer of 1994, PSA interviewed state assessment staff and SSI staff in the 25 states.[6] Based on data collected through these interviews, the PSA study team examined the potential of existing state assessment systems to produce the data needed to track the effects of SSI efforts on student performance. Specifically, it asked whether achievement in mathematics and science was assessed statewide, and whether testing data were available at the appropriate grade levels, were reported for the appropriate unit of analysis, and were consistent for a sufficient period of time to be useful in evaluating the SSI. It also examined whether the states had assessment systems that were aligned with the learning goals embodied in the SSIs. Their findings, based on the status of state testing programs during the 1993-94 school year, are outlined below.

Assessment in the SSI States

What is the character of the existing state assessment programs in mathematics and science? Tables 1 and 2 summarize PSA's findings about these programs in the SSI states for the 1993-94 school year. The most obvious conclusion to be drawn from these data is that more states assess students in mathematics than in science. As Table 1 shows, all but 3 of the 25 states assessed students in mathematics. The three exceptions were Nebraska, Montana, and Colorado which had no statewide assessment systems at all. In contrast Table 2 shows that only 16 SSI states collected any data on student achievement in science. In addition to Nebraska, Montana and Colorado, 6 other states had no statewide science assessment.

Student performance data for science is less available primarily because science has not been considered a "basic" subject. State assessment programs had their origins in the basic skills movement of the 1970s and typically included reading, writing, and basic mathematics, but not science. While state assessment programs have raised their standards and added more challenging content in recent years, many still do not include science.

Table 1: Required Statewide Assessments as of 1993–94: Mathematics

State	Years of SSI Funding	Assessment Program Name	Grades Tested	Item Type	Scoring	Reporting Levels
Arkansas	93-98	Stanford Achievement Test-8 (SAT-8)	4,7,10	MC	N	D,S,I
		Arkansas Minimum Performance Test	3,6,8	MC	C	D,S,I
California	92-97	California Learning Assessment System	4,8,10	MC,P	PL	D,S
Colorado	93-98	No Statewide Assessment	—	—	—	—
Connecticut	91-96	Connecticut Mastery Test	4,6,8	MC,P	C	D,S,C,I
		Connecticut Academic Performance Test	10	MC,P	C	D,S,I
Delaware	91-96	Delaware Interim Assessment	3,5,8,10	MC,P	PL	D,S,I
Florida	91-96	Grade Ten Assessment Test	10	MC	N	D,S,I
		High School Competency Test	11	MC	C	D,S,I
Georgia	92-97	Iowa Test of Basic Skills/TAP	3,5,8,11	MC	N	D,S,I*
		Curriculum-Based Assessment	3,5,8,11	MC	C	D,S
Kentucky	92-97	Kentucky Instructional Results Information System	4,8,12	P,PF	PL	D,S,C
Louisiana	91-96	Louisiana Educational Assessment Program	3,5,7,10	MC	C	D,S,I
		California Achievement Test-5	4,6	MC	N	D,S,I
Maine	92-97	Maine Educational Assessment	4,8,11	MC,P	N	D,S,I
Massachusetts	92-97	Massachusetts Educational Assessment Program	4,8,10	MC,P	PL,N	D,S
Michigan	92-97	Michigan Educational Assessment Program	4,7,10	MC	C	D,S,C,I
Montana	91-96	No Statewide Assessment	—	—	—	—
Nebraska	91-96	No Statewide Assessment	—	—	—	—
New Jersey	93-98	Early Warning Test	8	MC,P	PL	D,S,I
		High School Proficiency Test	11	MC,P	C	D,S,I
New Mexico	92-97	Iowa Test of Basic Skills*	3,5,8	MC	N	D
		High School Competency Exam	10	MC	C	D
New York	93-98	Pupil Evaluation Program	4	MC,P	C	D,S
		Regents Compentency Tests**	9	MC,P	C	D,S,I
North Carolina	91-96	End-of-Grade Tests	3–8	MC,P	N	D,S
		End-of-Course Tests	9–12	MC,P	N	D,S
Ohio	91-96	State Proficiency Test	9,12	MC	C	D
Puerto Rico	92-97	APRENDA (Spanish version of SAT)	1–9	MC	N	R,S,I
South Carolina	93-98	Basic Skills Assessment Program	3,6,8,10	MC	C	D,S,C,I
		Stanford Achievement Test-8	4,5,7,9,11	MC	N	D,S,C,I
South Dakota	91-96	Stanford Achievement Test-8	4,8,11	MC	N	D
Texas	92-97	Texas Assessment of Academic Skills (TAAS)	3,8,10	MC	C	D,S,I
		TAAS—End-of Course	9–12	MC	C	D,S,I
Vermont	92-97	Portfolios	4,8	PF	PL	D
		Uniform Assessment	4,8	MC,P	PL	D
Virginia	92-97	Iowa Test of Basic Skills	4,8,11	MC	N	D,S
		Literacy Passport Test	6–8	MC	C	D

Key: Item Type: MC=Multiple Choice, P=Performance and Open-ended Items, PF=Portfolios, Projects, and Experiments
Scoring: N=Norm-referenced, C=Criterion-referenced, PL=Performance Levels
Reporting Levels: R=Region, D=District, S=School, C=Classroom, I=Individual

*At the 11th grade, Georgia uses the Test of Achievement and Proficiency (TAP). Students taking the TAP are matrix sampled so that scores cannot be reported at the individual student level. Item type and scoring remain the same.

**To receive a diploma, students must demonstrate competency in mathemathics by passing either the Regents Competency Test or a Regents examination in mathematics. The two assessments listed in this table are those likely to be most relevant for assessing the SSI.

Table 2: Required Statewide Assessments as of 1993–94: Science

State	Years of SSI Funding	Assessment Program Name	Grades Tested	Item Type	Scoring	Reporting Levels
Arkansas	93-98	Stanford Achievement Test-8 (SAT-8)	4,7,10	MC	N	D,S,I
		Arkansas Minimum Performance Test	3,6,8	MC	C	D,S,I
California	92-97	California Learning Assessment System	5,(8,10)	MC,P	PL	D,S
Colorado	93-98	No Statewide Assessment	—	—	—	—
Connecticut	91-96	Connecticut Academic Performance Test	10	MC,P	C	D,S,I
Delaware	91-96	No Statewide Assessment	—	—	—	—
Florida	91-96	No Statewide Assessment	—	—	—	—
Georgia	92-97	Curriculum-Based Assessment	3,5,8,11	MC	C	D,S
		Test of Achievement and Proficiency (TAP)	11	MC	N	D,S
Kentucky	92-97	Kentucky Instructional Results Information System	4,8,12	P,PF	PL	D,S,C
Louisiana	91-96	Louisiana Educationa Assessment Program	3,5,7,10	MC	C	D,S,I
		California Achievement Test-5	4,6	MC	N	D,S,I
Maine	92-97	Maine Educational Assessment	4,8,11	MC,P	N	D,S,I
Massachusetts	92-97	Massachusetts Educational Assessment Program	4,8,10	MC,P	PL,N	D,S
Michigan	92-97	Michigan Educational Assessment Program	4,7,10	MC	C	D,S,C,I
Montana	91-96	No Statewide Assessment	—	—	—	—
Nebraska	91-96	No Statewide Assessment	—	—	—	—
New Jersey	93-98	No Statewide Assessment	—	—	—	—
New Mexico	92-97	Iowa Test of Basic Skills	3,5,8	MC	N	D
		High School Competency Exam	10	MC	C	D
New York	93-98	Pupil Evaluation Program	3,6	MC	C	D,S,I
		Regents Compentency Tests**	9	MC,P	C	D,S,I
North Carolina	91-96	End-of-Grade Tests	3,6,8	MC	N	D,S
		End-of-Course Tests	9–12	MC	N	D,S
Ohio	91-96	No Statewide Assessment	—	—	—	—
Puerto Rico	92-97	No Statewide Assessment	—	—	—	—
South Carolina	93-98	Basic Skills Assessment Program	3,6,8	MC	C	D,S,C,I
South Dakota	91-96	Stanford Achievement Test-8	4,8,11	MC	N	D
Texas	92-97	TAAS—End-of Course	9–12	MC	C	D,S,I
Vermont	92-97	No Statewide Assessment	—	—	—	—
Virginia	92-97	Iowa Test of Basic Skills	4,8,11	MC	N	D,S

Key: Item Type: MC=Multiple Choice, P=Performance and Open-ended Items, PF=Portfolios, Projects, and Experiments
Scoring: N=Norm-referenced, C=Criterion-referenced, PL=Performance Levels
Reporting Levels: R=Region, D=District, S=School, C=Classroom, I=Individual

*The science portion of the ITBS is not required as part of New Mexico's statewide assessment program; however, state assessment staff report the 99 percent of the state's districts administer the test.

**To receive a diploma, students must demonstrate competency in science by passing either the Regents Competency Test or a Regents examination in science. The two assessments listed in this table are those likely to be most relevant for assessing the SSI.

Taking Stock of the SSIs

This is the second issue of *CPRE Policy Briefs* to examine the National Science Foundation's Statewide Systemic Initiatives Program. The first, "Reforming Science, Mathematics, and Technology Education: NSF's State Systemic Initiatives," provides an overview of the SSI program. It also describes the policy strategies and implementation problems being faced by the 25 state SSIs.

Limitations of State Assessment Data

Knowing that student assessment results in science and mathematics are available is not sufficient to determine their usefulness for evaluating the SSIs. The student performance data must also be consistent with the emerging standards in mathematics and science and fit the goals of the SSI. To determine whether the results of state assessment programs can be used to assess state initiated reforms in math and science education, policymakers need to address such questions as:

- Do the state tests assess rigorous content, problem-solving, and applications of knowledge in the disciplines?

- How well-aligned is the content of the state tests with the curriculum and instruction promoted by the SSI?

- Are students tested at the grade levels targeted by the SSI?

- Does the state report scores at the appropriate level of the educational system (e.g., if the reform effort concentrates on schools, does the state report student achievement data at the school level)?

- Does the state administer tests in mathematics and/or science that yield comparable data statewide?

- Will the state's assessment program remain consistent for long enough to allow for its use in evaluating effects of reform efforts on student achievement over time?

The Character of State Assessment. PSA found that despite the considerable national fanfare about the introduction of portfolios and performance-based assessments and the raising of academic standards, state testing programs were still relying heavily on multiple-choice formats and were focused on basic skills. As the tables show, the assessments included a mix of item types. Tests composed solely of multiple-choice items were the most common. In mathematics, 12 of the 25 SSI states were using multiple-choice tests exclusively, while 9 others had at least one test that combined multiple-choice items with open-ended questions or performance events. In science, 10 states were using tests that were solely multiple-choice while 5 were using tests with mixed formats. Only 2 states, Vermont and Kentucky, were relying primarily on performance-based assessments, and Vermont was doing so only in mathematics. Clearly, state assessment programs had not yet abandoned more traditional forms of testing.

Yet, fundamental changes in state assessment policies and practices have taken place in states such as California, Connecticut, Kentucky and Vermont. In these states, performance-based assessment systems have been instituted as part of their broader reform strategies. However, with the possible exception of Connecticut where new forms of assessment have been introduced incrementally, these new approaches to assessment are surrounded by controversy. Public concern about the character of the new assessment system in California produced political turmoil that led to its suspension, and, probably will result in its replacement with a more conventional assessment system. And critical studies have raised questions about the technical adequacy of the new system in Kentucky and generated public debate about the future of the state's assessment program. (The sidebars on pages 6, 7, and 8 highlight the cutting-edge assessment systems being used in Connecticut, Kentucky, and Vermont.)

Meanwhile, other states like Arkansas, Delaware, and Massachusetts are developing new performance-assessment systems. In Arkansas, a 1991 education reform act eliminated the state's Minimum Performance Test and replaced it with a 12th-grade exit exam to be aligned with the state's new curriculum frameworks and learner outcomes. In addition, the state is instituting benchmark exams at grades 4 and 8 and is piloting a portfolio assessment. Delaware adopted an Interim Assessment Program in 1992-93 to be used until work on curriculum frameworks and a new performance-based system could be completed. The Interim Assessment, obtained from a commercial testing company, includes open-response items. In Massachusetts, the 1993 educational reform act called for phasing out the state's testing program and developing curriculum frameworks and coordinated assessments, including a "competency determination" as a requirement for high school graduation.

Some states are revising existing assessment programs. Michigan, which recently upgraded its mathematics assessment to make it consistent with the NCTM standards, is also upgrading its science assessment for 1995. Louisiana plans to revise its existing state assessments in mathematics and science after new curriculum frameworks are completed. North Carolina is revising and expanding its end-of-grade assessments in mathematics and science.

Another group of states including Florida, New Jersey, South Caro-

Connecticut's State Assessment Program

Connecticut, a state with a strong tradition of local control, has used its state assessment program to stimulate changes in local curricula. Connecticut has a two-pronged approach to assessment: the Connecticut Mastery Test (CMT) and the Connecticut Academic Performance Test (CAPT), both developed by the state. The CMT is a criterion-referenced test in mathematics and language arts, given every fall to students in grades 4, 6, and 8. First administered in 1985, the CMT was revised for 1994-95 to incorporate more problem-solving and higher-order skills, primarily by using more demanding and realistic open-ended questions.

Implemented in 1994, the CAPT assesses mathematics, science and language arts and is given every spring to students in grade 10. The CAPT includes multiple-choice and performance items. As part of the science assessment, for example, students are given a hands-on laboratory activity before the administration of the test and then are asked follow-up questions that relate directly to the hands-on task.

The CAPT also includes an interdisciplinary assessment section, in which students read and analyze a set of source materials that focus on "a controversial topic of social significance." Students then write a persuasive piece, using the source materials, to support their position on the issue. CAPT scores are included in students' transcripts.

The results of these state assessments are reported at the individual, classroom, school, district, and state levels. They also are included in school profiles that are published annually by the Department of Education and receive broad media coverage.

lina, and Virginia appear to be content to rely on their basic skills tests. A few states like Montana and Nebraska may never create statewide assessment systems because of strong traditions of local control and public concern about content and costs.

In addition to differences in the types of assessment items used, state tests also differ sharply in their scoring methods. Norm-referenced tests are constructed to compare achievement across large groups of students and are not linked to specific school or district curricula. Criterion-referenced tests, on the other hand, measure student mastery of specific curriculum objectives and provide valid performance comparisons only across schools and districts using those objectives. They also can be used to assess the quality and effectiveness of a curriculum's implementation. Both types of tests are used widely in state assessment programs.

Of the 25 states studied, 12 were using at least one mathematics testing program that was norm-referenced; 13 were using at least one testing program that was criterion-referenced. In science, the ratio was approximately the same. Few states report scores in terms of student performance levels, a method of scoring associated with newer, performance-based tests which require students to apply knowledge by solving a problem or conducting an experiment. Six states report at least some student results in terms of performance levels in mathematics, and only three do so in science.

Content Alignment. The PSA study found that the degree of content alignment between state assessments and SSI goals for curriculum and instruction ranged widely. At the high end of alignment are those SSIs which are integral parts of broader reform efforts. In Kentucky, for example, many individuals participating in the state's SSI also play active roles on state curriculum and assessment committees. SSI staff see the state's new assessment system as closely aligned with the curriculum and instruction they wish to promote and have designed the SSI to support its implementation.

At the other end of the scale, there is weak alignment, and changes in state assessment programs are lagging far behind other reform efforts. For example, while it is clear that state tests designed to measure minimum competency cannot capture outcomes in mathematics and science that the SSIs are attempting to promote, states such as Florida, New Jersey, New Mexico, and Virginia continue to administer them. In Virginia, for example, the SSI's definition of math and science achievement includes demonstrated student understanding of conceptual connections in mathematics and science; inquiry, problem-solving and decision-making; and the value of mathematics and science. However, Virginia's primary state assessment instrument, the Iowa Test of Basic Skills, cannot measures the SSI's success in promoting these outcomes because it focuses mainly on achievement in basic skills.

Similarly Puerto Rico's multiple-choice, norm-referenced APRENDA testing program does not measure the higher-order thinking skills the SSI is promoting in its middle and elementary schools. The commonwealth's SSI has overcome

Kentucky's State Assessment Program

Adopted in 1990, the Kentucky Instructional Results Information System (KIRIS), is designed to provide data for both school accountability and instructional improvement. KIRIS is a high-stakes, performance-based assessment system keyed to four student proficiency levels; it is aligned with a new state curriculum framework based on the state's six learning goals and 57 academic expectations. The KIRIS assessment has three parts:

1. Open-response items covering reading, math, social studies, science, arts and humanities, and practical living/vocational studies are included in the assessment. Open-ended questions in each subject tap multiple areas of content knowledge and require several steps to answer. The test allocates approximately 15 minutes for each open-ended response.

2. Students must also complete a one-hour performance event in math, science, social studies, arts and humanities, or practical living/vocational studies. The students work on an experiment or problem-solving activity in small groups or pairs and write their answers individually.

3. Students also submit portfolios containing five to seven examples of their best work in both writing and mathematics. Mathematics portfolios are submitted in grades 5, 8, and 12; writing portfolios are submitted in grades 4, 8 and 12. High school portfolios are under development. Scores from the mathematics portfolios, assessed statewide for the first time in 1992-93, will be incorporated into the accountability index starting 1994-95. Writing portfolio scores account for 14 percent of the cognitive part of the accountability index.

Various components of this assessment system are administered in grades 4, 5, 8, 11, and 12. Results are reported as the percentages of students attaining the four proficiency levels. KIRIS results are combined with other indicators including attendance and graduation rates to produce a score for each school on an accountability index. Each school is given a target, based on its previous baseline, for improving its score on the index over a two-year period. Schools receive rewards on assistance depending on their progress.

this deficiency by developing its own assessments. New Jersey's Early Warning Test and High School Proficiency Test also assess minimum competency, not the kinds of higher-order, performance-based learning outcomes for which the state's SSI strives. Conversely, in these states the SSIs may push for reforms in curriculum, pedagogy, and classroom assessment that may do little to ensure that student scores increase on more traditional state tests.

Grade-Level Coverage. Most SSIs are K–12 initiatives. However, of the 22 states with statewide mathematics assessments, 4 lack adequate coverage to evaluate K-12 state systemic initiatives; that is, they do not test students in one or more of the elementary-, middle- and high-school levels. Ohio and Florida provide statewide testing only at the high-school level, Vermont currently assesses students only in grades 4 and 8, and New Jersey administers a statewide test only in grades 8 and 11.[7] Of the 16 states with statewide science assessments, 4 fail to cover all three levels. Connecticut and Texas test students only at the high-school level, South Carolina only in grades 3, 6, and 8, and New York only in grades 4 and 9.

Reporting Levels. To be useful in evaluating an SSI, the analysis of the results of the statewide assessment must match the SSI's intervention strategy. If the strategy targets individual schools, an assessment program that reports student-achievement data only down to the district level—such as those in Ohio, New Mexico, and Vermont—will not help to isolate the impact of reform efforts. Reform efforts that target individual teachers also present special problems. In Arkansas, for example, the state reports student-achievement data at the district, school and individual student levels; but it may be impossible to sort scores by classrooms in order to evaluate effects of training on individual teachers. The task becomes even more difficult when teachers do not have self-contained classrooms, a common arrangement at the secondary-school level.

These assessment problems can be overcome. The Louisiana Department of Education has worked with the SSI to modify its testing forms so the state can link individual student scores with teachers who have received SSI training. But remedies require deliberate planning and add additional costs.

A related problem is that while appropriate data may exist, they may be difficult to obtain. School-level data are hard to obtain in such states as Ohio and New Mexico, where each district must decide whether to release the information to the public. In Ohio, the process of obtaining permission to examine school-level data from local superintendents has hindered the progress

Vermont's State Assessment System

Vermont's assessment system relies on portfolios in mathematics and writing and the Uniform Assessment. The system aims to encourage effective instructional practices and provide information for public accountability and program evaluation.

In 1990, the state education agency began piloting portfolios in writing and math in grades 4 and 8. The portfolio system was implemented statewide in 1991-92. All schools and districts now participate in the voluntary program. The state collects and scores samples of mathematics portfolios in grades 4 and 8, writing portfolios in grades 5 and 8. The state board of education will implement mathematics portfolios in the high schools in 1995-96 after the end of a three-year pilot program.

Each portfolio is unique, containing five to seven best pieces of work selected by the students from their regular class work. Students are expected to revise the entries and may work on them as long as they choose. The scoring rubric used to assess the math portfolios encourages the submission of the complex, multi-step problem-solving activities advocated in the NCTM standards.

The scoring of the portfolio sample collected by the state is done by teachers who have been trained within regional networks established by the state. The central scoring of the state sample produces results by district.

The regional networks train teachers in portfolio management and scoring. Local districts also hold conferences to score all of their portfolios. About 100 schools plan to score portfolios from the 1994-95 school year and will report their results in the fall. Eventually, school-level scores will be reported for all schools.

The Uniform Assessment covers mathematics and writing. It is designed to supplement the portfolio process with an on-demand, standardized assessment. The mathematics section includes complex multiple-choice items taken from the National Assessment of Educational Progress and one open-ended problem. Vermont expects to replace the Uniform Assessment with a more performance-based assessment in the near future and plans to add a standardized science assessment in 1995-96.

of the state SSI evaluation. As a consequence, the Ohio SSI is administering its own performance assessment in the classrooms of teachers who attend its intensive summer institutes in mathematics and science.

Consistency and Stability of the Measures. Twenty-four of the 25 states surveyed required some kind of standardized testing of students at specific grade levels or grade-level clusters in 1993-94. The one state that did not, Colorado, plans to institute new testing requirements in 1996-97. Of the 24 SSI states currently requiring some form of testing, at least 6 (FL, MT, NE, NJ, OH, and SD) allow some local choice about the tests to be administered at some or all of the tested grade levels. The PSA study did not attempt to determine how useful these "local-choice" testing programs would be for evaluation of systemic reform, but they do raise a problem of equating results on different tests.

Constantly changing state assessment programs further compound the difficulties of evaluating the impact of the SSIs on student performance. Changes in the nature or timing of the assessments make it difficult to determine changes in student performance over time. For example, Massachusetts is in the midst of a transition to an entirely new assessment program. Changes are also underway in California. In New York, the futures of the Pupil Evaluation Program and the Regents Competency tests are unclear. Overall, 14 of the 22 states with mathematics assessments and 12 of the 16 states with science assessments plan to change or expand at least one of their tests during the next few years. Such changes make it difficult to follow trends in student achievement.

Comparability of Assessments Across States. The variety of state SSI objectives and the diversity of state assessment programs limit the use of student-assessment data for comparisons of SSI programs across states. Comparisons can be made across the small number of states that use the same commercially developed tests, but these tests, although still in use in many local districts, are being de-emphasized as states change their statewide assessment programs. In 1993-94, among the 25 SSI states, only 8 administered norm-referenced tests purchased from national publishers; 4 used the Stanford Achievement Test statewide, 3 used the Iowa Test of Basic Skills and 1 used the California Achievement Test. At least 5 other states required districts to administer a commercially published test chosen from a list of state-approved assessments. All told, about half of the SSI states were using these tests.

Summing up: State Capacity for Evaluating the SSIs

The PSA study found only six state assessment programs in mathematics and four in science that met all of its criteria for evaluating the SSIs. That is, assessment programs which test at the appropriate grade levels, report data down to the appropriate level of the educational system, are expected to provide data over five years of the life of the SSI, and have test content and formats aligned with the curriculum and instruction promoted by the SSI.

It is clear that there are only a few states in which available student-assessment data are adequate for an evaluation of systemic reform in math and science education. In others, reformers will have to work with serious data limitations until their state assessment systems can be aligned more closely with the goals of systemic reform efforts. In most cases, the SSIs would have to administer additional student assessment in order to provide a complete picture of their effects on student performance in mathematics and science. Only a handful of SSIs are doing so. The potential for using student-assessment data collected by the states to evaluate the SSI as a federal strategy is limited.

Issues Facing State Policymakers

The PSA study suggests that the data needs of policymakers exceed the capacity of existing state assessment programs to measure progress toward reform goals. This is particularly true for science. Only 16 of the 25 SSI states currently have any statewide assessment in science. And only a few of those appear to be aligned with the emerging science standards. Standards-driven reform requires new assessments that are better aligned with the changes being sought in curriculum and pedagogy. While considerable work is underway to develop such assessments, formidable barriers must be surmounted to develop and implement valid and reliable assessments that are consistent with the emerging standards in mathematics and science. The development of adequate state assessment programs requires addressing some thorny issues such as those discussed below.

Purposes of Assessment

States are seeking new assessment systems that can simultaneously serve multiple purposes: providing good data for comparing the performance of schools, providing diagnostic feedback to help teachers improve their practice and meet the needs of individual students, and providing strong incentives for desired changes in practice. This seems reasonable on its face, but it may be difficult in practice. Assessments measuring individual student achievement and providing instructional feedback may not necessarily be good tools for accountability. And it may be hard to keep the multiple purposes in balance. Thus a high-stakes assessment may drive curriculum and pedagogy in unanticipated, and undesired, directions, while resulting in short-term improvements in performance undermining long-term goals for improving practice.

Validity of Methods

Are the new alternative assessments more valid than multiple-choice tests as their proponents claim? That is, do they more accurately measure the domains of knowledge or attributes of individuals that the reforms in classroom practice are intended to emphasize? Do they provide information that is consistent with other measures of the same domain? Are they good predictors of the future performance in school or at work?

In the case of performance assessment and portfolios, the answers to these and other questions about validity are not yet clear. Reliability (the consistency of the measurement from one time to the next) is also an issue for some of the new assessments. Providing reliable scores for individual students on performance assessments may take more time and be more costly than conventional multiple-choice tests. Portfolios raise special concerns about reliability. Efforts to "improve" portfolios in order to increase the reliability of the scores are viewed by some teachers as undermining their value as instructional tools and levers for classroom reform.

Consequences

Closely related to the issue of purpose is the question of the appropriate stakes to be associated with assessment. If the assessment system is linked to accountability, who should be held accountable, what should the consequences be, and who should bear them? Should there be rewards as well as sanctions? Should money be used as a reward? Should school takeovers or closings be considered as ultimate sanctions? What kinds of data are needed to justify the distribution of rewards and sanctions? These issues are among the most emotional and hotly debated questions about assessment policy.

High-stakes tests are viewed by some as narrowing the curriculum

and negatively affecting both students and teachers. Conversely, others argue that low-stakes tests offer little incentive for teachers to make the difficult changes required in curriculum and pedagogy or for students to work harder and select more difficult courses of study. The question of who should bear the consequences is also debated. Some see schools and their staffs as bearing primary responsibility for student performance. Others argue that students and schools should share responsibility and the consequences. In most states, there are stakes for students (graduation, promotion, school records) and for teachers and schools (public reporting, program changes, state intervention or takeover). How to define stakes so that the press for reform is strong but not distorting is the issue.

Breadth vs. Depth

There is some tension between the reform slogans "high standards for all," "less is more," and "hands-on learning," and the use of statewide standardized assessments. The slogans suggest that all children should meet a uniform set of academic standards and that the standards should be consistent with a constructivist pedagogy that focuses on understanding of key concepts and applications. This approach trades off breadth for depth of understanding, and ideally would leave teachers and schools some discretion about these curricular decisions so that the areas of depth can be related to students' backgrounds, interests, and prior learning. This poses serious challenges for the design of state assessment systems. If the assessment covers too broad a domain of knowledge, teachers may experience difficulty helping students prepare for it. On the other hand, researchers have found that when there are high stakes associated with a test, the provision of specific guidance about what will be tested often narrows the curriculum to what is covered by the test and eliminates local discretion. There appears to be no simple resolution to this dilemma, but balance must be sought between instructional reforms promoting depth at the expense of breadth of curriculum coverage and the assessment system.

Professional Development

The emerging assessment systems make substantial demands on teachers. Teachers not only need to understand the requirements of the new systems, but in many instances, they are expected to change their practice, enhance their subject-matter knowledge, develop new curricula, and serve as overseers and assessors in the new process. They also are on the front line in dealing with parents who may direct their concerns and misgivings about the new assessments to teachers (especially where teachers do the scoring as with portfolios). They need opportunities to acquire the necessary knowledge and skills, to practice new strategies, and to interact with other teachers about what works and how to solve common problems. In short, a radically restructured and refocused system of professional development is needed. The system must be intensive, continuous, and connected to classroom practice.

Development Time

Technically adequate performance assessments will take time to develop and perfect. Vermont has been working on portfolio assessment in math for five years and only now is on the verge of meeting conventional psychometric standards. And it has only begun to implement assessments for the high school. Kentucky's new assessment system is in its fourth year and has undergone critical review by two external groups.

These time frames for development should be seen as normal, and leaders in Vermont and Kentucky should be commended for seeking and using independent evaluations of their efforts. However, impatience and resistance to reform often join forces to focus public attention on the inadequacies of new assessments rather than on their potential benefits and the progress being made to solve their technical and logistical problems. Policymakers need to allow sufficient time for new assessments to be developed, technical problems to be addressed, and legitimacy to be established with educators, parents, and the public at-large.

Costs

New performance-based tests have proved to be more expensive to develop, administer and grade than conventional multiple-choice tests. The demand for increasing investments in assessment comes at a time when state budgets, and local district budgets, are under considerable stress. It is hard to argue for increasing funds for state assessment programs when state aid to schools is being cut. Interstate consortia such as the New Standards Project or the State Collaborative on Assessments and Student Standards (SCASS) initiated by the Council of Chief State School Officers have the potential of reducing development costs for new assessments by spreading them across multiple states. However,

they are only now beginning to produce products for the states to use, so it is not yet clear if they will live up to their promise.

Instructional Time

Performance assessment also takes more time to administer, so more instructional time is lost to assessment. In the case of portfolios, time devoted to revising and polishing final pieces can run into weeks rather than hours if the stakes are high. Increasing time for assessment purposes is not only costly but can generate resentment among teachers, parents, and students, and increase resistance to reforms.

Public Opinion

The new performance assessments are being opposed by some parents. Concern about the new assessment procedures, or about their results, has been expressed at one time or another by parents in highly regarded suburban schools, parents of high school students, parents of children identified as gifted, and parents with lower incomes and levels of education.

It appears that the public has some allegiance to multiple-choice tests because they are familiar and parents feel that they understand the results. Parents also may fear that their children will suffer in some way or lose some competitive advantage from the implementation of new measures of achievement. Public opposition to the new assessments in some states may also be related to the initial small proportions of students able to meet higher standards. This may be particularly true in suburban communities long accustomed to hearing only good news about the test scores in their local schools.

Summing Up: Improving State Capacity for Assessment

These are not simple issues. They are not just technical issues. They will not be solved overnight. Resolving them is confounded by debates within, and outside of, the educational community. There are those who object to standardized assessments and believe the only acceptable assessment happens in the classroom. Such assessments may serve valuable instructional purposes but are virtually useless for the purposes of evaluation and accountability. Some distrust all tests (especially those administered by government agencies), and others remain convinced of the value of traditional tests.

There are also those who fear that new standards and new assessments will introduce new biases into the system. They want to see a new generation of tests characterized by higher standards and better ways of assessing student understanding and ability to apply knowledge. Others believe that state assessment systems can, and should lead reform.

Will We Be Able to Evaluate the Impact of the SSIs?

State assessment programs vary widely in their capacity to produce student performance information that will be useful for evaluating the SSI reform efforts. The PSA study team found only six state assessment programs in mathematics and four in science that met all of their criteria. A handful of other states are now developing such assessments. They are shifting away from minimum-competency standards and multiple-choice formats to new, more complex assessments that examine student understanding of key concepts and their ability to apply knowledge. However, most of these new assessment programs will not come online quickly enough to contribute to the evaluation of the SSIs. Clearly the potential for using student assessment data collected by the states to evaluate the SSI either as a state or a federal strategy is limited. Yet the question of whether the significant investment in the SSIs has paid off in increased student performance demands an answer.

One possible strategy is the use of special assessments in those schools and classrooms most directly involved in SSI activities. Ohio and Puerto Rico have elected to use this strategy which can be implemented faster and at less cost than changing an entire state assessment program. At least in this manner, policymakers can determine whether the SSI strategy being used in their state has the potential for raising student performance.

A second strategy might be for a state to customize a commercially developed, off-the-shelf mathematics and/or science test and administer it in the 3rd, 5th, and 7th years of the SSI.[8] Such a test could be administered to schools targeted by the SSI as in Ohio or Puerto Rico or to a broader sample. It would be cheaper and faster than developing a new state test and could provide useful interim results.

However difficult the technical and political challenges to broadening and strengthening state assessment systems, the nation will not be better off by being uninformed about our children's accomplishments and competence in mathematics and science (or in other subject areas). Ignoring the

gaps between the achievements of children in other developed nations and our children, between boys and girls in our classrooms and schools, and among racial and ethnic groups, will not reduce them or ameliorate their social and economic consequences.

The nation needs good measures of student performance in mathematics and science. In our political system, that need can be met only by state policymakers who are willing to stay the course and help put new standards and assessments in place.

Endnotes

1. Rhode Island is no longer participating in the SSI program. The 25 current SSIs include Puerto Rico, referred to as a state for the purposes of this report.

2. This brief summarizes findings reported in *Assessment Programs in the Statewide Systemic Initiatives (SSI) States: Using Student Achievement Data to Evaluate the SSI* by K. G. Laguarda, J. S. Breckenridge, and A. M. Hightower (Washington, DC: Policy Studies Associates, June 1994).

3. American Association for the Advancement of Science, *Benchmarks for Science Literacy for All Americans* (Washington, DC: Author, 1993); National Council of Teachers of Mathematics, *Assessment Standards for School Mathematics* (Reston, VA: Author, 1993); National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards for School Mathematics* (Reston, VA: Author, 1989); National Council of Teachers of Mathematics, *Professional Standards for Teaching Mathematics* (Reston, VA: Author, 1991); National Science Teachers' Association, *Scope, Sequence, and Coordination of Secondary School Science: Volume 1, the Content Core* (Washington, DC: Author, 1993); National Research Council, *Draft National Science Education Standards* (Washington, DC: Author, 1994).

4. From SSI National Evaluation field notes written by Andrew Zucker, SRI International.

5. The national evaluation is being conducted under contract with SRI International. The Consortium for Policy Research in Education (CPRE), Policy Studies Associates (PSA), and the Council of Chief State School Officers (CCSSO) are sub-contractors.

6. Laguarda, Breckenridge, and Hightower. *Assessment Programs.*

7. The Vermont State Board of Education recently extended the mathematics portfolio to high schools. It will be implemented voluntarily in the 1995-96 school year but most high schools are expected to participate.

8. Although SSIs have been funded by NSF for only five years, many are expected to continue work beyond the funding period with state and private support. ■

The Consortium for Policy Research in Education

CPRE Policy Briefs are published occasionally by the Consortium for Policy Research in Education. The Consortium operates two separate, but interconnected research centers: The Finance Center and The Policy Center. CPRE is funded by the U. S. Department of Education's Office of Educational Research. The Policy Center of CPRE is supported by grant #OERI-R117G1007; the Finance Center of CPRE is supported by grant #OERI-R117G10039.

For further information on CPRE publications contact Pat Michaels, publications assistant, at CPRE, Carriage House at the Eagleton Institute of Politics, Rutgers University, 86 Clifton Avenue, New Brunswick, NJ 08901-1568; 908/932-1331.

CPRE POLICY BRIEFS

Non-Profit Org.
U.S. POSTAGE
PAID
New Brunswick, NJ
Permit No. 1017

Printed on recycled paper.