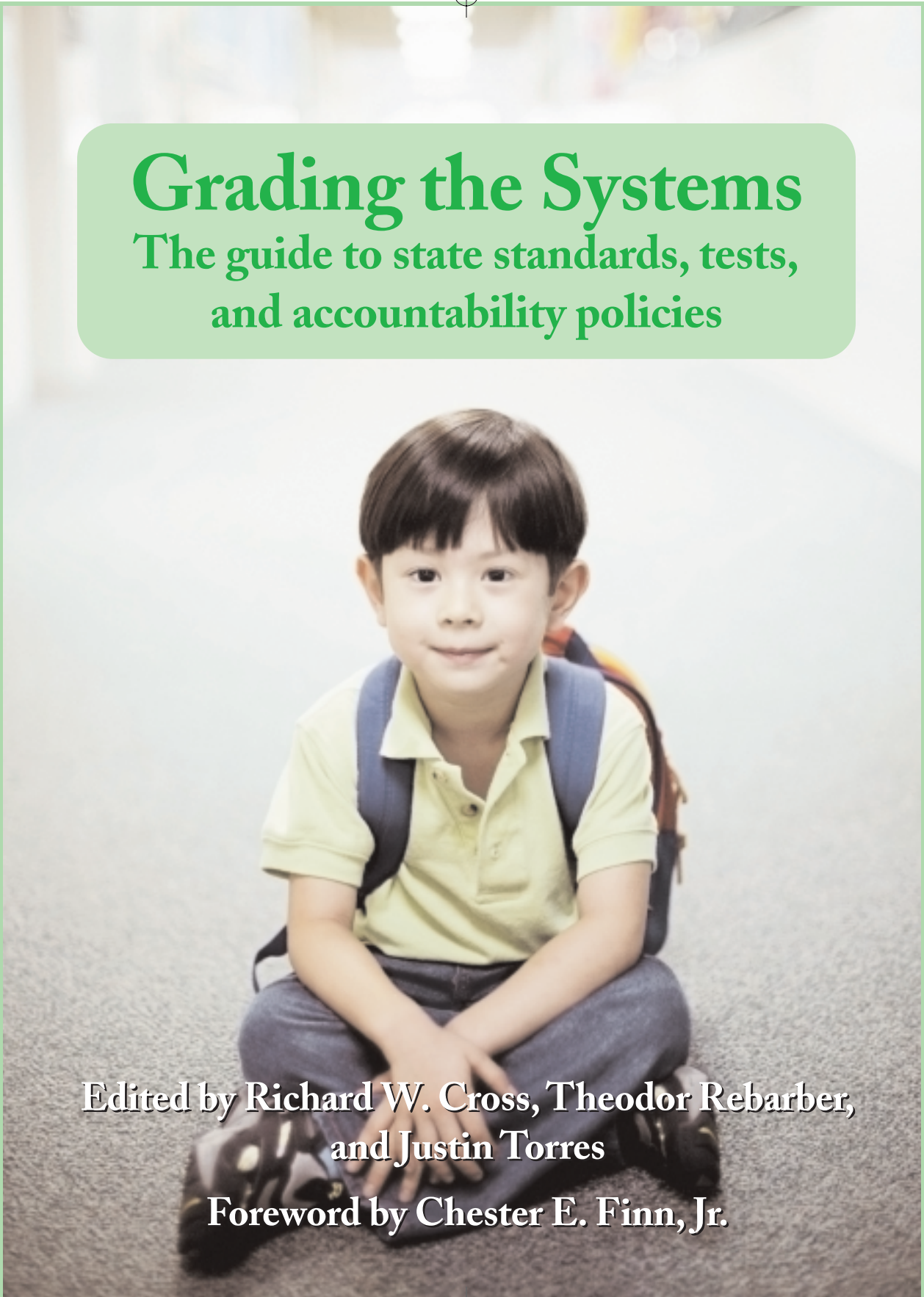
Grading the Systems

The guide to state standards, tests, and accountability policies



Edited by Richard W. Cross, Theodor Rebarber, and Justin Torres

Foreword by Chester E. Finn, Jr.

Grading the Systems

The guide to state standards, tests, and accountability policies

Edited by Richard W. Cross, Theodor Rebarber, and Justin Torres
Foreword by Chester E. Finn, Jr.

1627 K Street, NW
Suite 600
Washington, D.C. 20006
(202) 223-5452
(202) 223-9226 Fax
www.edexcellence.net

Accountability*Works*

1225 19th Street, NW
Suite 400
Washington, D.C. 20036
(202) 261-2610
(202) 261-2638 Fax
www.accountabilityworks.org

January 2004

Contents

Electronic copies of this report are available on the web at www.edexcellence.net. Additional copies are available for $10 each by calling 410-823-7474 or on the web at www.edexcellence.net/foundation/publication/order.cfm. Expanded state summaries, further details, and additional information on study methods and procedures are available on the web at www.accountabilityworks.org.

Design by Katherine Rybak. Printed by District Creative Printing, Upper Marlboro, Maryland.

Foreword

Long before anyone heard of the No Child Left Behind act (NCLB), I was given to comparing standards-based education reform to a "tripod," all three of whose legs must be sturdy if the entire structure is not to tumble down.

The first leg is the standards themselves. Has the responsible body—nearly always a state—done a good job of spelling out the skills and knowledge that it wants its schools to teach and its children to learn, grade by grade (or cluster of grades by cluster)? Are the academic standards clear and specific? Actionable by teachers? Understandable by parents? Do they cover the essential content without over-prescribing: without pinning all schools into a curricular straightjacket, thus eliminating the possibility of school diversity and parental choice, and without shackling teachers to prescriptive lesson plans, materials, and instructional methods when they might better use their own professional judgments? Besides getting the content right, does the state have a suitably ambitious sense of "how good is good enough," i.e., just how much of those skills and knowledge a child must really master in order to be said to have met the standards? It's one thing to set forth a spectacular menu of "wouldn't it be nice if they knew this" items; it's quite another to be clear about how much really needs to be learned for standards to be attained.

All three legs of the education "tripod" must be sturdy if the entire structure is not to come down.

The second leg is the testing or assessment system and other feedback loops by which the state (or school system, principal, or parent) determines just how well a child or classroom or school is doing vis-à-vis the standards. Do the tests "align" with the standards, i.e., do they really probe the essential skills and knowledge that the standards prescribe? Is the difficulty of the test items suitably matched to the scope and ambition of the standards? And is the "cut-off" or scoring system carefully calibrated to the "how good is good enough" decisions? Is the test administered properly so as to minimize shenanigans? Are there clear and consistent rules about which children (and schools) take it and who is exempt? Is it scored and analyzed in ways that enable important comparisons to be made, such as a child's and school's progress from year to year, and a school's performance in relation not only to absolute standards but also to other schools attended by similar youngsters?

The third and final leg is what some people term the "accountability system" itself and others call the "consequences." Because standards-based reform is inherently a behaviorist strategy for influencing people and institutions to attain pre-determined goals and produce certain results, it needs a well-crafted set of incentives, interventions, and rewards that apply at every level. Are these suitably balanced both for children *and* for the adults throughout the system? Do rewards come to those who attain the state's standards? Are there feasible interventions to spot and change the results of those who aren't quite making it? Are there appropriate sanctions for those who fail? And is all of this handled in a reasonable and fair way such that, for example, a child's fate does not hinge overmuch on a single score earned on a single test on a single day? (Everybody has bad days. One of those ought not blight a person's whole life.)

Balancing the Tripod

If all three legs of this tripod are sturdy and of equal length, standards-based reform has a reasonable chance of succeeding in its heroic quest to strengthen the achievement of American children. No, it's not the whole story. It's more like the educational exoskeleton than the vital internal organs of teacher knowledge and skill, well-wrought instructional materials, an orderly learning environment, engaged parents, and much more. The standards themselves don't teach or learn. Many other elements must also be in place, elements that some people term the "opportunity to learn." But the standards tripod also helps those responsible for the other elements to calibrate them to the desired results. For example, teacher training and certification should be aligned with a state's academic standards—and a teacher's preparation should also include a full introduction to assessment and accountability as they operate in that jurisdiction.

Just about every state was embarked on standards-based reform even before the enactment of No Child Left Behind in 2001, but NCLB has "made it official." It's now indisputably the law of the land. Though states remain free to choose their own standards and tests, they must in fact specify them—and report them to Washington. Their test results (in reading and math) are "audited" by the National Assessment of Educational Progress. And their accountability systems must now include, though not be limited to, the tracking of "adequate yearly progress" (AYP) as spelled out by NCLB and the cascade of interventions that NCLB mandates for schools and districts that fail to "make AYP."

Though states remain free to choose their own standards and tests, they must in fact specify them.

Many analysts and organizations have tried to evaluate states on one or more legs of the standards-based "tripod." We at the Thomas B. Fordham Foundation have scrutinized state standards in core academic subjects, as have ACHIEVE and the American Federation of Teachers. ACHIEVE has also examined state tests in relation to their standards and has sometimes looked at a state's entire accountability system—though such reviews have been conducted at the invitation of individual states and have not led to interstate comparisons. The Princeton Review has reported on state "testing systems." And so forth.

Assessing the Need

What's been missing up to now is a holistic examination of the entire standards-based reform tripod across a large number of U.S. states. The reason it's been missing is that it's so difficult to do. Though state standards are relatively accessible, many states are secretive with their tests—and even more so when it comes to their "cut scores" or "passing scores" on those tests. Studying the alignment of tests with standards, and analyzing the rigor of tests, are technically challenging ventures. And nowhere, to my knowledge, is there a tidy collection of the "consequences" associated with each state's accountability system. Moreover, all these things are in constant flux, partly in response to NCLB, partly because of state-generated pressures to update the content, fiddle with the tests, make the "passing level" easier or harder, etc.

With greater hope than confidence, we at the Thomas B. Fordham Foundation decided to venture into this deep and rather murky pond. We sought out Theodor Rebarber and Richard Cross, leaders of a small

but clear-headed organization named AccountabilityWorks. With their varied mix of experience in policy, educational management, and psychometrics, Rebarber and Cross were admirably qualified to undertake such a project, ambitious as it was.

We aren't a big enough foundation to go it alone on an endeavor of this magnitude and were gratified when the Smith Richardson Foundation agreed to join. Besides having deeper pockets, Smith Richardson has a distinguished record of supporting rigorous social science that bears on the understanding of educational effectiveness and student achievement. We could not have asked for a better partner.

The study itself has been slow to mature. We hoped to bring it out some time ago. But the authors were painstaking and refused to be rushed by eager funders. They needed to recruit other experts and conduct a series of complicated analyses. Criteria had to be specified. Facts had to be checked. So be it.

The product, we think, was worth waiting for. To our knowledge, it is the most comprehensive and rigorous appraisal ever undertaken of the entire standards-based "tripod" across a large number of states (30, to be precise). The vast bulk of this immense study is available on the Internet at www.edexcellence.net. It is highly detailed and state-specific, far more than was practical to print. (Yes, it's also a little wonky!)

What We Found

This is the most comprehensive appraisal of the entire "tripod" across a large number of states.

What we've undertaken in *this* report, with extensive editorial work by Fordham research director Justin Torres, is to bring out the study's highlights (and sometimes lowlights).

These come in two forms. First, Messrs. Rebarber and Cross take up the six key measures by which they gauged the adequacy of a state's accountability for K-12 education: the quality of its standards (in reading and math); the content of its tests (where such tests were obtainable); the alignments between standards and tests; the tests' rigor, or how high the cut score was set; their "technical trustworthiness" (e.g., is the scoring protocol "reliable"); and the accountability policies themselves, both before NCLB comes into play and after that statute's additional requirements are put in place.

On each of these six measures, some states fared well indeed. Massachusetts, Pennsylvania, and Virginia each got high marks on three of the six. A few other jurisdictions came within swinging distance of an overall decent performance. But many states did dismally and the multi-state averages can most charitably be termed "fair." (If NCLB is properly implemented—a sizable "if"—the multistate average for "accountability policies" will rise almost to the "solid" level.)

The second way in which this report presents its findings is by profiling each of the 30 states, including a quasi-report card that seeks to rate the state on all six of the major indicators. You'll find a lot of "N/As" there, however, because in many states it was impossible to obtain some of the information needed to complete this analysis and, in a few instances, the information had become obsolete (because of the recent adoption of a new test, for example) and the authors prudently opted not to judge something that

was no longer operative. (By and large, the materials that they evaluated were current as of January 1, 2003.) Because this field is in such flux in many states, partly due to NCLB's influence, all analyses of this type face the challenge of taking a snapshot of a moving target.

These state-specific report cards feature the six topics noted above, evaluated on a five-point scale from "very poor" (1) to "outstanding" (5). You won't find many 1s or 5s. (In truth, there is just one of the latter: Massachusetts did an exemplary job with its standards, though the accountability policies of a few states will be raised to that level if No Child Left Behind is fully implemented—a big "if," as we know.) Rather, you'll see a great many 2s ("poor"), 3s ("fair"), and 4s ("solid"). That's partly because the states in this study rarely did dreadful or terrific jobs on any of these criteria, and partly because each score itself incorporates much averaging of sub-scores. (All this is explained in the methodology section starting on page 38.)

Broadening Our Horizons

What do I make of it all? Three conclusions leap out.

First, it shouldn't be nearly so difficult to conduct such analyses. In the vast majority of cases, specific elements of a state's system of standards, tests, and accountability could not be evaluated because the state would not release basic technical information that it ought to release (with appropriate safeguards). States need to become far more transparent than most are about every aspect of their accountability systems. This is the public's business and the public has a right to see and understand the system by which its schools, children, and teachers are being judged, monitored, and held to account.

Many states did dismally and the averages can most charitably be termed "fair" to "poor."

Second, at a time when "standards-based reform" is America's principal vehicle for education improvement (joined, of course, by choice-based reform), when so much money and energy are being lavished upon it, and when so many hopes are vested in its ability to leverage major gains in school effectiveness and student achievement, the "tripod" on which rests this movement to raise student achievement is perilously shaky. Most states have some semblance of all three legs in place but nowhere can we say with confidence that this essential instrument sits firmly atop the ground.

Third, despite all the hoopla and hand-wringing occasioned by the federal NCLB act, the real action on standards-based reform remains at the state level, and it's the states that must regularly be scrutinized, challenged, judged, praised, and rebuked for whether or not they've got this complex mechanism running smoothly. Though it's running better in some places than others, this pioneering analysis says to me that there's not one state in America today that has got the whole thing right—and in lots of them more needs to be done before one can have confidence that standards-based reform has a reasonable shot at delivering the results that our children deserve and that our voters and taxpayers are counting on.

Many people pitched in to make this report a reality, and in the section on methodology we include a long list of reviewers and analysts who lent their expertise. My thanks to them all. From our end, we would

like to note with special appreciation the contributions of Marc Magee, who helped us pick out the most important elements of each state's analysis; and Marci Kanstoroom, former Fordham research director, who shepherded this project through its early stages.

Phoebe Cottingham and Mark Steinmeyer were the epitome of the gracious funder, which truly characterizes the Smith Richardson Foundation—engaged when they need to be, hands-off when it is best for the project, and always adding value to the final product. And finally, we thank the staff of AccountabilityWorks, which took its time to do a serious investigation and resisted the inevitable pressure to rush things.

The Thomas B. Fordham Foundation, based in Washington, D.C., supports research, publications, and action projects of national significance in elementary/secondary education reform, as well as significant education reform projects in Dayton, Ohio, and vicinity. It is neither connected with nor sponsored by Fordham University. Electronic copies of this report are available on the web at www.edexcellence.net. Additional copies are available for $10 each by calling 410-823-7474 or on the web at www.edexcellence.net/foundation/publication/order.cfm. Expanded state summaries, further details, and additional information on study methodology are available on the web at www.accountabilityworks.org.

Chester E. Finn, Jr.
President
Thomas B. Fordham Foundation
Washington, D.C.
January 2004

General Findings

By Theodor Rebarber and Richard W. Cross

Summary Ratings by State

	Standards	Content	Alignment	Rigor	TT&O*	Accountability before NCLB	Accountability after NCLB
AL	3.6					3.3	4.0
AR	2.3	3.5			3.3	2.3	3.4
AZ	3.4	3.6			3.0	2.3	3.4
CO	3.3	3.8	3.9	2.2	3.5	3.1	3.8
DC	3.2	3.1	2.9	2.8	3.3	2.9	3.7
GA	3.6	3.7	3.8		1.0	2.9	4.1
HI	2.3	3.1	1.5		2.8	1.6	3.3
ID	2.3	3.2	2.6	2.7	3.3	2.2	3.4
IL		3.6		2.0	4.5	3.2	3.8
KY	3.5	2.9			3.7	3.3	3.8
MA	4.7	3.0	3.5	3.7	4.1	3.1	3.8
ME	2.5	2.8	2.7	3.0	2.8	1.8	3.3
MI	2.3	2.9	2.7		2.2	2.9	3.9
MN	2.8	3.3	3.3	4.0	4.2	2.8	3.8
MT	2.2	3.6	2.3	1.5	3.8	1.9	3.3
NC	3.4	3.8	2.9		2.4	4.3	4.5
ND	2.5	3.0	3.2	2.8	3.8	1.6	3.3
NH	2.6	2.8	3.2		4.0	2.4	3.4
NM	2.4	3.4	1.8	3.3	2.6	3.1	3.9
NY	2.5	3.9	2.8	2.7	4.3	3.6	4.1
OH	4.0	3.2	3.1	2.5	3.4	3.2	4.1
PA	3.5	3.6	3.8	1.7	4.4	3.2	4.0
RI	1.5	2.6	2.4		1.4	2.3	3.4
SD	2.5					1.7	3.4
TX	3.5					3.8	4.3
VA	3.1	3.5	3.8	3.0	4.3	3.1	4.1
VT	2.2	2.6	2.9		1.0	2.1	3.5
WA	2.2	3.4			3.4	2.4	3.4
WI	3.1	3.0	2.8	2.2	3.6	2.6	3.6
WV	3.3	3.1	3.1	1.7	3.0	2.5	3.4
U.S. Average	2.9	3.3	3.0	2.4	3.2	2.7	3.7

**Testing trustworthiness and openness*

We are often told that American education is in a new era, one marked by high standards; regular, scientific, and in-depth assessment of student achievement; and stringent accountability measures that hold students, schools, and districts accountable for what is being taught and learned.

Yet is this true? Are state standards truly up to the task of educating every child? Are state tests aligned with those standards—that is, are the tests actually testing what is being taught? Are accountability systems really providing incentives—both positive and negative—for high performance? Do state K-12 accountability systems really work as advertised? Where are states exceeding expectations, and where are they falling short?

To answer these questions, the Thomas B. Fordham Foundation and AccountabilityWorks set out to evaluate state accountability systems. We decided to look at six broad measures of a good K-12 accountability system:

- *Standards for student knowledge and skills*
- *Test content*
- *Alignment between standards and tests*
- *Rigor of the tests*
- *Technical trustworthiness and openness of the tests*
- *Accountability policies (before and after implementation of No Child Left Behind).*

Standards and tests were evaluated in reading and math at the elementary, middle, and high school levels. Ratings are assigned on a 1-5 scale, with 5 as "outstanding," 4 as "solid," 3 as "fair," 2 as "poor," and 1 as "very poor." Our analysis is based on standards, tests, and accountability systems in place as of January 1, 2003, and we have not included in this version of the report evaluations of any elements that have changed since that date.

All told, this project evaluated 30 state accountability systems, including each state's academic standards, assessments (in both reading and math and for elementary, middle, and high school), and the accountability policies that define success and failure and provide both positive and negative incentives for success. The 30 accountability systems evaluated may best be described, on average, as mediocre. That is, they are not terrible, but neither are they yet up to the job of "leaving no child behind." In almost no way can any of these systems be described as outstanding across the board. That overall judgment, however, masks substantial differences—among states, among various dimensions of accountability, and among grades and subjects.

Keep in mind, as you read the general findings and more specific state analyses that follow, that a state that received a "4" or "solid" in any measure has the basis of an adequate set of standards, tests, or accountability policies. However, "solid" does not imply that these elements are fully sufficient—they may, in fact, contain significant gaps. Only those rare elements that rate a "5" (in most cases, grades have been rounded to the nearest whole number) should be looked to as models by other states. (A wealth of specific information on each state is available on the web at www.accountabilityworks.org.)

At the outset, we must make one general observation: There is a disturbing lack of openness to external evaluation on the part of too many state education agencies (which is why this study only addresses 30 states). State accountability systems function in a charged political environment where every error is criticized—in some cases by those who would prefer to see those systems disappear. Yet it is unfortunate that accountability systems intended to shed light on the performance of schools and districts are themselves often shielded from public scrutiny. We commend the leadership of those states that chose to cooperate with this study, especially the five jurisdictions that, with appropriate precautions, made available secure test forms for review (Colorado, Illinois, Michigan, New York, and Pennsylvania).

While this study has gone to unprecedented lengths to analyze academic standards and tests in a systematic and quantifiable manner, relying on professional expertise, we frankly admit that there remains an element of judgment in such matters. Others may favor different criteria or weigh some things differently, which may yield different overall scores and ratings. We have made our methodology, analyses, and criteria available so that a reader can make up his or her own mind about the findings.

Standards

Average: 2.9

The average quality of state standards is only fair, with no significant differences between grades or subjects.

Among the 30 states, Massachusetts' standards in reading and math stand out as the best in the nation. Ohio's are also solid (4), and three others are borderline solid: Alabama (4), Georgia (4), and Kentucky (4). Additionally, Arizona (3), North Carolina (3), Pennsylvania (3), Texas (3), and West Virginia (3) deserve honorable mention for their standards in these subjects.

Standard Ratings by Subject and Grade

Subject	Grade EL	MS	HS	Average
Math	2.8	3.0	3.0	2.9
Reading	3.0	2.7	3.0	2.9
U.S. Average	2.9	2.8	3.0	2.9

Standard Ratings by Subject and Rating Category

	Coverage	Intelligibility	Overall
Math	2.6	3.0	2.9
Reading	2.5	3.3	2.9
U.S. Average	2.5	3.1	2.9

Intelligibility and specificity are characteristics found in all of the high quality standards. What sets Massachusetts apart from the others, in addition to exceptional intelligibility, is the breadth of coverage of the essential skills in both reading and math. The Massachusetts reading standards for high school, for example, cover more than 90 percent of the essential reading skills for the grade level. Their supporting text is excellent, adding to the intelligibility of already well-written standards, and the state is one of only two in the study to include a recommended reading list in the appendix. Other states, such as Ohio, display solid features in one subject (reading), but not in both. The reverse was true for Alabama, which had solid ratings in math, but not in reading. (For a discussion of the "reference standards" used in evaluating state standards and test content, please see page 39.)

Of the 30 states reviewed, only two—Massachusetts and the District of Columbia—include book lists of any kind (either illustrative or mandatory). Since a major element of reading skill level consists of the challenge and sophistication of the texts a student can read, this constitutes a substantial omission. Similarly, knowledge of important texts and authors is useful for college classes in literature as well as general literacy. We are well aware of the political risks of identifying mandatory lists of specific texts and authors, but more states should consider identifying model lists in their reading standards, given the importance of texts in establishing reading expectations.

Test Content

Average: 3.3

Overall, test content is fair. There were no "outstanding" states.

Three states received overall solid ratings for test content: Colorado (4), North Carolina at the elementary and middle grades (4), and New York (4). Six other states are borderline solid, including Arizona (4), Georgia at the elementary and middle grades (4), Illinois (4), Montana (4), New Mexico (4), and Pennsylvania (4). Also deserving of honorable mention are Arkansas (3) and Virginia (3).

Higher-performing states had better high school tests, particularly in math, where other states' high school math tests were generally inferior. A distinguishing feature of both Colorado's and New York's tests was the relatively consistent and strong showings in the quality of math item design (with one slip at the middle school in Colorado), a particularly important and challenging feature of a math test. Coverage of essential skills in both math and reading in the top performing states was not uniformly solid across subjects or grades, hence these top performers were still short of outstanding performance. (For more information on the "reference standards" used in determining essential skills, see page 39.)

The tests assesed for borderline states such as Arizona and Montana were norm-referenced tests—

commercially developed tests purchased off the shelf, such as the Stanford-9—which typically have higher quality in item design, but inferior math content coverage appropriate to high school. The coverage of the most important math skills in algebra (or pre-algebra in the lower grades) and geometry was appreciably better in the higher performing state tests. With the exception of New York, test items that tap computation ("number sense" skills) in the lower grades comprised a significant portion of the better tests (nearly 45 percent of items at elementary level). Given the importance of computational mastery to all subsequent study in mathematics, this large proportion of number sense items is appropriate. Custom written state-developed math tests, which were generally rated lower than norm-referenced tests at the lower grades, had much smaller concentrations of key task types, such as number sense.

In math, the quality of the content is clearly better at the elementary and middle grades than in high school (3.6, 3.3, and 2.7, respectively). This difference is entirely an effect of the widely used norm-referenced tests, purchased off the shelf by some states. The math content quality of the state standards-based tests changes little between the grades, while the math content of norm-referenced tests (NRT) is significantly better than the state-developed criterion-referenced tests (CRT) at the elementary and middle grades (3.8 vs. 3.5 at the elementary level and 3.8 vs. 3.1 at the middle school level) and significantly worse at the high school grades (1.9 vs. 3.2). This disparity is caused by the fact that—with notable exceptions—high school norm-referenced tests contain only limited amounts of high school math, more often recycling middle-school level skills.

Math Test Content Rating, Norm-referenced vs. criterion-referenced Tests

	Grade			
	EL	MS	HS	Average
NRT	3.8	3.8	1.9	3.2
CRT	3.5	3.1	3.2	3.2
U.S. Average	3.6	3.3	2.7	3.2

In the three major content subcategories in elementary and secondary math—geometry, algebra, and number sense (e.g., computation and factoring)—there are significant differences between CRTs and NRTs. At the elementary level, criterion-referenced tests cover a substantially lower proportion of essential skills in number sense than do norm-referenced tests (58 percent vs. 73 percent). This is noteworthy given many American students' struggle with Algebra I and the fact that these fundamental skills are prerequisites for mastering high school algebra.

In algebra, state criterion-referenced tests are noticeably superior to norm-referenced tests at the early high school grades (50 percent vs. 29 percent). There is plenty of room for improvement, however, for all U.S. tests in this area.

In geometry, state criterion-referenced tests are noticeably better than norm-referenced tests at both elementary and early high school grades (67 percent vs. 54 percent in elementary, 49 percent vs. 29 percent in early high school). There is much room for improvement in geometry for all U.S. tests at the middle and high school grades—the average U.S. middle school test covers only about 44 percent of important skills in geometry.

Reading Test Content Rating Norm-referenced vs. criterion-referenced Tests

	Grade			
	EL	MS	HS	Average
NRT	3.4	3.3	3.2	3.3
CRT	3.6	3.3	3.3	3.4
U.S. Average	3.5	3.3	3.3	3.4

With respect to the inclusion of inappropriate or frivolous math items, there are significant differences between tests. At the elementary level, the proportion of inappropriate items is 9 percent on criterion-referenced tests, compared to 5 percent on norm-referenced tests. At the middle

grades, the proportion of inappropriate items is 9 percent on citeria-referenced tests compared to 4 percent on norm-referenced tests.

In the three reading content subcategories—vocabulary, comprehension, literature—norm-referenced tests do a substantially better job of covering important skills than criterion-referenced tests in vocabulary. There is almost no difference between the two with respect to comprehension and literature. Overall, both types of assessments could be improved.

The differences in reading test content between top performing states and others are measurable but not dramatic. The coverage of key skills in reading comprehension and literature is not appreciably different between the top performers and others, with each addressing about one-half of the key skills. Two other commonalities among the best tests are a more sustained focus on vocabulary in the lower grades and a higher concentration of literature items, averaging nearly 30 percent of items devoted to important literary tasks in the top group compared to 22 percent overall. Passage quality is another feature that distinguishes top performing states from others. Reading passages in these states were consistently more challenging and more likely to include reading selections of high literary quality.

With respect to the inclusion of frivolous or otherwise inappropriate items on reading tests, the differences between the grades are slight. The proportions of inappropriate items are 3 percent at elementary grades, 3 percent at middle grades, and 2 percent at high school grades. There may be slightly more such items overall on criterion-referenced assessments than on norm-referenced tests (3 percent on state criterion-referenced tests vs. 1 percent on norm-referenced tests).

Alignment of Tests to Standards

Average: 3.0

Alignment of tests to standards remains a problem in a number of states. The overall average is "fair," with no significant difference by subject or grade level.

A number of states achieved solid alignment grades, including Colorado (4), Georgia (4), Pennsylvania (4), and Virginia (4). Massachusetts also deserves honorable mention (3). More so than in the other rating areas, there are clear and consistent differences in the alignment scores between the highly rated states and others across all grades in both math and reading. The most highly rated states outperform the average by a substantial margin (3.8 vs. 3.0). In the highest-rated states, item alignment—the precision with which each test item maps to one or more state standard—is solid (4.0) relative to all states (3.2). Coverage of state standards in the highly rated states is also solid (3.8) relative to the overall coverage rating (3.0). As one might expect, the top-performing states all have state-developed criterion-referenced tests, which contributes to their solid performance.

A number of unexpected findings crop up when you assess the alignment of state tests and standards.

To begin, while state-developed criterion-referenced tests designed to measure state standards are obviously better aligned than off-the-shelf norm-referenced tests (3.2 vs. 2.5), the difference is not nearly as large as one might expect. States that have spent scarce taxpayers dollars to develop a custom assessment designed solely around their standards should be securing much better alignment. Under NCLB, even states with off-the-shelf norm-referenced tests are being required to augment these assessments with additional items, as needed, to ensure alignment with state standards. Most states, regardless of which type of assessment they used prior to NCLB, need to improve alignment.

Perhaps even more fascinating, in cases where the alignment between standards and tests is not strong (3.4 or lower), the content of the test is superior to the content of the standards 80 percent of the time. In cases of poor alignment (2.4 or less), the test content is superior to the standards content 100 percent of the time. Traditionally, psychometricians have put the onus of alignment on test developers—it is considered their job to align with state standards—and in principle one can hardly argue with that view. But in most actual cases of poor alignment between state standards and tests, our findings suggest that it is the standards, rather than the tests, that need to be rectified. This counterintuitive finding is often due to the fact that too many state standards are vague or incomplete, while the test development process often clarifies, or even adds, important content.

Alignment Average Ratings by Test Type and Grade Level

	Grade			
	EL	MS	HS	Average
NRT	2.6	2.6	2.4	2.5
CRT	3.2	3.1	3.2	3.2
U.S. Average	3.0	3.0	3.0	3.0

Alignment Average Ratings by Test Type and Subject

	Math	Reading	Average
NRT	2.4	2.7	2.5
CRT	3.3	3.0	3.2
U.S. Average	3.1	2.9	3.0

Test Rigor

Average: 2.4

States managed only a "poor" rating on test rigor, as determined by the "cut score," or percentage of correct answers required for a student to achieve proficiency on a test.

Of the six major dimensions of testing and accountability we reviewed for each state, this one contains the lowest scores. Even a number of states with fairly strong test content water down the impact of that content with low expectations for how students will perform on those tests.

The only state that earned a strong overall rating for rigor, meeting our high expectations across multiple tests, is Massachusetts (4). New Mexico also deserves honorable mention for just missing an overall strong rating in rigor (3). (Minnesota earned a "4" on rigor, but we were only able to evaluate one test, making the grade incomplete.) Two of the widely used norm-referenced tests are offered by their vendors with vendor-established proficient cuts, which we

Rigor by Test Type, Math

	EL	MS	HS	Average
NRT	3.5	2.5	3.5	3.2
CRT	2.6	2.2	2.3	2.4
U.S. Average	2.7	2.3	2.5	2.5

Rigor by Test Type, Reading

	EL	MS	HS	Average
NRT	1.5	3.0	3.0	2.5
CRT	2.6	2.6	2.3	2.5
U.S. Average	2.5	2.6	2.4	2.5

compared to the state-established proficient cuts. Overall, the average vendor cuts for norm-referenced tests are significantly more rigorous than the state cut scores (2.8 vs. 2.4), with the difference especially large in math (3.2 vs. 2.4).

Test Trustworthiness and Openness

Average: 3.2

Test Trustworthiness by State

	Overall	Reliability	Equating Info	Rater Reliability	Categorical Reliability	NAEP Read	NAEP Math	Security, Openness
AR	3	5	4	3	1	4		3
AZ	3	5	4	3	1	3	4	1
CO	4	5	5	1	3			4
DC	3	5	4	3	1		4	3
GA	1	1	1	1	1			1
HI	3	5	4	3	1			1
ID	3	5	5		2			1
IL	4	5	4	5	5			3
KY	4	5	5	5	2	3	3	3
MA	4	5	5	4	4			3
ME	3	4	1	1	5		3	3
MI	2	3	2	2	1			3
MN	4	5	5	4	3			4
MT	4	5	5		2			3
NC	2	1	1	1	1	5	5	3
ND	4	5	5	5	2			2
NH	4	5	5	4	4			3
NM	3	5	5	1	1			1
NY	4	4	5	4	5	4		4
OH	3	4	4	4	3			3
PA	4	5	4	5	4			4
RI	1	1	1	1	1			3
VA	4	4	4		5	4		4
VT	1	1	1	1	1			1
WA	3	5	5		2		4	1
WI	4	5	5	5	2			1
WV	3	5	4	3	1	3	4	1
U.S. Average	3.2	4.1	3.9	3.0	2.2	3.7	3.9	2.5

Overall, the technical trustworthiness and openness of state testing programs is fair.

The highest ratings in this category belong to Illinois (4), New York (4), Pennsylvania (4), and Virginia (4). Other states with strong scores include Colorado (4), Kentucky (4), Massachusetts (4), Minnesota (4), Montana (4), New Hampshire (4), North Dakota (4), and Wisconsin (4).

Evidence of general test reliability (what we call "internal consistency reliability") is very strong for most states, as is evidence of procedures to ensure that tests are truly parallel from year to year (and other forms of "equating").

Disturbingly, evidence on the particular measure of reliability relevant to whether or not students have met state standards ("categorical reliability") is generally poor. Of the states studied, only Illinois, Maine, New Hampshire, New York, Pennsylvania, and Virginia document strong evidence in this important area. The average categorical reliability rating of this group of states (4.3) is the most important element in setting them apart from the general run of states in this trustworthiness and openness. Overall, all states manage only a poor (2.2) rating in this category.

A substantial number of states are not very open or user-friendly with information about their testing programs (even allowing for the issue of reviewers' access to secure items), including test security policy. In many states, this limits the basic information that any

researcher would need to make substantive judgments about the reliability and trustworthiness of these tests. Further, even for some commercially available norm-referenced tests, information about test characteristics (ordinarily published in the test technical manual) is often unavailable for months or years after the test has been put into use.

In the higher-performing states, the openness rating of solid (3.6) is the second most distinguishing feature, relative to the overall borderline poor performance (2.5). These high-performing states made their testing policy, procedures, and technical information widely available and user-friendly. Test trustworthiness had the largest overall range in scores across the states, with some states receiving very poor ratings, due in large measure to the limited information provided in this area.

Accountability Policies

Average before NCLB: 2.7
Average after NCLB: 3.7

Prior to the passage of the No Child Left Behind act, state accountability policies on average were only fair, bordering on poor. NCLB, if properly implemented, would increase the average accountability ranking significantly.

Prior to NCLB, only 3 of the 30 states studied had strong accountability policies: New York (borderline 4), North Carolina (4), and Texas (4). If NCLB is fully and properly implemented, 18 states, a majority of our group, will have solid accountability policies (4 or higher). Even the accountability policies of all the other remaining states will be near solid (high 3). No Child Left Behind, however, does not improve states equally, since some important parts of a solid accountability system, including sanctions and incentives for individual students and adults, are left unaddressed by the law.

These highly rated states have solid ratings in three of the four major rating categories in contrast to the overall ratings, Performance Goals (4.0 vs. 2.8) Assessment & Evaluation (4.2 vs. 3.2) and Intervention & Adult Consequences (4.2 vs. 2.5). Even in Student Consequences, where the high performing states average only a fair (3.1) rating, this is still well above the average rating of 1.9 across all states in this study. Of the highly rated states, North Carolina achieved the best rating, a borderline solid performance (3.7) in addressing individual student consequences.

Assuming NCLB changes, such as sanctions for schools and districts, are fully added to what states were already doing, states' policies will be relatively solid in the areas of setting performance goals for schools and districts as well as in analyzing and reporting the data. State policies will be fair in the areas of intervening in failing schools and districts as well as assisting struggling students. Yet, even after NCLB, incentives and consequences for adults and for individual students remain weak throughout the states.

States were only fair (2.8) in establishing school performance targets prior to NCLB, but the new law

will strengthen this area considerably (3.8). Enhancements include accountability targets for all schools as well as for all school subgroups. Assessment and evaluation policies are also radically improved (from 3.2 to 4.6) through NCLB requirements such as grade-by-grade testing.

Accountability Policies by State (before & after NCLB)

	Overall Average		Performance Goals		Assessment & Evaluation		Incentives & Consequences for Adults		Incentives & Consequences for Students		Assistance for Struggling Students	
State	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
AL	3.3	4.0	3.0	4.0	3.3	4.5	3.5	3.5	3.0	3.0	4	5
AR	2.3	3.4	3.0	3.8	2.8	4.5	1.5	2.5	1.3	1.3	1	4
AZ	2.3	3.4	2.5	3.5	3.2	4.5	1.0	2.5	1.3	1.3	1	4
CO	3.1	3.8	4.0	4.3	3.3	4.7	4.0	4.0	1.3	1.3	2	4
DC	2.9	3.7	3.8	4.3	3.3	4.5	2.0	3.0	1.7	1.7	3	4
GA	2.9	4.1	1.3	3.3	3.7	4.7	1.5	3.0	4.3	4.3	4	5
HI	1.6	3.3	1.5	3.3	2.0	4.5	1.5	2.5	1.0	1.0	1	4
ID	2.2	3.4	1.8	3.5	3.3	4.7	1.5	2.5	1.3	1.3	1	4
IL	3.2	3.8	3.8	4.3	3.7	4.7	3.5	3.5	1.3	1.3	3	4
KY	3.3	3.8	3.8	4.3	3.8	4.7	3.5	3.5	1.3	1.3	3	4
MA	3.1	3.8	2.8	4.0	3.5	4.5	2.5	2.5	3.0	3.0	3	4
ME	1.8	3.3	1.3	3.3	2.7	4.7	1.0	2.5	1.0	1.0	2	4
MI	2.9	3.9	3.8	4.3	3.2	4.8	2.0	2.5	2.7	2.7	1	4
MN	2.8	3.8	3.3	4.3	3.3	4.5	1.5	3.0	2.0	2.0	2	4
MT	1.9	3.3	1.3	3.3	3.2	4.7	1.5	2.5	1.0	1.0	1	4
NC	4.3	4.5	4.0	4.8	4.7	4.8	4.0	4.0	3.7	3.7	5	5
ND	1.6	3.3	1.0	3.3	2.3	4.5	1.5	3.0	1.0	1.0	1	4
NH	2.4	3.4	2.8	3.5	3.0	4.7	2.5	2.5	1.0	1.0	2	4
NM	3.1	3.9	4.0	4.3	2.7	4.5	3.5	3.5	2.3	2.3	3	4
NY	3.6	4.1	4.0	4.3	3.7	4.7	4.0	4.0	2.3	2.3	4	5
OH	3.2	4.1	2.8	3.5	3.2	4.8	3.0	3.0	3.7	3.7	4	5
PA	3.2	4.0	3.5	4.3	3.5	4.5	4.0	4.5	2.3	2.3	1	4
RI	2.3	3.4	3.0	3.8	2.7	4.5	2.5	2.5	1.0	1.0	1	4
SD	1.7	3.4	1.0	3.3	2.7	4.7	1.5	3.0	1.0	1.0	1	4
TX	3.8	4.3	4.0	4.0	4.2	4.8	4.5	4.5	3.3	3.3	1	4
VA	3.1	4.1	1.5	3.3	3.7	4.8	3.5	4.5	3.3	3.3	4	5
VT	2.1	3.5	2.3	4.0	2.7	4.5	2.5	3.0	1.0	1.0	1	4
WA	2.4	3.4	2.5	3.5	3.3	4.5	1.5	3.0	1.3	1.3	1	4
WI	2.6	3.6	3.0	4.0	3.5	4.7	2.0	3.0	1.0	1.0	2	4
WV	2.5	3.4	3.0	3.5	3.2	4.7	2.5	2.5	1.0	1.0	1	4
U.S. Average	2.7	3.7	2.8	3.8	3.2	4.6	2.5	3.1	1.9	1.9	2.1	4.2

Incentives and consequences for adults are also improved a bit (from 2.5 to 3.1) as a result of NCLB, but there remains much leeway for states to implement strong policies in this area. Incentives and consequences for individual students were a major policy deficit prior to NCLB and are entirely unaffected afterward (1.9 before and after), because NCLB is silent in this area.

A notable but often overlooked enhancement wrought by NCLB will be assistance to struggling students, which will improve markedly due to the supplemental education services provisions of the new law (raising a score of 2.1 to 4.2).

State Summaries

Alabama

Alabama's standards are solid overall but vary considerably among elementary, middle, and high school, and between reading and math. For example, high school math standards are excellent, with consistently well-written examples and an appropriate level of difficulty throughout. But middle school reading standards are only fair; the generally intelligible layout is undermined by the use of vague descriptions—for example, the directive to "value recognized written, spoken, and visual works of literature representative of various cultures and eras," which is neither specific nor verifiable—and the inclusion of only one-third of essential skills as defined by the reference standards. (More information on the reference standards used in evaluating state standards and test content is available at page 39.) Overall, in both reading and math the standards were more clear than comprehensive, with lower marks for content coverage.

Alabama

Standards	**4**	**Solid**
Test Content	**N/A**	**N/A**
Alignment	**N/A**	**N/A**
Test Rigor	**N/A**	**N/A**
Test Trustworthiness and Openness	**N/A**	**N/A**
Accountability Policies*	**3 (4)**	**Fair (Solid)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on Alabama Course of Study Content Standards. Our evaluations of elements relating to the state's tests are not included in this report, because the tests have changed since the analysis was performed. Go to www.accountabilityworks.org for more information.

Our evaluations of elements relating to the state's tests are not included in this report, because the tests have changed since the analysis was performed. An analysis of the test that has been replaced, the SAT-9 tests used in grades 4, 8, and 10, is included in the "Further Details" section available on the web at www.edexcellence.net.

Alabama's accountability policies prior to NCLB were fair, falling short in nearly every category, but seldom very short. Strengths include the state's assessment program, with its annual student testing in grades 3 through 8, and its system of incentives and sanctions for educators. The most notable deficiency was the lack of performance levels for students, with results reported only in terms of percentile rank without any independent proficiency scores. Should the state implement NCLB requirements, we expect its accountability policies to improve to solid.

Arizona

Arizona's standards are fair, with the best being high school math, which are well written and refreshingly direct. More typical are the middle school math standards, whose intelligibility is sometimes obscured by inflated language and items that are not grade-appropriate. For example, one middle school math standard requires that students "construct, use, and explain algorithmic procedures for computing and estimating with whole numbers, fractions, decimals, and integers." Most likely, a simpler task is intended, as jus-

tification and construction of an algorithm is far too complex an operation for middle school students. This standard also lacks the intelligibility of middle school reading, which for example direct students to "Identify information that is either extraneous or missing (e.g., directions, tools required, parts needed, illustrations, diagram sequence, bold face for relevant steps)." Important content receives sufficient coverage in math at the high school level but not enough at earlier grades, while the opposite is true for reading.

Arizona		
Standards	**3**	**Fair**
Test Content	**4**	**Solid**
Alignment	**N/A**	**N/A**
Test Rigor	**N/A**	**N/A**
Test Trustworthiness and Openness	**3**	**Fair**
Accountability Policies*	**2 (3)**	**Poor (Fair)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on Arizona's Instrument to Measure Standards (AIMS); SAT-9 tests in all grades and criterion-referenced tests in grades 3, 5, 8, and 11. Test content standards are based on the SAT-9, since the state did not make its CRT available for review. Test alignment and rigor were not evaluated because the AIMS was not made available for review. Go to www.accountabilityworks.org for more information.

The SAT-9 test content is generally solid, though high school reading passages suffer from being insufficiently challenging and skew too heavily toward non-fiction. Similarly, the high school reading test does not measure a student's skills in some important areas, like poetry. (More information on the reference standards used to evaluate state standards and test content is available at page 39.) Technical trustworthiness and openness is fair, with trend data going back to 1997 positively associated to the NAEP. However, there is no readily available information on test security.

Arizona's pre-NCLB accountability policies were poor. While the state's assessment system has an impressive mix of standards-based and norm-referenced tests at many grade levels, the state used a limited system of incentives and consequences for educators and had no system of incentives and consequences for students. Should Arizona implement NCLB requirements, we expect its accountability policies to improve to fair.

Arkansas

Arkansas' standards are poor, with a fetish for application to "real world problems" that detracts from the core content to be learned. This problem is especially apparent in middle school math, which contains numerous requirements for facile applications to the real world, such as the directive to "Apply computation (add, subtract, multiply, and divide) and estimation to real-world problems." Other standards are hopelessly vague; for example, "Describe, model, draw, construct, compare, and classify shapes in one, two, and three dimensions," which does not specify *which* shapes (triangles, rectangles, squares, etc.) or use proper terminology ("construction" is a technical term requiring the use of straight-edge and compass). Elementary reading standards are typical, with sufficiently clear writing undermined by flawed assumptions (for example, the premise that all textual meaning is derived from students' thought processes and is not inherent in the text) and coverage of only about one-

Arkansas		
Standards	**2**	**Poor**
Test Content	**3**	**Fair**
Alignment	**N/A**	**N/A**
Test Rigor	**N/A**	**N/A**
Test Trustworthiness and Openness	**3**	**Fair**
Accountability Policies*	**2 (3)**	**Poor (Fair)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on Arkansas' Standards for Accreditation; SAT-9 tests in grades 5, 7, and 10, and criterion-referenced tests in grades 4, 8, and end-of-course. Test content ratings are based on the SAT-9, since the state criterion-referenced tests were not made available for review. It was not possible to assess test rigor or alignment because the state's criterion-referenced tests were not made available for review. Go to www.accountabilityworks.org for more information.

third of essential grade-level skills. Almost two-thirds of the standards address worthwhile content, but less than half of the important grade-appropriate skills are covered by these standards. However, many of the state standards address core content only marginally, making our rating of 2 generously high. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.)

The best test content is found in elementary and middle school in both math and reading. Math especially provides solid coverage of the key math skills, and reading test passages include a variety of genres with an appropriate level of difficulty. The elementary level reading test provides solid coverage of essential skills, at a level somewhat higher than the middle school test. High school math and reading assessments provide insufficient coverage of core content. The reading test uses reading passages that are not particularly interesting or challenging. The testing system has a fair overall rating for trustworthiness, with test security policy readily available and reading scores positively associated with the NAEP.

Arkansas' pre-NCLB accountability policies were poor, with the state falling short of our expectations in nearly every area. A particularly glaring weakness is the complete lack of consequences for low-performing students and the limited number of disaggregated subgroups. A relative strength is the assessment program, as it covers most grades and subjects. Should the state implement NCLB requirements, state accountability policies should improve to fair.

Colorado

Colorado		
Standards	**3**	**Fair**
Test Content	**4**	**Solid**
Alignment	**4**	**Solid**
Test Rigor	**2**	**Poor**
Test Trustworthiness and Openness	**4**	**Solid**
Accountability Policies*	**3 (4)**	**Fair (Solid)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on Colorado Model Content Standards; criterion-referenced tests in grades 3-10 in reading and grades 5-10 in math under the Colorado Student Assessment Program (CSAP). Go to www.accountabilityworks.org for more information.

Colorado's standards are fair, with middle school math—marked by clear, specific, and measurable standards that address a fairly high proportion of essential skills—being the best. More typical are the elementary math standards, which contain useful definitions and are written in clear language, but contain significant gaps in content (only half of essential fourth grade skills are covered) and are grade-appropriate only two-thirds of the time. High school math standards are the worst of the group, with a excessive emphasis on "real-world math." For example, standard 5.2.1 reads, "Solve real-world problems involving multiple dimensions and express them using appropriate units of measurements." This is a simple measurement and units problem that is more suitable to the 8th grade or below. Arguably, the nine standards that make up section six of the document are all below grade level.

Colorado's standards-based tests provide a solid level of overall coverage. These are high-quality tests that generally have a minimal amount of frivolous or inappropriate items. The best test content is found in the elementary and high school math tests, with their strong assessment of essential math skills, items that map directly to standards, and minimal design problems. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.)

The overall alignment between these tests and Colorado's state standards is solid. The alignment for high school reading is typical, with the test covering about two-thirds of the state standards. But this high alignment is undercut by poor test rigor. Colorado—unfortunately enough given the efforts it has taken to design high-quality tests—consistently places the cut score for proficiency below the level appropriate for each grade. The worst offender is middle school reading, where the state's cut score of 50 percent is 30 points below the appropriate level. Test trustworthiness is solid, with high levels of both internal consistency and year-to-year consistency, but there is no documentation of the reliability of human scoring of open-ended questions.

Colorado's pre-NCLB accountability policies were fair, with the state falling short of a solid rating despite strong performance goals because there is no system of student consequences. Should the state implement NCLB requirements, we expect its accountability policies to improve to solid.

District of Columbia

The general structure of the District's standards—which are, on average, fair—appears rather straightforward, but the details betray a subtle confusion throughout that decreases their usefulness. In the worst case, three-fourths of the elementary math standards cover appropriate content, but the standards overall cover less than half of the most important grade-appropriate skills (and suffer from the recurrent fetish for "real-world application"). Middle school reading is more typical, with three-quarters of essential skills covered by specific and measurable standards, though supporting text does little to enhance understanding and in some instances actually detracts from the intelligibility of the standards. For example, each content standard is divided into muddled categories of "performance standards," "essential skills," and "technology integration." The District does a good job of clarifying some expectations, requiring students to read "at least thirty books or book equivalents," though we are unclear what a useful "book equivalent" would be.

District of Columbia

Standards	**3**	**Fair**
Test Content	**3**	**Fair**
Alignment	**3**	**Fair**
Test Rigor	**3**	**Fair**
Test Trustworthiness and Openness	**3**	**Fair**
Accountability Policies*	**3 (4)**	**Fair (Solid)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on District of Columbia's Standards For Teaching and Learning; SAT-9 tests in grades 4, 8, and 10. Go to www.accountability-works.org for more information.

The SAT-9 tests used by the District of Columbia in grades 4, 8, and 10 provide a fair level of content coverage. The best test content is found in the middle school math test, with a very high proportion of items focused on essential 8th grade math skills, while high school reading and math vie for the worst coverage of the key 10th grade math and reading skills in this study. In addition, reading is marred by uninteresting and unclear reading passages. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and the District of Columbia's standards is fair, with middle school math typical in covering about half of the district's standards.

The rigor of the test, which follows the SAT-9 publsher's established cut scores, is also fair, with proficiency cut scores falling about 10 percent lower across the board than is appropriate for the specific grade

level. Test trustworthiness and openness is fair; math scores on the test correlate with NAEP scores, but only limited test security information is available.

The District of Columbia's pre-NCLB accountability policies were fair, with a solid assessment program grounded in annual student testing between grades 1 and 11 undermined by a poor system of consequences for students and educators. Should the state implement NCLB requirements, we expect its accountability policies to improve to solid.

Georgia

While the average quality of Georgia's reading and math standards is just barely solid, beneath this average is a wide variance in quality. For example, while the coverage of the middle school math standards is outstanding, the intelligibility of these standards is poor, with vague, ill-defined terms leaving important gaps in its discussion of mathematical strategies. Standard 1 in middle school math indicates that students should solve "non-routine" problems, without defining "non-routine." Standard 3 recommends using "scientific calculator and computer skills to solve problems, to discover patterns and sequences, to investigate situations, and draw conclusions." This standard is completely backward: it puts pedagogy before content (technology is primary to the skill); never defines what is meant by "investigate situations"; and confuses inductive reasoning with mathematical proof. In the state's high school reading standards the pattern is reversed, with extensive supporting text leading to a solid level of intelligibility that is offset by a poor overall level of coverage.

Georgia

Standards	**4**	**Solid**
Test Content	**4**	**Solid**
Alignment	**4**	**Solid**
Test Rigor	**N/A**	**N/A**
Test Trustworthiness and Openness	**1**	**Very Poor**
Accountability Policies*	**3 (4)**	**Fair (Solid)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on Quality Core Curriculum (QCC) Standards; state-developed Criterion-Referenced Competency Tests (CRCT) in grades 1-8. It was not possible to assess test rigor because the state did not provide the necessary technical information. Georgia's standards are being revised as this report is published. Go to www.accountabilityworks.org for more information.

The state-developed criterion-referenced tests used by Georgia provide a borderline solid level of overall coverage. The elementary school reading test is typical of the state's overall content coverage, with barely more than half of essential reading skills covered, but with the best coverage found in the most important skill areas. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and Georgia's state standards is also solid. The alignment for elementary school math is typical, with the test covering more than two-thirds of the state standards.

However, the testing system has a very poor overall rating for trustworthiness and openness. The state's department of education was unwilling to provide the requested technical material for this study—a request that is well within the norms of the profession—and the longitudinal testing data that were provided by the state fell outside of the ranges of available NAEP data, making any comparison with independent trends impossible.

Finally, Georgia's accountability policies prior to NCLB were fair, a score that reflects the unevenness of the state's system. While Georgia has an outstanding assessment program, with annual student testing in reading and math in grades 1-8 and 11, and in social science and natural science in grades 3-8 and 11, the system of consequences for educators is poor, with only a meager positive incentive program and no system of monetary awards. However, it should be noted that the Georgia legislature has passed legislation that would correct these problems if implemented. Should the state implement NCLB requirements, we expect its accountability policies to improve to solid.

Hawaii

In most grades and in both reading and math, Hawaii's standards are poor—badly written, vague, and with numerous instances of problematic content. Examples abound: elementary school reading directs students to "interact thoughtfully with each other about texts that represent diverse perspectives," which is a grand slam of undefined, unassessable, and tendentious notions. Middle school reading standards cover less than half of essential reading and literature skills and further direct students to "read to understand human experience and the range of choices and possibilities in life." The high school math standards suffer from similar problems. The standards include a set of "benchmarks" designed to add detail. To put it simply, they fail in this regard. The geometry benchmarks illustrate the problem. There are only nine, covering the entire four years of high school. The first two read, "Make and evaluate conjectures about, and solve problems involving, classes of two- and three-dimensional geometric objects (e.g., 'Are all squares rectangles?')," and, "Use logical reasoning to create and defend valid geometric conjectures." These are much too broad to provide guidance for either assessment or instruction. Further, the first example, "Are all squares rectangles?" is a question fit for 4th or 5th grade, not high school. The next benchmark, "Represent transformations of objects in the plane with coordinates, vectors, or matrices; describe the effects of a given transformation," addresses content that is generally found only in pre-calculus courses. The intervening material on Euclidean geometry—the staple of any core set of geometry standards—is completely absent.

Hawaii		
Standards	**2**	**Poor**
Test Content	**3**	**Fair**
Alignment	**1**	**Very Poor**
Test Rigor	**N/A**	**N/A**
Test Trustworthiness and Openness	**3**	**Fair**
Accountability Policies*	**2 (3)**	**Poor (Fair)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on Hawaii Content Standards; SAT-9 tests in grades 4, 8, and 10. It was not possible to assess the rigor of the state's tests because the state does not report SAT-9 test results by proficiency level. Go to www.accountability-works.org for more information.

The SAT-9 tests used by Hawaii provide a fair level of content coverage overall. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) However, the alignment between these tests and Hawaii's standards is very poor. The alignment for elementary school math is typical, with the test covering about one-quarter of the state standards. The testing system has a fair overall rating for trustworthiness, but a lack of statewide longitudinal data means that comparisons with NAEP trends are not possible.

Finally, Hawaii's accountability policies prior to NCLB were poor. Despite annual progress reports on

schools, the system does not include any annual goals nor does it measure such basic indicators as attendance. While students are tested in grades 3, 5, 8, and 10 in math, reading, and writing, Hawaii did not categorize schools based on performance and offers no positive incentives for teachers or administrators. Should the state implement NCLB requirements, we expect its accountability policies to improve to fair.

Idaho

Idaho's math standards are generally poorly written and lack specificity. Only at the elementary level is the coverage of important skills fair, with almost two-thirds of the state standards addressing appropriate content, and about half of the most important skills covered and with a concentration on the most important number sense skills—a concentration that adds to the coverage rating. Still, the standards are beset with intelligibility problems, largely due to a lack of specificity and helpful examples.

Idaho

Standards	**2**	**Poor**
Test Content	**3**	**Fair**
Alignment	**3**	**Fair**
Test Rigor	**3**	**Fair**
Test Trustworthiness and Openness	**3**	**Fair**
Accountability Policies*	**2 (3)**	**Poor (Fair)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on Idaho Achievement Standards—math only—and ITBS tests in grades 4, 8, and 10. The Idaho English Language Arts standards have changed since our analysis was performed and are not included in this report. Alignment results are limited to math. Go to www.accountability-works.org for more information.

Idaho uses the ITBS, which is a good choice for a norm-referenced test for the elementary and middle school grades. The high school math test is another matter, however; it is packed with items more appropriate to middle school and contains little high school math content. The best test content is found in the middle school and high school reading tests, which cover nearly two-thirds of the most important reading skills. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) Given the limitations in the standards, the overall fair alignment rating between these tests and Idaho's state standards is about as high as one could reasonably expect. The alignment for elementary school math is typical, with the test covering about one-half of the state standards.

Testing rigor is also fair, but varies widely among the tests. For example, while the cut score for proficiency in middle school reading is outstanding, exceeding even the standards of our own independent evaluators, the cut point for proficiency in the middle school math test is very poor. Technical trustworthiness is fair as well. The state does not make readily available longitudinal trend data on the ITBS, nor has the state participated in NAEP testing since 1992, making trend comparisons impossible.

Despite an extensive assessment program in which students are tested annually from K through 11, prior to NCLB there was no clear categorization of schools by their level of performance and no framework for either educator or student performance consequences. Overall accountability policies were poor, though with the implementation of NCLB that score will rise to fair.

Illinois

Middle school math is characteristic of the state's overall test content coverage, with about three-quarters of essential grade-level math skills covered, solid item design, and few irrelevant or frivolous items. But at all three of the testing levels reviewed, a significant portion of the items are concerned with less important material, or, as with high school reading, are concentrated almost exclusively in one domain (reading comprehension). (See page 39 for more on the "reference standards" used in evaluating state standards and test content.)

Illinois		
Standards	**N/A**	**N/A**
Test Content	**4**	**Solid**
Alignment	**N/A**	**N/A**
Test Rigor	**2**	**Poor**
Test Trustworthiness and Openness	**4**	**Solid**
Accountability Policies*	**3 (4)**	**Fair (Solid)**

***(Scores after NCLB requirements in parentheses)**

Standards ratings are not included, since the state has added a set of Assessment Frameworks since our analysis was completed; tests are the Illinois Standards Achievement Tests (grades 5 and 8) and the Prairie State Achievement Examination (Grade 11). Illinois' grade 3 test is not evaluated in this study. Go to www.accountability-works.org for more information.

The rigor of the tests earned a poor rating, with the cut score for proficiency consistently placed 20 percentage points below what our analysis suggests is appropriate. However, the testing system has a solid overall rating for trustworthiness and openness, with particularly high ratings for internal consistency and the reliability of human scoring of open-ended items (for which it received the highest rating in our study, just shy of outstanding).

Illinois' accountability system was fair prior ot NCLB. While the state had a solid system of performance categorization and a solid assessment program, it was undermined by a lack of student consequences. Should NCLB be fully implemented, we expect its rating to rise to solid.

Kentucky

The average quality of Kentucky's standards is solid, with math standards generally better than reading and high school standards generally better than those for early grades. The standards for high school reading cover three-quarters of essential skills and consistently use clear language and helpful supporting text; for example, the instruction to "Analyze the effect of theme, conflict and resolution, symbolism, irony, analogies, and figurative language." Supporting text provides a clear explanation of what students need to do in order to be very good readers, including the important distinction that "Students must understand that various types of reading materials have different features that affect how they are read." Throughout, the coding of these standards is overly complicated, and coverage of essential skills is weaker than intelligibility, but this is generally high-quality work.

Kentucky		
Standards	**4**	**Solid**
Test Content	**3**	**Fair**
Alignment	**N/A**	**N/A**
Test Rigor	**N/A**	**N/A**
Test Trustworthiness and Openness	**4**	**Solid**
Accountability Policies*	**3 (4)**	**Fair (Solid)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on Kentucky's Core Content standards; CTBS-5 norm-referenced tests in grades 3, 6, and 9. It was not possible to assess either test alignment or test rigor because the state's criterion-referenced tests were not made available for review. Test content analysis applies only to norm-referenced CTBS-5 tests. Go to www.accountabilityworks.org for more information.

Kentucky uses an unusual version of the CTBS-5 norm-referenced tests in grades 3, 6, and 9, and a state-developed criterion-referenced test in other grades. Since the state-developed tests were not made available for review, the content coverage analysis applies only to the CTBS-5 tests, which were generally fair. The elementary school reading test is typical of the state's overall content coverage, with reading passage of appropriate length and difficulty, but only about one-half of essential grade-level reading skills are covered, with a higher level of coverage for the most important of these skills. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The testing system has a solid overall rating for trustworthiness and openness, with information available that addresses security procedures and limited information available on the relationship with NAEP reading and math trend data.

Kentucky's accountability policies prior to NCLB were fair. While the state has a particularly strong assessment program and solid performance goals for schools and districts, it falls well short in its system of consequences for educators and students, with little evidence that the incentive system for educators works and the only consequence for low student performance being remediation. Should the state implement NCLB requirements, we expect its accountability policies to improve to solid.

Maine

Maine		
Standards	**2**	**Poor**
Test Content	**3**	**Fair**
Alignment	**3**	**Fair**
Test Rigor	**3**	**Fair**
Test Trustworthiness and Openness	**3**	**Fair**
Accountability Policies*	**2 (3)**	**Poor (Fair)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on Maine's Learning Results Document in Mathematics and English Language Arts; criterion-referenced Maine Educational Assessment (MEA), used in grades 4, 8, and 10. Reading tests were not made available for analysis. Go to www.accountabilityworks.org for more information.

The average quality of Maine's standards is poor. Middle school math is emblematic of the problem: They are lumped together instead of broken out by grade, most of them are overly broad, and while 69 percent of them can be at least loosely related to essential skills, only 21 percent of the essential skills are covered, among the lowest percentages in this study. Intelligibility is also a major problem. One standard, for example, instructs students to "Support reasoning by using models, known facts, properties, and relationships," which is an excellent example of a "disembodied standard," since it abstracts mathematical reasoning from specific mathematical subjects. Reading standards are only marginally better and suffer from the misguided assumption that students create meaning entirely separate from texts.

The state-developed criterion-referenced tests used in grades 4, 8, and 10, called the Maine Educational Assessment (MEA), provide a fair level of overall coverage. However, this assessment was based only on the math tests, as the state did not make the reading tests available for analysis. The elementary school math test is typical of the state's overall math content coverage, with few irrelevant or frivolous items and coverage of about one-half of essential grade-level skills. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and Maine's standards is also fair, ranging from the three-quarters of state standards covered by the

elementary school math test to the one-half of state standards covered by the high school math test.

The rigor of the tests was fair. While the elementary and middle school math tests earned a solid rating, with cut scores placed just a few points below what our analyses suggest is appropriate, the overall average was brought down by the very poor cut score for the high school math test, which was 25 percentage points below what was appropriate for the grade level. Overall, the testing system earned a fair rating for trustworthiness and openness, with a solid rating for internal consistency, but a lack of evidence on the reliability of the human scoring of open-ended items.

Maine's accountability policies prior to NCLB were poor. While Maine has a fair to solid assessment program in place, with student testing in a wide range of topics in grades 4, 8, and 11, there are no goals for schools or districts and no consequences for students or educators. Should the state implement NCLB requirements, we expect its accountability policies to improve to fair.

Massachusetts

Massachusetts has much to be proud of when it comes to academic standards: the Bay State's are the best we surveyed. Massachusetts' standards focus on important content at all three levels (elementary, middle, and high). For the most part they are exceptionally clear, specific, and measurable, such as the following grade 8 reading standard:

> Interpret mood and tone and give supporting evidence in a text. For example, students read excerpts from *A Gathering of Days*, by Joan W. Blos, a novel written in diary form of the last year a fourteen-year-old girl lived on the family farm in New Hampshire. Students write in their own journals and then discuss in groups how the difficulties of the year . . . are reflected in the writing's tone, and the extent to which detail in the writing helps the reader to understand and identify with the text.

Massachusetts		
Standards	**5**	**Outstanding**
Test Content	**3**	**Fair**
Alignment	**3**	**Fair**
Test Rigor	**4**	**Solid**
Test Trustworthiness and Openness	**4**	**Solid**
Accountability Policies*	**3 (4)**	**Fair (Solid)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on Massachusetts Learning Standards; criteria-referenced Massachusetts Comprehensive Assessment System (MCAS), used in grades 3-8 and 10. Go to www.accountability-works.org for more information.

The state is one of only two in our study to include a recommended reading list in the appendix, which is particularly helpful in guiding teachers, parents, and students. The Massachusetts list is especially praiseworthy.

At the elementary level, Massachusetts' math standards achieve the highest rating possible. The division between Measurement and Geometry is particularly appropriate beyond the 4th grade, and it gives implicit recognition to the study of math as a liberal art and not simply as an extension of the applied sciences. Massachusetts' avoidance of reliance on calculators ensures that students develop a deep familiarity

with numbers, the bedrock of math at the elementary and secondary levels. Middle and high school math standards, though not rated quite as high as the elementary standards, are still strong.

The state-developed criterion-referenced tests used in grades 3-8 and 10, called the Massachusetts Comprehensive Assessment System (MCAS), provide a fair level of overall coverage. The 4th grade math test covers three-quarters of essential grade-level skills. However, because of a large number of irrelevant or otherwise inappropriate items, the proportion of test items devoted to these essential skills is low. In addition, the test had a greater than typical percentage of item design flaws and poorly written directions. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and Massachusetts' state standards is fair, ranging from the one-third of state standards covered by the elementary school math test to the two-thirds of state standards covered by the high school reading test.

The rigor of the tests earned a solid rating, with cut scores often falling only a few points below what our analysis suggests is appropriate for the grade level. Particularly noteworthy is the fact that the MCAS is the only state assessment in math that received strong marks for rigor at every grade level assessed. Overall, the testing system earned a solid rating for trustworthiness and openness, with a particularly high rating for internal consistency, but the state falls a bit short in the availability of information on test security and administration procedures.

Massachusetts' accountability policies prior to NCLB were fair. While Massachusetts had the foundation of a sophisticated system in place, with a strong assessment program, the effectiveness of these policies was undermined by the lack of comprehensive performance goals and limited educator and student consequences. Should the state fully implement NCLB, we expect its accountability policies to improve to solid.

Michigan

We should state at the outset that while Michigan's ratings across the board are weak, leadership changes at the state department of education and the testing division in the last two years hold out the hope of improvement. For example, current efforts to elaborate the state standards and align state tests, if successful, would likely cause ratings to rise.

Michigan

Standards	**2**	**Poor**
Test Content	**3**	**Fair**
Alignment	**3**	**Fair**
Test Rigor	**N/A**	**N/A**
Test Trustworthiness and Openness	**2**	**Poor**
Accountability Policies*	**3 (4)**	**Fair (Solid)**

***(Scores after NCLB requirements in parentheses)**

As they stand now, Michigan's standards are poor—poorly written, overly vague, and replete with problematic content. The middle school math standards are typical of the state's overall efforts, with coverage of only about 40 percent of essential skills and many of the standards open to numerous interpretations. For example, students are instructed to "Use patterns and generalizations to solve problems and explore new content," which is clearly too vague to be of any practical value in learning math. Elementary school math

is even worse, with very poor coverage and esoteric instructions such as, "Recognize that change is often predictable, but variable, and that patterns emerge that help to describe the change."

The criterion-referenced tests used in grades 4, 7, 8, and 10 provide a fair level of content coverage. The best content is found in the elementary school reading test, which devotes over 90 percent of its items to essential skills and utilizes high quality reading passages with strong literary merit. The high school math test, in contrast, devotes only 20 percent of its test items to essential skills, with a focus on simple probability problems crowding out the testing of fundamental skills in algebra and geometry. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and Michigan's state standards is fair. The alignment for middle school reading is typical, covering about one-half of the state standards. The testing system has a poor overall rating for trustworthiness and openness, with a fair rating for internal consistency, but poor or very poor ratings for other critical areas such as test equating, rater reliability, and categorical reliability due to the very limited information provided. The state did volunteer to make available secure information for this analysis, signifying a rare commitment (in comparison to other states) to reviewing and improving present systems.

Michigan's accountability policies prior to NCLB were fair, with a solid effort to categorize schools based on performance undermined by the lack of serious educator consequences and the lack of interventions for low-performing students in early grades. Should the state implement NCLB requirements, we expect its accountability policies to improve to solid.

Analysis based on Michigan Curriculum Framework: Content Standards and Benchmarks; criteria-referenced Michigan Educational Assessment Program (MEAP) used in grades 4, 7, 8, and 10. It was not possible to assess the rigor of the state's tests because the necessary technical information was not made available. Go to www.accountability-works.org for more information.

Minnesota

Minnesota's standards for reading—the only standards we were able to analyze—are fair, generally covering about half of the essential skills. But they do occasionally lapse in vagueness: the high school standard for Reading, Listening, and Viewing Complex Information, for example, directs that "A student shall demonstrate the ability to comprehend and evaluate complex information in varied nonfiction by reading, listening, and viewing varied English language selections containing complex information," which is impossible to measure. Again, we would suggest a recommended reading list. Throughout, the document suffers from a confusing design that makes it hard to follow.

Minnesota		
Standards	**3**	**Fair**
Test Content	**3**	**Fair**
Alignment	**3**	**Fair**
Test Rigor	**4**	**Solid**
Test Trustworthiness and Openness	**4**	**Solid**
Accountability Policies*	**3 (4)**	**Fair (Solid)**

***(Scores after NCLB requirements in parentheses)**

Our analysis of test content, alignment, and rigor is based on the state-developed reading tests used in grades 5 and 10 (7th grade reading and any math tests were not made available for review). The Minnesota Comprehensive Assessment provides a fair level of content coverage overall. The high school reading test is typical of the state's overall content coverage, with the poor showing of covering only one-third of essential reading skills somewhat mitigated by the fact that more than 90 percent of

its test items address these essential skills, with very good reading passages. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The alignment between these tests and Minnesota's state standards is fair, ranging from about one-third coverage for the high school reading standards to just over two-thirds coverage for the elementary school reading standards.

The rigor of the elementary school reading test (5th grade)—the only test that could be evaluated on this dimension—was solid. This is a challenging instrument, with a cut score for proficiency set at 76 percent, exactly what our analysis suggests is appropriate for the grade level.

The testing system earned a solid overall rating for trustworthiness and openness, with a high level of internal consistency and sound evidence for the equating of tests from year to year. Finally, Minnesota's accountability policies prior to NCLB were fair, a rating that reflects the unevenness of the state's system. While Minnesota has solid assessment policies, they are undermined by a system of categorization that does not evaluate whether a school is meeting its goals or falling short. In addition, student incentives and consequences are limited to the requirement that students pass an exit test as a condition of receiving a high school diploma. Should the state implement NCLB requirements, we expect its accountability policies to improve to solid.

Analysis based on Minnesota Curriculum Standards (for reading only); state-developed Minnesota Comprehensive Assessment (MCA) in grades 5 and 10 (7th grade was unavailable for review). Scores for content, alignment, and rigor are based solely on reading tests, because no other tests were made available. Go to www.accountability-works.org for more information.

A dynamic new superintendent took the helm of the Minnesota department of education in 2003. Based on the speed with which she has developed a promising set of new standards (unfortunately, too recently to be included in this study), it seems likely that a brighter educational future is on the horizon for Minnesota than is reflected here.

Montana

Montana's standards are a very mixed lot, with some decent work dragged down by several sets of standards that are among the worst in the country. For example, the elementary reading standards are fair, with coverage of almost three-quarters of essential grade level skills and language that is for the most part specific and clear. But elementary math standards are simply atrocious. Barely one-third of the standards address worthwhile content, and then only marginally—coverage that is among the poorest in this study. Unaccountably, there is no mention at all of fractions, decimals, or percentages, and the standards are replete with vague, inflated, or confusing instructions. Middle and high school math are only marginally better. Generally, these math standards are not salvageable; a total rewrite is warranted.

Montana

Standards	**2**	**Poor**
Test Content	**4**	**Solid**
Alignment	**2**	**Poor**
Test Rigor	**2**	**Poor**
Test Trustworthiness and Openness	**4**	**Solid**
Accountability Policies*	**2 (3)**	**Poor (Fair)**

***(Scores after NCLB requirements in parentheses)**

The ITBS-A norm-referenced tests used by Montana in grades 4, 8, and 11 provide a solid level of content coverage overall. However, the quality varies depending on the particular

test, with the best content found in the elementary school math test, which covers over three-fourths of the most important math skills, and the worst found in the high school math test, which, despite a minimal number of item design problems, provides coverage of a paltry one-fifth of essential grade-level skills. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and Montana's state standards is poor. The alignment for elementary school reading is typical, with the test covering a little more than one-third of the state standards.

Analysis based on Montana Content and Performance Standards; ITBS norm-referenced tests in grades 4, 8, and 11. Go to www.accountabilityworks.org for more information.

The overall rigor of the test is poor, with the cut score often falling 20 or more percentage points below what our own analyses suggest is appropriate. For example, the cut score for proficiency in middle school math is 47 percent, while our independent analysis suggested that 80 percent was appropriate. The testing system has a solid overall rating for trustworthiness, but information was not readily available on the state's security procedures.

Finally, Montana's accountability policies prior to NCLB were poor. Despite solid assessment policies, Montana has a very poor system of performance goals and fails to categorize schools based on their performance. In addition, consequences for educators and students are minimal, with no statewide intervention program and no evidence to suggest the tests are used for any purpose other than diagnostically. Should the state implement NCLB requirements, we expect its accountability policies to improve to fair.

New Hampshire

The average quality of New Hampshire's standards is fair, with a tendency toward vague language undercutting generally good coverage of the most essential skills. For example, elementary math standards direct students to "Explore, discuss, and describe properties of common two- and three-dimensional figures" or "Explore situations involving probability." Exploration is not a standard as such, but an undefined process that the student may undertake during study. The standards for middle school math are typical of the state's overall approach, with standards that generally lack precision and provide coverage of only one-half of the essential grade level skills, but that provide better coverage for the most important of these skills.

New Hampshire		
Standards	**3**	**Fair**
Test Content	**3**	**Fair**
Alignment	**3**	**Fair**
Test Rigor	**N/A**	**N/A**
Test Trustworthiness and Openness	**4**	**Solid**
Accountability Policies*	**2 (3)**	**Poor (Fair)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on New Hampshire Curriculum Framework and Educational Improvement and Assessment Program in grades 3, 6, and 10. It was not possible to assess rigor because the state did not provide the necessary technical information. Go to www.accountabilityworks.org for more information.

The state-developed New Hampshire Educational Improvement and Assessment Program provides a fair level of overall coverage. Elementary school reading is typical of the state's overall efforts, with coverage of a little less than one-half of essential reading skills. Throughout, there are item design problems and an unconscionable number of grammatical mistakes. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and New Hampshire's state standards is also fair, with the tests on average covering about one-half of the state standards. The testing system earned a solid overall rating for trustworthiness and openness, with high marks

for internal consistency, the equating of tests from year to year, and the reliability of human scoring on open-ended questions.

Finally, New Hampshire's accountability policies prior to NCLB were poor. While New Hampshire has a solid assessment program, with students in grades 3, 6, and 11 tested across a range of subjects, it is undermined by the lack of positive incentives for educators or students and the failure of the state to provide resources to encourage additional support for low-performing students. Should the state implement NCLB requirements, we expect its accountability policies to improve to fair.

New Mexico

Across every grade and subject, New Mexico's standards are poor. Math standards, for example, show an extraordinary emphasis on the development of facile problem-solving strategies, an emphasis so pervasive that the reader is 20 standards into the document before encountering a straightforward goal. The standards are also very vague and would not give much guidance to parents or teachers, such as directions to "Analyze the relationship between geometry and measurement" or "Create and extend patterns through use of manipulatives" [*sic*]. These are simply not assessable tasks, and a host of very important objectives are not covered, including factoring and working with even and odd numbers. The absence of important topics is especially egregious since the standards at times grapple with such minutiae as "Analyze architectural designs that are inherently used in bridge designs and structural formations for strength, light, and esthetic value." Reading standards are no better, and in some ways even less clear. For example, high school reading directs students to take on such vague tasks as "Analyze the ideas of others by identifying the ways in which writers achieve a sense of completeness and wholeness" or "Use language, literature, and media to understand the role of the individual as a member of many cultures."

New Mexico			
	Standards	**2**	**Poor**
	Test Content	**3**	**Fair**
	Alignment	**2**	**Poor**
	Test Rigor	**3**	**Fair**
	Test Trustworthiness and Openness	**3**	**Fair**
	Accountability Policies*	**3 (4)**	**Fair (Solid)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on New Mexico Content Standards; CAT-6 norm-referenced tests used in grades 3-9. Here, analyses relating to tests are based on assessments used in grades 4 and 8. The criterion-referenced New Mexico High School Standards Assessments were not made available for review. Go to www.accountabilityworks.org for more information.

The CAT-6 norm-referenced tests used by New Mexico in grades 3-9 provide a fair overall level of content coverage. A major contribution to test quality is made by the 8th grade math test, which devotes most of its items to essential skills and consequently covers three-quarters of all of the essential grade-level math skills. However, this quality is not consistent throughout the tests, with the 4th grade reading test, for example, covering only one-half of essential reading and literature skills, containing reading passages that are not challenging, and often failing to test student knowledge of basic literary elements, such as plot, theme, characters, imagery, and setting. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and New Mexico's state standards is poor, with the tests consistently covering less than one-half of the state standards for the subject and grade level.

The overall rigor of the test is fair, with the "cut score" for proficiency often falling 5 to 15 percentage points below what our own analysis suggests is appropriate. For example, the cut score for proficiency in the elementary math test is 70 percent, while our independent analysis suggested that 82 percent was appropriate. (The cut score for grade 4 reading actually exceeds our expectations.) The testing system also has a fair overall rating for trustworthiness, with solid internal consistency and year-to-year test equating, but poor information on categorical and rater reliability.

Finally, New Mexico's accountability policies prior to NCLB were fair. It had a sophisticated system of growth targets for schools and performance goals that not only include targets for dropout rates and attendance, but also treat student performance in terms of median and mean scores instead of simply as the percentage of students reaching a cut point. The only area of weakness is the state's assessment program, which reported results only at the state level and not for specific schools or districts. If NCLB requirements are fully implemented, we expect New Mexico's accountability policies to improve to solid.

New York

Overall, New York's standards in both subjects are poor. Across the board, math standards are generally vague, poorly written, and have noticeable holes in the coverage of core math concepts. The overall coverage of the most important skills areas is very poor, with about one-third of the state standards addressing frivolous content, and only about one-third of the most essential skills covered. Standards intelligibility suffered from insufficient definition, as many standards are subject to a wide range of interpretation. Strangely, New York's math standards place greater emphasis on thinking, writing, and talking *about* math—so-called "math chat"—than on actual performance, and do not cover important topics in anything like a comprehensive manner. Reading standards are somewhat better than math standards.

Standards	**2**	**Poor**
Test Content	**4**	**Solid**
Alignment	**3**	**Fair**
Test Rigor	**3**	**Fair**
Test Trustworthiness and Openness	**4**	**Solid**
Accountability Policies*	**4 (4)**	**Solid (Solid)**

New York

***(Scores after NCLB requirements in parentheses)**

Analysis based on New York's standards in Curriculum, Instruction, and Assessment; state-developed criterion-referenced tests used in grades 4, 8, and Regents Exams in high school. Go to www.accountabilityworks.org for more information.

The state-developed criterion-referenced tests used in grades 4, 8, and in high school provide a solid level of overall coverage. The elementary school reading test is typical, with outstanding coverage of essential skills, consistently high quality task types, and reading passages that are commendable for their literary merit. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and New York's state standards is fair, but the alignment varies considerably depending on the grade level. For example, while the elementary math alignment is solid, with the test covering 70 percent of the standards, the alignment for elementary reading is poor, with the test covering only 20 percent of the standards. The rigor of the tests is also fair, with the cut score for proficiency 10 to 20 percentage points below what our analyses suggest is appropriate for the specific grade levels. For example, while the state recommends a "passing" score on the 11th grade reading test of

58 percent, our analysis suggests that 72 percent is appropriate. The testing system has a solid overall rating for trustworthiness, with high levels of internal consistency, year-to-year equating of tests, and the reliability of human scoring of open-ended items.

New York's accountability policies prior to NCLB were solid, with well-developed performance goals and an assessment system with reports on individuals, schools, and districts. The state falls a bit short in the area of consequences due to a lack of positive incentives for students. Should New York fully implement NCLB, its rating will remain unchanged.

North Carolina

The average quality of North Carolina's standards is fair, but that is a somewhat misleading average of widely varying scores for elementary, middle, and high school reading and math. For example, the state's middle school math standards are outstanding—better than our expectations in almost every way, written in clear language that designates specific learning goals, and consistently focused on the most important skills. But the state's middle school reading standards are poor, with vague directives such as, "Use the stance of a critic to: construct or review," and coverage of only one-third of essential skills.

North Carolina

Standards	**3**	**Fair**
Test Content	**4**	**Solid**
Alignment	**3**	**Fair**
Test Rigor	**N/A**	**N/A**
Test Trustworthiness and Openness	**2**	**Poor**
Accountability Policies*	**4 (5)**	**Solid (Outstanding)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on North Carolina Standard Course of Study; criterion-referenced tests in grades 4, 7, and high school. This evaluation is based only on the 4th and 7th grade tests; high school tests were not available for review. It was not possible to assess rigor because the state did not provide necessary technical information. Go to www.accountabilityworks.org for more information.

The state-developed tests used by North Carolina in grades 4 and 7 provide a solid level of overall coverage (high school tests were not available for review). The elementary school math test is typical, with coverage of about two-thirds of essential math skills, quality task types, and few design flaws. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and North Carolina's standards is fair, with math tests demonstrating solid alignment with state standards while the reading tests had poor alignment, covering only a little more than half of grade-level reading standards. The testing system earned a poor overall rating for trustworthiness and openness because of limited information. However, North Carolina tests did show a strong consistency with the data trends of the NAEP.

North Carolina's accountability policies prior to NCLB were solid. The system includes a comprehensive assessment program at every grade level beginning at grade 3, clear improvement goals for every school, and incentives and sanctions tied to school performance against those goals. The few gaps included a relative lack of attention to districts and no clear goals for the performance of important student subpopulations. Should the state implement NCLB requirements, we expect its accountability policies to improve to outstanding.

North Dakota

The average quality of North Dakota's standards is poor. The standards for high school reading are typical of the state's overall approach, with less than one-half of essential skills covered by the standards and the use of vague and extremely broad language, making it difficult to know exactly what students are meant to know and be able to do.

Standards	**2**	**Poor**
Test Content	**3**	**Fair**
Alignment	**3**	**Fair**
Test Rigor	**3**	**Fair**
Test Trustworthiness and Openness	**4**	**Solid**
Accountability Policies*	**2 (3)**	**Poor (Fair)**

North Dakota

***(Scores after NCLB requirements in parentheses)**

The CTBS-5 off-the-shelf norm-referenced tests used by North Dakota provide a fair level of content coverage overall. The 4th grade reading test is typical of the state's overall content coverage, with reading passages of appropriate length and difficulty but coverage of only a little more than one-half of the essential grade-level reading skills, and a high percentage of items addressing these skills. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and North Dakota's state standards is also fair, but the alignment varies depending on the grade level and subject. For example, while the alignment for middle school reading is solid, with the test covering three-quarters of the standards, the alignment for elementary school math is poor, with the test covering just over one-half of the standards.

Analysis based on the North Dakota Standards and Benchmarks in reading and math in effect in the summer of 2002. Draft revisions are on the verge of being adopted as this report is being published. Test analysis of the CTBS-5 norm-referenced tests used in grades 4, 8, and 10. Grade 6 tests used by the state are not part of this analysis. Go to www.accountabilityworks.org for more information.

The rigor of the tests is fair, but varies considerably depending on the subject and grade level. For example, while the cut score for proficiency in the 4th school math test is solid, falling only 2 percentage points below what our analysis suggests is appropriate, the cut score for the reading test at the same grade level is very poor, falling almost 30 percentage points short. The testing system has a solid overall rating for trustworthiness, with high levels of both internal consistency and the reliability of human scoring of open-ended items.

Finally, North Dakota's accountability policies were poor prior to NCLB. While the state had a fair assessment program, with testing in grades 4, 6, 8, and 10 covering NCLB-mandated subjects, there is no evidence that North Dakota had any framework of accountability and its system was designed with a near total lack of consequences for educators and students. If NCLB requirements are fully implemented, we expect the state's rating to rise to fair.

Ohio

Altogether, Ohio's standards are solid, coming up to our expectations in most areas. With a few exceptions, language is direct, coverage of essential skills is high, and standards are clear and measurable. A few small worries: reading standards show a tendency toward excessively trendy language, whereby a student "constructs" meaning on his or her own. And we would recommend the inclusion of a reading list to add to the specificity of these standards. But overall, a job well done.

Less well done are the state's criterion-referenced tests used in grades 4, 6, and 9, which provide a fair level of overall coverage. The high school reading test is typical of the state's overall content coverage, with less than half of essential reading skills covered, mitigated somewhat by the better coverage found in the most important skill areas. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and Ohio's state standards is also fair. The alignment for elementary school math is typical, with a loose association between most items and a standard, and coverage of about half of the state standards in total.

Ohio

Standards	**4**	**Solid**
Test Content	**3**	**Fair**
Alignment	**3**	**Fair**
Test Rigor	**3**	**Fair**
Test Trustworthiness and Openness	**3**	**Fair**
Accountability Policies*	**3 (4)**	**Fair (Solid)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on Ohio's Academic Content Standards; criterion-referenced tests in grades 4, 6, and 9. Go to www.accountabilityworks.org for more information.

The rigor of the tests is fair, with the cut score for proficiency falling 10 to 15 percentage points below what our analysis suggests is appropriate. For example, the cut score for the middle school reading test is 67 percent, while our analysis suggests that 85 percent is appropriate. The testing system also earned a fair overall rating for technical trustworthiness and openness. While the tests scored high marks for internal consistency and equating of test forms from year to year, there was a lack of evidence supporting the categorical reliability of the tests and there was an absence of technical data in general on the high school proficiency test.

Ohio's accountability policies prior to NCLB were fair. The state had a very strong system of performance categorization, placing districts into five well-defined and clearly delineated categories based on their achievement, but this system did not include individual schools. In addition, performance goals are based on the percentage of students passing without taking into consideration achievement gains for all students. Should the state implement NCLB requirements, we expect its accountability policies to improve to solid.

Pennsylvania

Pennsylvania's standards are fair, with elementary and middle school math warranting solid ratings. Generally, math standards are better than reading standards, which tend to be overly vague and broad (including one high school standard—"Read and understand works of literature"—that may be the single most absurd standard of any analyzed in this study). The standards for middle school math are typical of the state's overall approach, with the majority written in clear and distinct language that focuses on grade-appropriate knowledge. However, though none of these standards is brief and a few are quite prolix, they manage to cover only about half of the important grade level skills.

The state-developed Pennsylvania System of School Assessment (PSSA) tests used in grades 5, 8, and 11 provide a solid level of overall coverage. The middle school reading test is typical of the state's overall content coverage, with coverage of nearly two-thirds of essential reading skills, but containing reading pas-

sages that are often too easy for the grade level and have little literary merit. The middle school math test, however, has perhaps the best content coverage of any assessment included in this study, though there are problems with the wording of some items. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and Pennsylvania's state standards is also solid. The alignment for high school math is typical, with three-quarters of the standards addressed in the state test either directly or indirectly.

Pennsylvania		
Standards	**3**	**Fair**
Test Content	**4**	**Solid**
Alignment	**4**	**Solid**
Test Rigor	**2**	**Poor**
Test Trustworthiness and Openness	**4**	**Solid**
Accountability Policies*	**3 (4)**	**Fair (Solid)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on Pennsylvania Academic Standards; criteria-referenced Pennsylvania System of School Assessment (PSSA) used in grades 5, 8, and 11. Go to www.accountability-works.org for more information.

Having developed these generally high-quality tests, however, Pennsylvania sets the cut score far too low, often 10 to 20 percentage points below what our analysis suggests is appropriate, resulting in a poor rigor rating. For example, the cut score for passing the elementary school reading test is 58 percent, while our analysis suggests that 76 percent would be more appropriate. The testing system earned a solid overall rating for trustworthiness and openness, with high marks for internal consistency and the reliability of human scoring of open-ended items.

Pennsylvania's accountability policies prior to NCLB were fair. While Pennsylvania has solid assessment policies with extensive information on student performance in sub-disciplines, the framework for student consequences is poor, with the specific consequences for high school students left to the discretion of districts. Should the state implement NCLB requirements, we expect its accountability policies to improve to solid.

Rhode Island

No two ways about it: when it comes to standards-based accountability, Rhode Island is pretty much a mess. The Ocean State's standards are poor, with some standards—such as elementary math and reading—being very poor. (Elementary math includes a directive to "Have an intuitive understanding of whole numbers," which can't even be called a standard, properly understood.) At each grade, there is little in the reading standards' supporting text that is of use to instructors: no recommended reading lists, classroom scenarios, or activities, etc. The standards for middle school math are typical of the state's overall approach, with only one-third of essential skills covered by standards that are extremely vague and susceptible to problematic content, including vague directives such as, "Identify and justify an appropriate representation for a given situation."

Rhode Island		
Standards	**2**	**Poor**
Test Content	**3**	**Fair**
Alignment	**2**	**Poor**
Test Rigor	**N/A**	**N/A**
Test Trustworthiness and Openness	**1**	**Very Poor**
Accountability Policies*	**2 (3)**	**Poor (Fair)**

***(Scores after NCLB requirements in parentheses)**

The New Standards tests used in grades 4, 8, and 10 provide a fair, bordering on poor, level of content coverage overall and have the dubious honor of being the lowest-rated of all off-the-shelf tests reviewed in this study. The middle school reading test exemplifies the problems, with coverage of only one-third of essential reading skills and very low coverage of the vocabulary identified as high priority for the grade level. The high school test instructions are often open to broad interpretation and reading passages display mundane and formulaic choices of low interest to high school students. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and Rhode Island's state standards is also poor. The alignment for elementary school math is typical, with only about one-half of the state standards tested either directly or indirectly. It was not possible to assess the rigor of the state's tests because the necessary technical information for the New Standard tests it uses was not available (even though New Standards is sold commercially by a test publisher!). The testing system has a very poor overall rating for trustworthiness and openness, due largely to the fact that technical manuals were not available for review (a basic professional responsibility of any testing program) and to the lack of data suitable for comparisons with NAEP trends.

Analysis based on Rhode Island's Standards and State Frameworks; New Standards norm-referenced tests used in grades 4, 8, and 10. It was not possible to assess the rigor of the state's tests because the necessary technical information for the New Standards tests was not made available. Go to www.accountabilityworks.org for more information.

To close out this sorry tale, Rhode Island's accountability policies prior to NCLB were poor. Despite a fair system of performance goals and a fair set of educator consequences, the overall rating is dragged down by the failure to classify schools in terms of achievement results and a practically nonexistent framework of student consequences. Should the state implement NCLB requirements, we expect its accountability policies to improve to fair.

South Dakota

South Dakota

Standards	**3**	**Fair**
Test Content	**N/A**	**N/A**
Alignment	**N/A**	**N/A**
Test Rigor	**N/A**	**N/A**
Test Trustworthiness and Openness	**N/A**	**N/A**
Accountability Policies*	**2 (3)**	**Poor (Fair)**

***(Scores after NCLB requirements in parentheses)**

South Dakota's standards are fair, but that is the average of widely variant ratings across all levels and subjects. Math standards are generally better than reading—especially in high school—with generally fair coverage and intelligibility. Reading standards suffer from broad, vague language, little supporting text, and few classroom scenarios or activities. Vague language abounds, such as the directive to high schoolers to "use the reading process to understand directions and procedures."

Analysis based on South Dakota Content Standards. Evaluations of the state's tests are not included because the tests have changed since the analysis was performed. Go to www.accountabilityworks.org for more information.

Our evaluations of elements relating to the state's tests are not included in this report, because the tests have changed since the analysis was performed. An analysis of the test that has been replaced, the SAT-9 tests used in grades 4, 8, and 11, is included in the "Further Details" section available on the web at www.accountabilityworks.org. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.)

South Dakota's accountability policies prior to NCLB were poor. While the state has a fair assessment program, with tests in nearly every grade that include subjects beyond reading and math, it had no performance goals for schools or districts, did not categorize schools or districts based on performance, had only a limited system of educator consequences, and no system of student consequences. Should the state implement NCLB requirements, we expect its accountability policies to improve to fair.

Texas

The average quality of Texas' standards is fair. The best are the end-of-course high school reading standards, which are clear, specific, and measurable, and almost entirely focused on essential skills. More typical of the state's overall efforts are middle school reading standards, which are sufficiently clear, but cover less than one-half of grade level essential skills. Throughout, the standards suffer from poor organization that makes them hard to follow. The document also includes a "Key to Understanding Standards" that is itself hard to follow and is of very little assistance in understanding the standards.

Standards	**3**	**Fair**
Test Content	**N/A**	**N/A**
Alignment	**N/A**	**N/A**
Test Rigor	**N/A**	**N/A**
Test Trustworthiness and Openness	**N/A**	**N/A**
Accountability Policies*	**4 (4)**	**Solid (Solid)**

Texas

***(Scores after NCLB requirements in parentheses)**

Our evaluations of elements relating to the state's tests are not included in this report, because the tests have changed since the analysis was performed. An analysis of the test that has been replaced, the Texas Assessment of Academic Skills (TAAS), is included in the "Further Details" section available on the web at www.accountabilityworks.org. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.)

Analysis based on Texas Essential Knowledge and Skills. Our evaluations of elements relating to the state's tests are not included in this report, because the tests have changed since the analysis was performed. Go to www.accountability-works.org for more information.

Texas' accountability policies prior to NCLB were solid, with a comprehensive testing system focused on a broad range of topics, a clear and useful system of performance evaluations, and an incentive program in which schools that achieve outstanding results can receive financial rewards. Should Texas fully implement NCLB requirements, we expect its accountability policies to maintain their solid rating.

Vermont

Vermont's system of standards, testing, and accountability has serious problems across the board. Middle school math standards are the best, but most standards across all grades and subjects are poor. The standards for elementary math are typical, with nearly two-thirds of essential skills covered, undermined by vague and overly broad language that limits the usefulness of the standards themselves, such as "Create

and use a variety of strategies and approaches to solve problems, and learn approaches that other people use." Elementary reading is especially bad, with confusing and tedious supporting text. Reading and some literature standards are also poorly organized and unaccountably separated from other literature standards by a large body of text consisting of standards for other disciplines.

Vermont

Standards	**2**	**Poor**
Test Content	**3**	**Fair**
Alignment	**3**	**Fair**
Test Rigor	**N/A**	**N/A**
Test Trustworthiness and Openness	**1**	**Very Poor**
Accountability Policies*	**2 (4)**	**Poor (Solid)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on Vermont Framework of Standards and Learning Opportunities; New Standards norm-referenced tests used in grades 4, 8, and 10. It was not possible to assess the rigor of the state's tests because the necessary technical information for the New Standard tests was not made available. Go to www.accountability-works.org for more information.

The New Standards tests used in grades 4, 8, and 10 provide a fair—bordering on poor—level of content coverage overall and hold the dubious distinction of having the worst ratings of all off-the-shelf tests reviewed in this study. The middle school reading test exemplifies the problems, with coverage of only one-third of essential reading skills and very low coverage of the vocabulary identified as high priority for the grade level. The high school test instructions are often open to broad interpretation and reading passages display mundane and formulaic choices of low interest to high school students. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and Vermont's state standards is fair. The alignment for elementary school reading is typical, with 60 percent of the state standards tested either directly or indirectly. It was not possible to assess the rigor of the state's tests because the necessary technical information for the New Standard tests was not available (even though New Standards is sold commercially by a test publisher!). The testing system has a very poor rating for trustworthiness and openness, due largely to the fact that technical manuals were not available for review (a basic professional responsibility of any testing program) and to the lack of data suitable for comparisons with NAEP trends as a result of the state's limited participation in the testing program.

Vermont's accountability policies prior to NCLB were poor. Despite a fair regimen of assessment and performance categorization, the system was undermined by limited consequences for educators and no consequences for students. Should the state implement NCLB requirements, we expect its accountability policies to improve to solid.

Virginia

Virginia

Standards	**3**	**Fair**
Test Content	**3**	**Fair**
Alignment	**4**	**Solid**
Test Rigor	**3**	**Fair**
Test Trustworthiness and Openness	**4**	**Solid**
Accountability Policies*	**3 (4)**	**Fair (Solid)**

***(Scores after NCLB requirements in parentheses)**

The average quality of Virginia's standards is fair. Elementary school math is by far the best; there, despite some deficiencies in pre-algebra and geometry, an impressive three-fourths of the state standards are related to essential content, especially to number sense. The standards for middle school reading are more typical, with most of the standards written in sufficiently clear language and coverage of a little more than half of

the essential grade level skills. The reading standards are sometimes marred by irrelevant directives such as, "The student will become a skillful interpreter of the persuasive strategies used in mass media," found in middle school reading. The addition of a reading list would greatly enhance the intelligibility of the literature standards.

The Standards of Learning tests used in grades 3, 5, 8, and as end-of-course tests in high school provide a fair level of overall coverage. The 8th grade reading test is typical of these tests, with coverage of a little more than one-half of essential skills, consistently high quality task types, and solid reading passages of a length and difficulty appropriate for the grade level. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and Virginia's standards is solid, with most tests covering more than three-quarters of the grade-level state standards. The end-of-course high school algebra exam is particularly impressive in this respect, covering an outstanding 95 percent of the state's Algebra 1 standards.

However, the rigor of the tests is only fair, with the cut score for proficiency often falling 20 percentage points below what our analyses suggest is appropriate for the specific grade levels. For example, the same high school algebra exam that provides such outstanding coverage allows a passing score for students who answer only 54 percent of the questions correctly. Our analysis suggests that 72 percent is appropriate given the difficulty of the test. The testing system has a solid overall rating for trustworthiness, with high levels of internal consistency and year-to-year equating of tests, and an outstanding level of categorical reliability (which determines whether students are meeting achievement levels).

Virginia's accountability policies prior to NCLB were fair. While the state did a solid job of categorizing schools based on performance and has put in place a solid assessment program, the system paid no attention to year-to-year improvements. Should the state implement NCLB requirements, we expect its accountability policies to improve to solid.

Analysis based on Virginia Standards of Learning; criterion-referenced Standards of Learning tests used in grades 3, 5, 8, and as end-of-course tests in high school. Go to www.accountability-works.org for more information.

Washington

The average quality of Washington's standards is poor, a rating that is an average between the generally fair standards for reading and the generally very poor standards for math. The standards for elementary school reading are typical of the state's overall reading standards, with three-quarters of essential reading skills covered by standards that are usually clear, specific, and measurable. Middle school reveals some of the many problems of the state's math standards, which are meant to be read in conjunction with "Benchmarks" that do little to clarify the standards. For example, standard 4.1 reads, "Gather information." Bad enough, but then the

Washington		
Standards	**2**	**Poor**
Test Content	**3**	**Fair**
Alignment	**N/A**	**N/A**
Test Rigor	**N/A**	**N/A**
Test Trustworthiness and Openness	**3**	**Fair**
Accountability Policies*	**2 (3)**	**Poor (Fair)**

***(Scores after NCLB requirements in parentheses)**

Analysis based on Washington's Essential Academic Learning Requirements; ITBS tests in grades 3, 6, and 9 (criterion-referenced tests used in grades 4, 7, and 10 were not made available for review). It was not possible to assess the rigor of the state's tests because it does not report ITBS test results by proficiency level. It was not possible to assess the alignment between the state's standards and its tests because it would not provide the necessary technical information. Go to www.accountabilityworks.org for more information.

benchmark adds, "Develop and follow a plan for collecting information." What this has specifically to do with math, as opposed to any other human endeavor, is unclear. Only slightly better is standard 5.1, which reads, "Relate concepts and procedures within mathematics." The benchmark essentially repeats this vague standard, saying, "Relate and use conceptual and procedural understanding among a variety of mathematical content areas." High school math is typical of the breadth of the state's math standards, with only one-quarter of essential grade level math skills covered.

Washington State uses the ITBS tests in grades 3, 6, and 9, and a state-developed criterion-referenced test in grades 4, 7, and 10. Since the state tests were not made available for review, the content coverage analysis applies only to the ITBS tests, which were generally fair. The 3rd grade reading test was typical of the overall quality of the tests, with coverage of 58 percent of essential grade level reading skills. Reading passages were generally of high quality but were a little too short for the grade level and disproportionately emphasized some less important skills, such as determining the answer to inference questions. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The testing system has a fair rating for trustworthiness and openness, with high marks for test reliability and year-to-year test equating offset somewhat by the lack of documentation on categorical reliability—the consistency of results described by proficiency level.

Washington State's accountability policies prior to NCLB were poor. Despite an extensive assessment program in which a mix of state-developed tests and ITBS tests cover reading, math, writing, and communications at various grade levels, there was no system of educator consequences and no real consequences or positive incentives for students. Should the state implement NCLB requirements, we expect its accountability policies to improve to fair.

West Virginia

The average quality of West Virginia's standards is fair, with solid standards in elementary and high school math and poor standards in high school reading. The standards for elementary school math are typical of the state's overall approach, with two-thirds of the well-written standards addressing appropriate content and just over half of essential grade level skills covered. Coverage is reduced in part by the limited number of standards that address core pre-algebra and geometry skills. Seventy-three percent of high school reading standards address the most essential skills and concepts, but the standards are so broad and vague that it is difficult to understand exactly what students are meant to know and be able to do. There are also a number of irrelevant items, such as the directive to "identify common careers found in fiction (novels) and evaluate their desirability." The high school math standards are a model of intelligibility, however, such as the directive to "sim-

West Virginia

Standards	**3**	**Fair**
Test Content	**3**	**Fair**
Alignment	**3**	**Fair**
Test Rigor	**2**	**Poor**
Test Trustworthiness and Openness	**3**	**Fair**
Accountability Policies*	**3 (3)**	**Fair (Fair)**

***(Scores after NCLB requirements in parentheses)**

plify numerical expressions and evaluate algebraic expressions using grouping symbols and order of operations."

The SAT-9 tests used by West Virginia in grades 4, 8, and 10 provide a fair level of content coverage overall. The best test content is found in middle school math, with a very high proportion of test items focused on essential 8th grade math skills, while the worst is found in high school reading, which provides poor coverage of 10th grade reading skills and utilizes uninteresting and unclear reading passages. (See page 39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and West Virginia's state standards is fair. The alignment for high school reading is typical, with the test covering about one-half of the state standards.

Analysis based on West Virginia's Instructional Goals and Objectives; SAT-9 tests used in grades 4, 8, and 10. Go to www.accountability-works.org for more information.

The quality of these tests, however, is undercut by the low cut score set by the state, about 20 percent lower across the board than is appropriate for the specific grade level. The high school reading test is typical, with the state using 66 percent as the passing grade while our analysis suggested that 83 percent is appropriate. The testing system has a fair overall rating for trustworthiness and openness, with math and reading scores on the test relating reasonably well with NAEP score trends for the state.

West Virginia's accountability policies prior to NCLB were fair, with a solid assessment program grounded in annual student testing in reading and math between grades 1 and 11 undermined by a merely fair system of educator consequences and a non-existent system of student consequences. Should the state implement NCLB requirements, we expect its accountability policies to maintain their fair rating.

Wisconsin

The average quality of Wisconsin's standards is fair. The standards for elementary school math are typical of the state's overall approach, with over three-quarters of state standards focused on appropriate skills and nearly half of essential grade level skills covered. The standards are well structured and generally well written, but sometimes exhibit problematic content, such as elevating use of a calculator to the level of a math standard. Reading standards, especially in elementary and middle school, are of generally better quality, though we would suggest the inclusion of classroom scenarios, reading lists, and illustrative examples to improve intelligibility of the documents.

Wisconsin		
Standards	**3**	**Fair**
Test Content	**3**	**Fair**
Alignment	**3**	**Fair**
Test Rigor	**2**	**Poor**
Test Trustworthiness and Openness	**4**	**Solid**
Accountability Policies*	**3 (4)**	**Fair (Solid)**

***(Scores after NCLB requirements in parentheses)**

The CTBS-5 norm-referenced tests used by Wisconsin in grades 4, 8, and 10 provide a fair level of content coverage overall. The 4th grade reading test is typical of the state's overall content coverage, with reading passages of appropriate length and difficulty and coverage of a little more than half of essential grade-level reading skills, but a higher level of coverage for the most important of these skills. (See page

Analysis based on Wisconsin Model Academic Standards; CTBS-5 norm-referenced tests in grades 4, 8, and 10. Go to www.accountabilityworks.org for more information.

39 for more on the "reference standards" used in evaluating state standards and test content.) The overall alignment between these tests and Wisconsin's state standards is also fair, with the test on average covering about one-half of the grade level subject standards. Like many other states, however, Wisconsin undercuts these tests with low rigor, setting the cut score for proficiency 10 to 30 percentage points below what our analysis suggests is appropriate. For example, the cut score for the 8th grade reading test is set at 62 percent, while our analysis suggests that, for the grade level and difficulty of the test, 87 percent is appropriate. The testing system has a solid rating for trustworthiness, with high levels of both internal consistency and the reliability of human scoring of open-ended items.

Wisconsin's accountability policies prior to NCLB were fair. While the state had a solid assessment program, with testing in grades 4, 6, 8, and 10 covering a wide range of subjects, the only real intervention required by the state was the development of a school improvement plan and there is no evidence that this approach has ever raised the performance of a school. Should the state implement NCLB requirements, we expect its accountability policies to improve to solid.

Methods and Procedures

Note: More extensive information on study methodology, the rating system, and expert reviewers is available on the web at www.accountabilityworks.org.

Ratings

Individual ratings in each category and subcategory consist of whole numbers on a 1-5 scale, with "5" the highest score possible and "1" the lowest, indicating the following:

- **Five: meeting or exceeding the study's high expectations (outstanding performance),**
- **Four: largely meeting the study's high expectations, though significant gaps remain (solid performance),**
- **Three: partial achievement of the study's high expectations (fair performance),**
- **Two: only minimal achievement of the study's high expectations (poor performance), and**
- **One: little or no achievement of the study's high expectations (very poor performance).**

Analyses that yielded decimal scores were rounded to the nearest whole number for individual state ratings but multi-state averages are often reported to the nearest tenth of a point (e.g., 4.1).

Standards

State standards were evaluated separately in reading and math at elementary, middle, and early high school grades (where possible, 4th, 8th, and 10th grades). Each set of standards was rated separately on two qualities, intelligibility and coverage, which were then averaged into an overall score. Reviewers rated each quality on several criteria.

Intelligibility

- Is the standard clear, specific, measurable, and free of needless jargon?
- Is the language of the standard substantive and definite, leaving no doubt of what the boundaries are (i.e., of what is being asked of the student or teacher)?

- Does the standard contain misguided or trendy content or occurrences of plain error?

Ratings were lowered if there were consistent errors throughout the standards, inclusion of confusing explanatory material, or because of other systematic problems. A solid rating of "4" in intelligibility reflects an average between 3.5 and 4.4 of the intelligibility scores for individual standards, with overall adjustments as appropriate.

Coverage

- How well do the standards in a particular grade and subject cover the essential skills for that grade and subject?
- To what extent do the standards in a particular grade and subject focus on essential math or reading (including literature) content? For example, do they provide excessive attention to topics of lesser significance?

Standards are mapped to a set of "reference standards" developed by AccountabilityWorks, which cover the essential skills in each area and at all grade levels, according to our expert reviewers. Reference standards provide detailed descriptions of important math and reading skills at three academic milestones (late elementary, late middle school, and the middle of high school). Their level of specificity would allow the educated layman or teacher to clearly distinguish between two or more skills within the subject area and grade. They are strictly limited to subject content and do not address how the content is to be taught, the types of materials used in instruction, or the amount of time devoted to any particular topic.

Standards were further organized according to three priority levels that reflect the importance of each skill to developing an overall mastery of the subject. The definition and priority level of each of the reference standards were set by the raters and reviewed by the panels in reading and math, comprised of experts in those subjects, as well as curricular design. The reference standards were used in the evaluation of state standards content and test content coverage to determine the extent to which the most important skills were addressed. (Further information on these reference standards is available upon request.)

From these mappings, the percentage of state standards mapping to reference standards and the percentage of reference standards covered are computed and averaged. For example, if a state has 20 standards at a given grade and subject, of which 14 map to the reference standards (70 percent), resulting in 60 percent of the reference standards at the grade and subject being covered, the raw coverage score would be 65 percent. Based on this number, as well as rater judgment, a final coverage rating of 1 to 5 is assigned. A solid rating of "4" in coverage is indicated by a raw score between 60 and 69 percent.

Test Content

The test content raters evaluated each test on two general criteria: a rating for overall content coverage and either item design quality (math) or passage quality (reading).

Content Coverage

In math, the content coverage rating is the more heavily weighted of the two in the overall test content analysis. In reading, the content coverage rating and the passage rating are equally weighted. Several criteria were evaluated to determine the content coverage rating:

- Do the test items collectively address the most important math or reading (including literature) content, according to the reference standards?

- Does the test rely on the best math test item types, again based on the reference standards?

A solid rating of "4" in test content is indicated by a raw score between 60 and 69 percent.

Item Design Quality

- How many items are unclear or have other problems?

A solid rating of "4" in item design means only 1 to 4 percent of items were flagged for significant problems. This rating applies only to math tests, because there were very few item design flaws of a substantive nature in reading tests.

Passage Quality

- On tests, are the reading level and difficulty of reading passages, the variety of passage genres, and the literary quality of the passages appropriate to each grade level?

The highest ratings in passage quality would require reading levels challenging for the grade (though some below-grade passages are acceptable to assess accurately students functioning below grade level) and would include reading selections from works of American and English literature known for their literary quality and cultural significance. A solid rating of "4" in passage quality indicates a range of grade-appropriate passages as well as selections of appropriate literary quality.

Alignment

The alignment analysis determines the extent to which a test measures the knowledge and skills defined by the state standards. The alignment rating contains two components, "item alignment" and "standards coverage," each addressing a different dimension of alignment.

Item Alignment

- Does each test item map to one or more state standards, and how closely?

For example, if an item consists of a vocabulary question that requires students to use context clues to determine the meaning of an unknown word and the state has a well-defined standard that refers to this skill, the item would rate well (score of 2 on a 0-2 scale). If the state has a long list of content included within a standard, together with this particular skill, the item would receive most of the credit (1.5 on the same scale). If, however, the closest state standard refers to vocabulary but not to using context clues, the item would map only loosely to the standard and therefore receive only a middling score. If there are no state standards focused on vocabulary, the item would receive the lowest score. Item alignment is heavily impacted by the intelligibility and precision, or lack thereof, of the state standards. The individual item alignment scores for all items are averaged into a single item alignment raw score of 0-2. A solid rating of "4" in item alignment is indicated by an average raw score between 1.5 and 1.74.

Standards Coverage

- To what extent are the state's standards covered by the items on the test, and to what extent are the most *important* of the state's standards covered?

The analysis includes only those standards that can be readily assessed via a large-scale assessment ("assessable standards"); standards that are meant to be evaluated through classroom activities, such as how well a student discusses a work of literature, are not counted against the alignment score. The main consideration for standards coverage is a straight computation of the percentage of the state's assessable standards covered on its tests. Further, grades can be reduced or increased depending on how many of the most *important* standards are covered (according to the discretion of the rater). In reading, if more than 30 percent of test questions are based on passages that are significantly below the ostensible grade level of the standards, a reduction is applied to the coverage score (since the test should be aligned to the grade specified in the standards). A solid rating of "4" in standards coverage indicates between 80 and 89 percent of assessable standards covered on the test reviewed.

Rigor

- Is the cut score, or percentage of correct answers on a particular test the state requires for a student to be labelled "proficient," appropriate to the grade level and difficulty of the test?

We compare the cut score that the state sets for any particular test to the cut score recommended by the study raters for that test. The assumption employed by the study raters in establishing their recommended cut scores is the level of performance achievable on challenging subject matter by the great majority of students if they work hard and benefit from strong instruction (assuming they worked similarly hard and received similarly strong instruction in preceding grades). Thus, the target is a high—but achievable—standard based on the potential of the great majority of students. Rater analyses of test rigor were carefully reviewed by the principal investigators for consistency and fairness. A solid rating of "4" in rigor is indicated by a state-mandated cut score falling within +/- 5 percent of the appropriate cut score.

Testing Trustworthiness and Openness

A full technical evaluation of the quality of a testing program is a complicated and extended project that requires the inspection of the technical digest (usually published by the test developer) and lengthy interviews with the personnel involved in the test development process. The trustworthiness and openness ratings included in this study serve a different, but still valuable, purpose: to provide an indication of the level of confidence that the general public should have in the reported test results based upon a more limited inspection of the information available from the state and, if applicable, the test publisher. With respect to technical characteristics of the test, the emphasis is on those elements that are readily quantifiable and for which there is relative consensus on the criteria for success (e.g, test reliability). All of the information reviewed for this purpose is published or otherwise made available by the state (or vendors to the state) in an accessible form. Beyond the basic elements reviewed here, there are other critical features related to test development and administration, not reflected in these ratings, that may contribute—positively or negatively—to the overall quality of a testing program. Access to this kind of information and the conclusions that could be drawn from it are beyond the scope of this study.

The rating categories analyzed here reflect important test features on which information should be readily available for researchers and citizens alike. In instances where a state did not provide important information, its rating was lowered. While the absence of such information does not indicate that the technical quality of the testing program is necessarily deficient, it inevitably lowers public confidence in the results reported. Public confidence in a testing program cannot rely *solely* on the assurances of testing administrators or on private reviews by experts hired by the same officials, but must be based on openness to informed public scrutiny.

The following items, averaged together, yield a trustworthiness and openness rating:

Internal Consistency

Reliability is a quality of the consistency of test results. Several measures, or indices, of reliability tap consistency in different ways. Internal Consistency (IC) reliability indices measure the extent to which the items in a test perform consistently among themselves. The ratings here take into account the fact that tests with open response items tend to have somewhat lower reliability ratings. An IC index (in math and reading) of at least 0.85 is considered solid, and would receive a rating of 4 on a scale of 1 to 5.

Test Equating

Under this category, evidence is reviewed on adequate equating of tests. Equating fixes the several forms of a test to a common scale even when the difficulty of the forms is not perfectly equal. Horizontal equating applies to test forms at the same grade level. Vertical equating applies to the several test forms between grade levels. Equating procedures are highly technical. A thorough evaluation requires close inspection of results from the equating process; the criteria employed in this study do not go into such depth, but review evidence that proper technical procedures were followed. Points are assigned based on the following criteria:

- Is there documentation of procedures demonstrating that horizontal and vertical (where appropriate) equating has been utilized?

- Is there a specified and appropriate equating model?

- Is there documentation that the assumptions of the equating model have been met?

Inter-rater Reliability

Under this category, the evidence of inter-rater reliability (IRR) for manually scored items such as essays or other open response items is evaluated. Manually scored items require human judgment. This introduces a potential source of scoring error not present in machine-scored multiple-choice items (unless the scoring machine is mis-programmed!). There are several accepted indices for IRR. For example, an IRR correlation of at least 0.8 (on 0-1.0 scale) earned a rating of "4," with significantly higher index scores earning a higher rating and significantly lower index scores earning a lower rating.

Categorical Reliability

States typically report student results in different ways: on a numerical developmental scale; sometimes in terms of national norms; more recently, also in terms of whether students meet state achievement standards. Categorical reliability indicates the consistency of student results relative to state achievement standards. The categorical reliability of the test depends upon the general (IC) reliability as well as the position of the cut scores. There are several indices of categorical reliability with different criteria for strong performance for each. For example a Kappa index of 0.6 (on a 0-1.0 scale) is solid and earns a rating of "4." Other evidence of categorical reliability may be obtained from inspection of test reliability at each point along the scale.

Comparison Between State Assessments and the NAEP

Where available, multi-year trends on state assessments are reviewed for (broad) consistency with trends over the same period on the National Assessment of Educational Progress (NAEP) in the same grades and subject. A rating of "5" indicates similar trends based on a significant number of data points. "4" indicates similar trends based on limited data. "3" indicates trends that are muddled, neither sharply inconsistent nor very similar. "2" indicates trends that are inconsistent, though based on limited data. "1" indicates trends that are inconsistent based on a significant number of data points. Given the limited availability of data, there are sufficient data to assign a score in this category to only about a third of participating states, mostly based on 4th grade data. Still, this is a valuable piece of evidence regarding trustworthiness where it exists. With the NCLB mandate that all states participate in NAEP, more such data should become available over time.

Rating of Policies Regarding Security and Openness

This category includes the quality and availability of policies and materials related to test development and administration, including:

- Accessible and available test security policies,

- accessible and available test blueprint or framework, and
- accessible and available test developer contracts.

Accountability Policies

A set of criteria related to accountability and testing policy were developed in consultation with the Thomas B. Fordham Foundation. These criteria address state policies in the following areas:

Setting Performance Goals for Schools and Districts

- Does the state have annual performance goals for all schools and districts?
- Are non-achievement outcomes, such as the graduation rate, incorporated into the school and district performance goals?
- Are school and district performance goals broken out by at least four student subpopulation groups (i.e., ethnicity, special education status, limited English proficient status, and income status)?
- Do school and district performance goals address both absolute performance and "value-added"?

Assessment and Evaluation of Student Achievement

- Are state assessments conducted at least once in high school and grade-by-grade in grades 3-8, in both reading and math?
- Do assessments cover at least the core subjects?
- Are assessment results broken out by six student population subgroups (i.e., ethnicity, special education status, limited English proficient status, income status, migrant status, and gender)?
- Do students receive individual assessment results?
- Are schools and districts categorized, at the least, as meeting or not meeting performance targets?
- Does the state employ an explicit statistical model for ensuring that school and district categorizations are statistically valid and reliable?

Incentives and Consequences for Schools and Educators

- Are there interventions in schools or districts that consistently fail to meet performance targets?
- Do positive incentives exist for successful educators and administrators (i.e., certificates of recognition, financial bonuses, or other benefits)?

Incentives and Consequences for Students

- Do positive incentives exist for students who excel in meeting state academic standards?

- Are high school diplomas based on reliable indicators of student academic accomplishment, including exit examinations?

- Do other consequences exist for students who do not meet state standards as determined by state assessments (e.g., promotion withheld, summer school, etc.)?

Assistance to Struggling Students

- Is assistance required for students who are not meeting state academic standards as determined by state assessments?

A solid rating of "4" in accountability policies indicates that state policies fully meet the above criteria. The individual ratings for each criterion are averaged into ratings for each category or cluster, and those individual ratings are then averaged into an overall rating. All "pre-NCLB" ratings are based on the state policies in place by the spring/summer of 2002. All "post-NCLB" adjustments to the ratings are based on the state's accountability policies, as originally evaluated, *plus* the assumption of full compliance with relevant NCLB policy requirements.

Personnel and Procedures

Raters, Subject Review Panels, Expert Advisors

All underlying analyses were carried out by raters with expertise in the relevant content area and, where applicable, the grade levels being reviewed. Panels and advisors consisting of nationally recognized experts reviewed and approved the evaluation criteria (e.g., the "reference standards" in reading and math) and, where possible, sample analyses.

The analyses of standards, test content, alignment, and test rigor in reading and literature were performed by Bonnie Armbruster (reading researcher); Sheila Byrd (English teacher and standards developer); Carol Jago (English teacher); Jeanette Roberts (English teacher); and Kathleen Madigan (higher education professor and reading teacher). The reading review panel membership included Douglas Carnine (reading research policy); Sandra Stotsky (state standards and higher education professor); and Joanna Williams (reading researcher).

The analyses of standards, test content, alignment, and test rigor in math were performed by Patrick Coulton (mathematician); Kristin Umland (mathematician); Sara Tarver (math instructional researcher); David Wright (mathematician); Richard W. Cross (math teacher and higher education professor); Joseph Almeida (math teacher); Michael Augros (math teacher); Merle Farrier (higher education professor); Mark Langley (math teacher); and Damian Waterbury (math teacher). The math review panel members were James Milgram (mathematician); Richard Askey (mathematician); and Robert Dixon (math instructional design and research).

The testing trustworthiness analyses were performed by AccountabilityWorks research director and study principal investigator Richard W. Cross (testing and measurement). Review of the criteria for testing trustworthiness, as well as input on alignment, was provided by psychometric advisor Susan Phillips (psychometrician and attorney).

AccountabilityWorks research associate Thompson McFarland worked with president and study co-principal investigator Theodor Rebarber (accountability policy making, research, management) on the accountability policy analyses. Criteria were reviewed by Chester E. Finn, Jr. (policy making and research), Andrew Rotherham (policy making and research), and Marci Kanstoroom (policy making and research).

Justin Torres, Jeanette Roberts, and Marc Magee provided extensive editorial assistance in the preparation of the final report.

Common Procedures

States were selected for inclusion in this study on the basis of whether a copy of the state-administered test in reading and mathematics—or an alternate form—could be obtained for analysis (since most of the dimensions evaluated involved the state test). Tests and standards were sought at the following levels:

- Reading and math at the elementary level (4th grade, or the nearest available grade),
- Reading and math at the middle school level (8th grade, or the nearest available grade), and;
- Reading and math at the high school level (10th grade, or the nearest available grade).

Standards were obtained from each state's website, but few states post previously administered tests.

A letter requesting relevant study materials was sent to all 50 states and the District of Columbia at the beginning of the study. In the case of states that do not release their tests and for which no alternate test form was readily available, procedures such as non-disclosure agreements and travel to the state to review the tests on-site were offered; several states agreed to participate under such secure conditions. Where there was no response from a state, or an incomplete response, at least one additional follow-up attempt was made by phone or email to the state testing division. (Usually, multiple follow-up attempts were made.) Often, a form of the test could be obtained without assistance from the state, as in cases where a state posts on its website test forms from previous years or where an alternate form of a nationally norm-referenced test was available from the test publisher. In the case of states using both a custom standards-based test and a norm-referenced test, the standards-based test was the one evaluated if it could be obtained. Where only the norm-referenced test could be obtained but the state also has a standards-based test, the norm-referenced test was evaluated as the state test if results from its administration are widely publicized as part of the state's accountability system (and the absence of the custom test was noted). On these occasions, the norm-referenced test was only evaluated in applicable categories, such as test content, but not on its alignment to state standards. Where a state uses a norm-referenced test as its sole assessment instrument for a given level and subject, the norm-referenced test was evaluated in all available categories, including alignment to state standards.

Comparisons between grades are generally avoided unless they are especially large because different raters conducted analyses at different grades and, although precautions were taken to minimize rater bias overall (e.g., secondary reviews by the research director or, in some cases, other raters), there was no formal multi-rater reliability check. In the few cases where comparisons between grades are included in the findings, additional analysis was conducted to confirm the results. In the cases where comparisons are made between different test types (norm-referenced test vs. criterion-referenced test), the averages are made across the tests rather than across states (i.e., a norm-referenced test used in multiple states is not weighted more heavily in the results). References to terms such as "significant difference" do not imply any statistical test, but only refer to a qualitative estimation of the trends in the data.

Where possible (in nearly all cases), analyses of standards, test content, alignment, and test rigor were conducted "blind." All information related to authorship/vendor, state of origin, and type of test (norm-referenced test vs. criterion-referenced test) was stripped from the tests, standards, and other study materials before these were distributed to raters and other reviewers. This procedure was applied to all states except those—Colorado, Illinois, and New York—where a site visit was required to obtain access to the test; in those cases, only the standards were evaluated without identifiers.

While extensive efforts—unparalleled, we believe, for such studies—were made to develop and use consistent and replicable procedures, clear and explicit criteria, systematic analysis of resulting data, and internal review and confirmation of findings, the nature of the research domain necessarily involves a substantial element of expert judgment in the analyses. Another study based on substantially different criteria or a different philosophy might develop different results.

Additional information about study criteria, procedures, and methodology is available upon request. Please contact AccountabilityWorks at 202-261-2610, or by email at contact@accountabilityworks.org.

Selected Recent Publications

Publications are available electronically at our website at www.edexcellence.net. Single printed copies of most publications are available at no cost by calling 410-823-7474, or visit www.edexcellence.net/foundation/publication/index.cfm. Additional copies are $10.

Effective State Standards for U.S. History: A 2003 Report Card

This groundbreaking and comprehensive state-by-state analysis of K-12 education standards in U.S. history was prepared by Sheldon Stern, former historian at the John F. Kennedy Presidential Library. It evaluates U.S. history standards in 48 states and the District of Columbia on comprehensive historical content, sequential development, and balance. (September 2003)

Where Did Social Studies Go Wrong?

This report consists of penetrating critiques by renegade social studies educators who fault the regnant teaching methods and curricular ideas of their field and suggest how it can be reformed. While nearly everyone recognizes that American students don't know much about history and civics, these analysts probe the causes of this ignorance—and lay primary responsibility at the feet of the social studies "establishment" to which they belong. (August 2003)

Terrorists, Despots, and Democracy: What Our Children Need to Know

This report includes the voices of 29 political leaders, education practitioners, and cultural analysts who discuss what schools should teach about U.S. history, American ideals, and American civic life in the wake of 9/11, the war on terror, and the liberation of Iraq. (August 2003)

Charter School Authorizing: Are States Making the Grade?

This report from the Thomas B. Fordham Institute is the first significant study of the organizations that authorize charter schools. The report examines 23 states and the District of Columbia to determine how supportive they are of charter schools, how good a job their authorizers are doing, and how policy makers could strengthen their states' charter programs. (June 2003)

Better Leaders for America's Schools: A Manifesto

This report, published jointly by the Thomas B. Fordham Institute and the Broad Foundation, contends that American public education faces a "crisis in leadership" that cannot be alleviated from traditional sources of school principals and superintendents. Its signers do not believe this crisis can be fixed by conventional strategies for preparing, certifying and employing education leaders. Instead, they urge that first-rate leaders be sought outside the education field, earn salaries on par with their peers in other professions, and gain new authority over school staffing, operations, and budgets. (May 2003)