



The Princeton Review

Testing The Testers 2003

***An Annual Ranking of State
Accountability Systems***

The Princeton Review

Testing The Testers 2003

An Annual Ranking of State Accountability Systems

<i>Executive Summary</i>	<i>3</i>
<i>Background and Principles</i>	<i>5</i>
<i>Methodology</i>	<i>7</i>
<i>Why Weight?</i>	<i>8</i>
<i>How to Read the Results</i>	<i>9</i>
<i>Changes from Last Year</i>	<i>12</i>
<i>Moving Forward</i>	<i>13</i>
<i>Conclusion</i>	<i>15</i>
<i>Appendix I: Detailed Scores</i>	<i>15</i>
<i>Appendix II: Indicators, Rubrics, and Explanations</i>	<i>17</i>
<i>Appendix III: Changes from 2002</i>	<i>23</i>
<i>Appendix IV: The Question of Proficiency</i>	<i>32</i>

Executive Summary

During the Winter of 2002-2003, The Princeton Review conducted *Testing the Testers 2003*, our second Annual Ranking of State Accountability Systems. Unlike other studies, ours is not primarily concerned with the rigor of academic standards or of the tests that measure them. Rather we looked at the overall accountability system, focusing closely on policy: is it consistent, secure, open to public scrutiny, and flexible enough to improve over time? Will it tend to create better, more effective schools?

As the stakes for testing rise, and with the pressure of the Federal *No Child Left Behind* Act (NCLB), accountability systems increasingly affect what gets taught and how. As a result they will strongly influence how schools develop over the next several years. Simply put, good accountability systems will tend to result in better schools, and bad systems will create worse ones. The purpose of *Testing the Testers* is to highlight good and bad accountability practice with the hope of helping the overall tide to rise. By “good” we mean accountability systems that will lead not only to improvement on test scores as well as on other measures of school quality, that will support educator professionalization, make school a more satisfying and rewarding experience for students, and importantly, that will be able to improve and adapt as political and pedagogical realities change. Raising test scores is not that difficult if raising scores is all you want to do, and are willing to sacrifice the rest of what school means in order to do so. That, to us, would be bad accountability.

We collected data on twenty-two relevant indicators from each state and the District of Columbia. Each indicator was grouped in one of four major criteria and states received a score of either zero, one, or two points depending upon how their program performed. The criteria were:

<i>Academic Alignment:</i>	High-stakes tests are aligned to academic content knowledge and skills as specified by the states’ curriculum standards.
<i>Test Quality:</i>	The tests are capable of determining that those curriculum standards have been met.
<i>Sunshine:</i>	The policies and procedures surrounding the tests are open, and open to ongoing improvement.
<i>Policy:</i>	Accountability systems will tend to affect education in a way that is consistent with the goals of the state.

These criteria were weighted at 20%, 15%, 30%, and 35% respectively and the raw scores scaled accordingly to give each state and the District of Columbia a ranking from one to fifty-one (the highest possible weighted score was 100). Each state was also assigned letter grades on the A-F scale for each of the four criteria.

THE TOP RANKED PROGRAMS						
Rank	State	Weighted Score	Alignment	Test Quality	Sunshine	Policy
			20%	15%	30%	35%
1	NY	88.5	B+	A	B	A-
2	MA	85.7	B-	A	A-	B+
3	TX	84.3	B-	B+	A-	A-
4	NC	84.0	B-	A	B	A-
5	VA	81.7	A	A	B+	B-
6	LA	81.0	B-	A	B+	B+
6	FL	81.0	B-	A	B+	B+
8	AZ	80.2	B-	A	C+	A-
8	OK	80.2	B-	A	B	B+
10	CA	79.7	B+	A	B	B-

THE BOTTOM RANKED PROGRAMS						
Rank	State	Weighted Score	Alignment	Test Quality	Sunshine	Policy
			20%	15%	30%	35%
41	KS	58.2	D	A	C+	C+
42	IN	56.8	D	A	C	C+
43	HI	55.5	C-	B+	C-	B-
44	WY	54.5	F	A	C	B-
45	ND	54.3	C-	B+	C-	C+
46	WI	53.2	C-	A	C-	C+
47	WV	52.2	D	A-	F	B-
48	SD	49.8	B-	A	F	C
49	RI	48.5	C-	A	F	B-
50	MT	29.0	F	B-	F	C-

Only Virginia received two A's, and no state received an A for either of our most significant criteria, Sunshine and Policy. Nearly 30% of states received overall scores of 65 or lower, and of the individual grades given to the bottom-performing twenty states, nearly 40% were C or lower. On the positive side, forty-six programs received grades of B+ or better for the quality of the test instruments themselves.

Although the rankings are affected by the weighting we applied (especially for those states in the middle three quintiles) most states tend to do things well or poorly with some consistency across all indicators, regardless of weighting. Most reasonable weightings (including no weighting at all) do not drastically alter the composition of the top or bottom rankings. Rankings for unscaled scores are presented in the body of this report, and readers are encouraged to download the data spreadsheet from <www.princetonreview.com/stat-study> and formulate their own weightings and judgements.

Background and Principles

Launched in 1981, The Princeton Review is known to most people as the company that prepares students for college and graduate school admissions tests like the SAT, LSAT, and MCAT. Some also know us for the work we do in helping students and schools with other aspects of the college and graduate school admissions process. Four years ago we formed our K-12 Services Division to help schools and districts cope with the increasing pressure of state-mandated high-stakes accountability programs. These services center on providing educators around the country with tools and professional development to help align their classroom practice with state standards so that they do not have to resort to conventional test preparation in their classrooms. The experience we have gained working with schools on a national basis gives us an overview of the different ways of implementing accountability that is both broad and deep. We understand the tests as psychometric instruments, and we are witness to the effects of policy in the classroom.

Our positions on accountability derive from this perspective: we are deeply-versed in testing but are not ourselves a testing company. Further, since our customers are mostly schools and districts we have little incentive to seek favor with the state departments of education which design and implement accountability systems. As a result, we have both the expertise and freedom to speak out regarding flaws in state-imposed testing mandates.¹

In this study we use the term “accountability” to encompass more than just the tests. Tests in and of themselves can—at best—do no more than provide information. The meaning and the usefulness of that information is very much dependent on a set of policy choices, just as policy choices determined the nature of the tests to begin with. Still more policy choices determine the impacts of testing on the overall character of education in a state—both good and bad—as well on the actual experience of school for individual students, families, and educators. Those who feel strongly about testing one way or the other are generally reacting to accountability policy rather than the tests themselves, and it is the hybrid of tests and policy which *Testing the Testers* seeks to judge.

Our specific intent in conducting this study annually is to shine the same spotlight on states’ accountability programs that those programs claim to shine on student, teacher, and school performance. Those who design and implement accountability should themselves be accountable and open to scrutiny.

We believe that test-based accountability for students, schools and, ultimately, educators is going to be an important part of education for the next several decades, and we prefer that it be done well rather than poorly. We, therefore, approached this study primarily from a systems perspective. We asked, first, is an accountability program likely to further the goals it sets for itself without a lot of unintended consequences? Second, does it provide means for educators and families to actually improve learning? These perspectives are embodied in the indicators we chose to assess the quality of each state’s efforts, as well as in the relative weights we assigned to them. For example, we believe that greater degrees of openness and disclosure surrounding all aspects of a testing program will result in its becoming more robust, more effective, more reliable, and more politically sustainable. This is reflected in the relatively high weight we gave to the third criterion, “Sunshine.”

Lastly, it is important to note that this study is not and does not intend to be a ranking of the overall

1. Full disclosure: During the school year 2002-2003 our K12 Services Division had a pre-existing contract with the Massachusetts State Department of Education (ranked second in our study) to provide online and offline tutorial programs for some students, intended to improve their math and English/language arts skills as measured by the Massachusetts Comprehensive Assessment System (MCAS). During the same year our Admissions Services Division was in the midst of a four-year contract with the State of Hawaii (ranked forty-third) to provide college guidance software to high schools and conduct applicant outreach and process electronic applications for the state university. Finally, our Test Prep Division has a contract with the Kentucky Department of Education (ranked twenty-first) to provide online SAT preparation for the Kentucky Virtual High School.

quality of states' educational systems, of the rigor of their standards, of the adequacy or equity of their financing, or of their level of student attainment. Iowa, for example, is generally regarded as having excellent schools though they rank near the bottom on accountability. Mississippi, on the other hand, has an excellent system of accountability despite having what many would consider to be serious gaps in school quality.

Nor is *Testing the Testers* intended to track or mirror the degree of states' compliance with NCLB, though there is some overlap between our indicators and the law's requirements. Of the nine states whose plans had been approved by the Department of Education as of this writing, only three are in our top ten. Further, the fact that a plan has been approved does not mean it has been implemented, and we give credit only for policies in effect this year.

The criteria and indicators are listed below, along with the weight that each contributed to a state's ranking (percentages are rounded up and so sum to more than 100). Descriptions of the scoring rubrics used for each indicator along with some explanatory comments can be found in Appendix II. The full dataset of state responses can be downloaded from <www.princetonreview.com/statestudy>.

Criterion #1: Tests are aligned to academic content knowledge and skills as specified by the states' curriculum standards (20% of Rank).

- 1a) Are standards granular enough that a small number of test items can reasonably measure a student's mastery of that granule? (6%)
- 1b) Is there substantial overlap between the published curriculum standards and those that are actually tested? (8%)
- 1c) Do states allow schools or students to choose from among a number of tests?(6%)

Criterion #2: The tests are capable of determining that those standards have been met (15% of Rank).

- 2a) Are the items well written and the tests scored accurately and completely? (4%)
- 2b) Do the tests include multiple item types (e.g., multiple choice, open ended, computation, performance activities, etc.)? (5%)
- 2c) Are items validated before the test is assembled, and does someone other than the test developer review items before the test is constructed? (2%)
- 2d) Were the achievement cutoff points (and scoring curve, if applicable) established before the first live administration of the test? (2%)
- 2e) Are the cutoff points (and scoring curves, if applicable) for various grades and subjects consistent enough across subjects and years to enable comparison? (3%)

Criterion #3: The policies and procedures surrounding the tests are open, and open to ongoing improvement (30% of Rank).

- 3a) Are contract terms with the testing companies readily available for public inspection? (2%)
- 3b) Are detailed test specifications readily available? (5%)
- 3c) Is there a reasonable level of security around the test and scoring procedures, and are there due process guidelines for students and educators accused of cheating? (6%)
- 3d) Are complete test scores released to the public in a timely manner? (6%)
- 3e) Is the test released every year? (8%)

3f) Does the state publish detailed information on the disparate performance of different groups? (5%)

Criterion #4: Accountability systems affect education in a way that is consistent with the goals of the state (35% of Rank).

4a) Does the state track and judge schools by value-added analysis? (5%)

4b) Does the state judge schools by multiple indicators, such as graduation, promotion, attendance, and violence rates? (5%)

4c) Do tests have stakes for students, and do students have opportunities to re-take the test if necessary? (4%)

4d) Is test data regarding individual students distributed to educators and families in useful detail? (5%)

4e) Is school-level performance data shared with the public along with explanations and contextual detail appropriate for a general audience? (4%)

4f) Are support programs in place to assist students and schools in overcoming their deficiencies? Are there consequences based on the performance of these programs? (5%)

4g) Does the state give districts and schools the latitude to meet performance standards in reasonably flexible ways? (4%)

4h) Does the state maintain publicly available data warehouses to evaluate educational progress over time? (4%)

Methodology

Our ratings are based on information provided by each state, including legislation, press releases, overviews for parents, testing manuals, reports explaining the significance of test scores, and telephone interviews we conducted. We also relied on the findings of the American Federation of Teachers' study, *Making Standards Matter 2001* <<http://www.aft.org/edissues/standards/msm2001/>> for some information regarding the alignment of state standards to state assessments.

Upon conclusion of the data-gathering, we sent draft findings to each state's Director of Assessment for a review of their accuracy and completeness. Each state was invited to submit comments or corrections, and we then made any necessary changes in our analysis to reflect the additional or corrected information. As a quality control measure half of the corrected state results were then sent to local superintendents for verification that the policy on the ground in fact reflected the policy on the book. Any discrepancies were reconciled through further interviews. Six state departments of education—Iowa, Kansas, Maryland, Montana, Oklahoma and Tennessee—chose not to respond to our requests for confirmation, some emphatically so.²

With the data in hand we then awarded scores of zero, one, or two points for each indicator based on our judgment of each state's performance. If we could not find the required information and the state refused or was otherwise unable to provide it, we assigned a score of zero: after all, our underlying premise, and that of accountability in general, is that knowing is always better than not knowing. Still, while we consider silence and failure to be equivalent from a policy perspective others may not, and so we have tagged with an asterisk those

A	3.84 and higher
A-	3.50 and higher
B+	3.17 and higher
B	2.84 and higher
B-	2.50 and higher
C+	2.17 and higher
C	1.84 and higher
C-	1.50 and higher
D	1.17 and higher
F	-

2. Information for Arkansas and Oklahoma was however confirmed by local school districts.

zero ratings that were assigned because information was not forthcoming.

Finally, letter-grades for each criteria were assigned by equating the scores to the classic A-F distribution.

Why Weight?

Rather than simply rank programs by the sum of their raw scores we chose to apply a weighting framework to ensure that some criteria did not overwhelm others in determining outcomes. For example, Criterion #1: Alignment, *“Tests are aligned to academic content knowledge and skills as specified by the states’ curriculum standards”*, comprises three indicators while Criterion #4: Policy, *“Accountability systems affect education in a way that is consistent with the goals of the state”* comprises eight. Using raw scores would effectively give Policy nearly three times as much weight as Alignment in determining a state’s final rank.

At the same time, we believe that some criteria do make relatively greater contributions to robust and productive systems of educational accountability than others. Criterion #4, *“Accountability systems affect education in a way that is consistent with the goals of the state”*, is the very reason these programs exist, while Criterion #3, *“The policies and procedures surrounding the tests are open, and open to ongoing improvement”*, largely determines the system’s ability to heal and improve itself. There are flaws and limitations inherent in any complex institutional process, particularly one so significantly driven by political and social imperatives. We felt that giving extra weight to the Sunshine and Policy criteria was appropriate and thus they contribute a combined 65% of a state’s ranking.

We weighted the indicators that make up each criterion along similar lines, judging the proportional contribution of each to the broader technical and policy framework of its criterion. We believe, for example, that the use of value-added analysis is critical to making accountability systems both useful and fair. For this reason we weighted that indicator more heavily than, say, the use of consistent scoring curves across tests but gave it less weight than we did to a strong overlap between curriculum standards and what is actually tested.³

This subjectivity is not unusual: Weighting is an implied component of any state accountability system that reports more than one indicator of school quality or that uses multiple criteria to determine whether a student is promoted. Thirty-seven state-sponsored “school report cards”, for example, list inputs and outcomes other than test scores to help create a more complete picture of each school. They include things like graduation rates, percentage of certified teachers or of those with advanced degrees, absentee rates for students and staff, incidences of violence, and the like. These tend to be highlighted or deemphasized on report cards depending on what the state sees as their relative importance, and many states that hold schools accountable on multiple measures weight them differently to come up with an overall judgment.

It is interesting to note that, despite the attention we devoted to weighing the indicators relative to one another, the overall rankings, especially at the top and bottom, are not drastically affected by whether weightings or raw scores are used. States that perform well generally do so broadly, as do states that perform poorly. There are some significant exceptions, however. Ohio was most negatively affected by our

3. Arizona is alone among states in moving away from value-added analysis. Its new Achievement Profiles for high schools, released this past fall, no longer list data on One-Year Growth.

weightings, having been bumped 9 slots lower than its unweighted score would have earned, with Michigan and Minnesota next, losing eight steps each. Tennessee was the only state to benefit substantially from our weightings, gaining six steps, though the additional three steps gained by California enabled it place in our top ten.

Readers interested in fully separating our subjective weightings from the underlying data are encouraged to download our ranking spreadsheet and weight the indicators according to their own priorities.

How to Read the Results

It is not surprising that most of the top rankings are held, as they were last year, by states with the longest experience with test-based accountability. New York, Massachusetts, Texas, North Carolina, Virginia, Florida, and Arizona have spent many years and many tens of millions of dollars grappling with the myriad psychometric, logistical, and political issues with which other states are just beginning to struggle. Having worked out the kinks in the basics of accountability like curricular alignment and test quality, they are in a position to concentrate on policy enhancements. Relative newcomers like Louisiana and Oklahoma have clearly leveraged the experience of the early adopters to jump-start their own high-quality programs.

Accountability is clearly more difficult than mere testing: Thirty-nine states received A's for Test Quality, including four of the ten worst programs overall. But only Virginia (because of its leadership role in testing flexibility) received an A for Alignment, and no state received an A for either Sunshine or Policy. Strong scores in those criteria were scarce and heavily-concentrated: six of the seven awards of A- went to states in the top ten.

The fact that some states have been doing accountability better and longer than others does not mean it is any less contested in their communities. New York, Massachusetts, and Virginia continued to be roiled by protests and boycotts from academics and from parents in wealthy communities who feel strongly that test-based accountability impinges on the freedom of educators and parents to define and measure learning as they deem best.⁴ Texas, on the other hand, has consistently enjoyed broad support for the notion of test-based accountability, though there continues to be debate over how it is implemented and the true meaning of the rise in state-determined proficiency levels (see Appendix IV for more on this topic). Louisiana likes to point to its accountability system as evidence of its commitment to turning around a historically weak educational system.

Of course, there is always the Iowa exception. Alone among states, Iowa's official position has been that it ought to have only the most minimal role in determining performance standards or accountability systems for local schools. Philosophically opposed to state oversight, it has left nearly all matters regarding accountability to the districts themselves, a policy which enjoys wide public support (and which until fairly recently was solidly in the mainstream of America's unique local-control philosophy).⁵ Iowa's low scores on our measures are thus not the result of poor design or execution (as is the case with the other states in the bottom ten), but of conscious and consistent policy decisions. Iowa's system is less a failure than a conscientious objection and we felt it was inappropriate to include them in the rankings.

4. In communities where the public conversation is shaped by families of generally high-achieving students the sentiment is often against test-based accountability, which is seen as a form of teacher-bashing and a dumbing-down and narrowing of the curriculum. Support for testing appears much higher in poor communities with large numbers of under-performing schools and students. Parents there often see it as the sole form of quality control and academic discipline over their schools.

5. Indeed, our indicator 1c, Test Choice, rewards districts having the autonomy to choose their own tests. The other indicators presume though that this discretion will exist within a more uniform and coherent system of state oversight.

Nevertheless, Iowa's stance has some negative consequences. Leaving the issue of testing to districts with limited resources has meant that most students in Iowa take the Iowa Test of Basic Skills (ITBS) and the Iowa Tests of Education Development (ITED). It is then left to each district to translate those results into educational inputs. The practical consequence is that, while the tests are sound as instruments (Iowa received an A- for Test Quality), there is no statewide policy safety-net to ensure that the tests are well-administered or that the information they generate is used to improve educational outcomes. This inability to know for certain what policies each Iowa district is subject to is the reason that Iowa received so many scores of zero for "could not be determined", a full 25% of the total for all states combined. Other states in the bottom ten do not appear to share Iowa's philosophical objections. Rather, they have simply done a poor job thus far of designing and implementing the systems they deploy. In some instances this may be due to constraints imposed by the legislature. If resources are truly scarce it may appear cost-prohibitive to develop a test designed to map closely to the state's curriculum standards or to release the entire test each year, both of which are basic building blocks of a strong testing program.⁶ Some states have legislative restrictions sponsored by teachers' unions that prohibit the use of value-added analysis to determine year-over-year improvements at the classroom or building level.

Rank	State	Weighted total	Alignment	Test Quality	Sunshine	Policy	Unweighted Score (of 44)	State
		20%		15%	30%	35%		
1	NY	88.5	B+	A	B	A-	38	NY
2	MA	85.7	B-	A	A-	B+	38	MA
3	TX	84.3	B-	B+	A-	A-	38	NC
4	NC	84.0	B-	A	B	A-	37	TX
5	VA	81.7	A	A	B+	B-	37	LA
6	LA	81.0	B-	A	B+	B+	37	FL
6	FL	81.0	B-	A	B+	B+	36	VA
8	AZ	80.2	B-	A	C+	A-	36	AZ
8	OK*	80.2	B-	A	B	B+	36	OK*
10	CA	79.7	B+	A	B	B-	36	MS
11	SC	79.5	B-	B+	B-	A-	36	PA
12	MS	78.7	B+	A	B+	B-	36	MN
13	PA	78.0	B+	A	C+	B+	35	CA
14	UT	77.2	B-	B	B+	B+	35	SC
15	MN	77.1	B-	A	B-	B	35	UT
16	CO	76.7	B-	B+	B+	B-	35	IL
17	NV	76.5	B-	B+	B	B+	35	OH
17	TN*	76.5	B-	A	B-	B+	34	CO
19	IL	76.2	B-	A	B-	B	34	NV
20	ME	76.0	B-	A	B	B-	34	ME
20	OR	76.0	B-	A	B	B-	34	WA

6. The total cost of testing programs are typically between one-one-hundredth and one-one-thousandth of per-pupil spending, a small amount to pay for the information and, ideally, quality control it yields.

Rank	State	Weighted total	Alignment	Test Quality	Sunshine	Policy	Unweighted Score (of 44)	State
22	OH	75.8	C	A	B+	B	34	MI
23	KY	74.7	B-	A	B-	B	33	TN*
24	WA	74.5	B-	A	B-	B-	33	OR
25	AK	74.2	B-	A	B	B-	33	KY
26	MI	73.0	C-	A	B-	B+	33	-AK
27	ID	72.6	B-	A	B	B-	33	AR*
28	NJ	72.3	B-	B+	B-	B	32	ID
29	AR*	72.0	D	A	B+	B	32	NJ
30	CT	71.1	B-	A	B-	B-	32	CT
31	NE	70.0	A	A	D	B-	32	VT
32	VT	69.6	C-	A	B-	B	31	NE
33	AL	67.5	C-	A	B	B-	31	MO
34	MO	67.0	C-	A	B-	B-	31	MD*
35	MD*	66.2	D	A	C+	B+	31	NM
36	DE	65.6	C-	B	B-	B+	30	AL
37	NM	65.5	D	A	C+	B+	30	NH
38	NH	64.7	C-	A	C+	B-	29	DE
39	DC	62.5	C	A	C+	C+	29	DC
40	GA	61.7	B-	A	B-	C-	28	KS*
41	KS*	58.2	D	A	C+	C+	28	IN
42	IN	56.8	D	A	C	C+	27	GA
43	HI	55.5	C-	B+	C-	B-	26	HI
44	WY	54.5	F	A	C	B-	26	WY
45	ND	54.3	C-	B+	C-	C+	25	ND
46	WI	53.2	C-	A	C-	C+	24	WV
47	WV	52.2	D	A-	F	B-	23	WI
48	SD	49.8	B-	A	F	C	22	RI
49	RI	48.5	C-	A	F	B-	20	SD
	IA*‡	42.2	C+	B-	F	C+	19	IA*‡
50	MT*	29.0	F	B-	F	C-	16	MT*

*indicates states that chose not to validate their information

‡ Iowa is excluded from the rankings. See the discussion in this section

We should note that our rankings are fairly well-correlated with state size, about 0.50. Partly this is because the larger states have been doing accountability longer. Certainly, resources are also an issue: Montana simply cannot put the same dollars against these systems as New York. The answer to that problem, however, cannot be for smaller or less wealthy states to simply do a poorer job for their students. Rather, they should join together with other states in consortia that allow them to spread the burdens over a broader base.

These are political issues with political resolutions. One of our hopes for *Testing the Testers* is that parents will use it to bring pressure to bear on elected and appointed representatives to explore innovative approaches that will lead to the best accountability systems possible.

Changes From Last Year

Testing the Testers is intended to change over time to adjust to the world in which it operates. Since its purpose is to help drive changes in outlook or practice, it is appropriate to re-evaluate and re-set the bar each year. For example, *No Child Left Behind* requires states to build accountability systems with some of the same features which we consider positive and for which we award points. Since the potential loss of Title I funding (the ultimate sanction for failure to comply with NCLB) is already concentrating the minds of policy makers on these indicators, we will tend to de-emphasize or eliminate these over time so that other elements of good practice can be highlighted and rewarded.

Similarly, as the majority of states become competent at the basic mechanics of standard-setting and test administration we will de-emphasize these criteria in order to give more weight to policy differences. This year, for example, we decreased the weighting of Criterion #2, Test Quality, from 20% to 15% and increased that of Criterion #4, Policy, from 30% to 35%. Appendix III describes the changes in the composition and weighting of the indicators since last year.

Weighting and methodological issues aside, we're gratified to see a general improvement in the quality of state programs from last year across nearly all comparable indicators. Last year was one of significant transition for many states, who had specified but not yet fully-implemented important changes in many areas of their accountability systems. For example, on Indicator 1b, *Is there substantial overlap between the published curriculum standards and those that are actually tested?*, forty-one states received the highest score this year compared with only twenty-nine on our first study. Tennessee is perhaps the most dramatic example of this rising tide: significant improvements instituted during the tenure of recently-departed Commissioner Faye Taylor took effect this year, bringing the system she inherited up from a weighted rank of forty-one on last year's study to twenty-two today. It will be interesting to see whether her successor will be able to build on this record of improvement.

The general rise in quality is, ironically, responsible for much of the downward movement in the rankings. Where states have fallen relative to others it is usually not because their programs have gotten worse but, rather, because others have gotten much better or because a general rise in proficiency has driven a change in the composition or weighting of the indicators. Further, many more states accepted our invitation to review their data in detail this year and so it is probably a more accurate representation of actual accountability practice.

Careful readers may object that while we award points for states who employ value-added analysis of their data the annual changes in our indicators preclude such year-over-year comparisons of the states themselves. Fair enough. *Testing the Testers* is criterion-referenced, as it were, and those criteria will change each year as the general policy climate as well as the state of best practices, evolves. Even so, seven of the best and six of the worst from last year's study reappear this year.

Moving Forward

One of the interesting effects of NCLB has been a muting of many of the debates over accountability. Partly that is because having an accountability system is no longer a local or state-level option. Partly it is because the logistical aspects of NCLB compliance are so daunting in so many states. Partly, to be sure, it is because failure to comply with its numerous provisions or to make AYP goals has such unprecedented and immediate consequences. NCLB has concentrated the minds of state and local educators on what must be done but not, unfortunately, on what could be done. Future versions of *Testing the Testers* will attempt to move beyond rewarding basic proficiency or mere compliance with statute in order to more fully spotlight what we consider to be particularly sophisticated, effective, or humane approaches to rigorous systemic accountability.

Examples of these might include the efforts underway in California and, to a lesser extent, in Florida to align state high school exit and state college entrance exams. Certainly it makes little sense from the standpoints of pedagogy, accountability, or resource allocation to tell high school students that one test qualifies them to leave high school while another one, entirely unrelated, qualifies them to enter college. Rigorous course-end tests like the New York State Regents should be an acceptable substitute for the SAT or ACT, which say nothing about a student's academic abilities. In the Regents example, these might first be acceptable only to the New York State University system but there is no good reason why other state systems should not accept them as well. The equating of various state year-end or exit tests to one another for the purposes of college admissions is not a terribly complex or difficult problem. Ideally in our view, all states would seek to provide students and schools with the greatest degree of test choice (and hence of curricular choice) that is consistent with a reliable measurement of educational achievement. This is embodied in our indicator 1c, "Test Choice." Virginia and Nebraska alone fulfill this idea at present, permitting students or districts to choose from an array of end-of-course tests as well as from non-state tests like the Advanced Placement and International Baccalaureate exams. Next year's study will more heavily reward states that link exit and entrance tests, as well as those providing real test choice.

NCLB and the state systems it codifies are exclusively concerned with summative testing. These summative tests by nature of their timing, administration, and high-stakes consequences, necessarily stand apart from the ongoing work of the school year they seek to measure. This is one reason why teachers have generally been hostile to them: they simply cannot provide information that helps them do a better job with the kids they're working with right now and hence will never be what teachers most need, useful classroom tools. Even worse, drill-and-practice for these tests throughout the year tends to displace some of the ongoing informal assessment that teachers have always used to help them gauge their students' strengths and weaknesses, thereby depriving them of a useful classroom tool.

This displacement occurs because most teacher-driven formative assessment is not tightly linked to the summative standards and tests on which they and their students will be judged and so is not rewarded in a high-stakes world. This can be remedied, however, by state-level efforts to provide teachers with content and tools to help them link their classroom curricula to the high-stakes summative tests. These can range from simple guidelines for the creation of periodic low-stakes benchmark tests, to sophisticated electronic systems that provide ongoing formative assessment linked to both textbooks and state tests. *Testing the Testers 2004* will reward state efforts to turn summative assessment from a crude bludgeon into a useful classroom tool.

Related to this is the need to tie assessment to targeted professional development. There are many, many problems with the way educator professional development is handled in most states, not the least of which is its general failure to target specific, demonstrated areas of weakness in a particular teacher's practice. What most teachers need—and want—is not two days in an auditorium in August for “math” professional development but one hour, this week, with a trainer so she can help her kids do better multiplying fractions with two-digit denominators. This ideal is, in fact, a possibility given the sort of ongoing assessment described above. Even summative data, however, can be put to much better use in giving each teacher more targeted assistance provided that state law permits the tracking of student performance by teacher. This is something we will look more closely at next year as well.

Conclusion

Accountability matters because, in a high-stakes world, you get what you measure. Just as flawed accounting practices can reward the wrong sort of corporate behavior, flawed testing systems can reward the wrong sort of educational practices.

Supporters and opponents of test-based accountability implicitly agree on at least one thing: it matters because it changes what schools do. The re-organization of school governance, policy, and administration induced by accountability programs will be at least as profound as that which took place in the early decades of the twentieth century and which gave us the schools we know today. The debates over accountability have been so heated precisely because it is seen as such a long lever, and because the tracks and templates we lay down now will have significant effects—for better and for worse—for decades to come.

Most of the positions against accountability were fairly Manichean: they were not proposals for doing it better but in general protests against doing it at all. That battle having been lost (through a change in venue to Congress) those who might have made progressive contributions to accountability-as-a-given find themselves by and large irrelevant, left preaching to a choir without a congregation. This is unfortunate, since it is accountability-as-a-given that most needs engaged alternative perspectives to keep it from becoming the stultifying and brittle educational monoculture of its opponents' bad dreams.

At The Princeton Review we believe in accountability. We believe that it is difficult to improve something for which you have no metrics. We believe that accountability is crucial to restoring public confidence in public schools. We believe that there are fundamental issues of equity regarding who receives a good education and who does not that will persist indefinitely in the absence of hard data brought to light. And we believe that all the other efforts to improve educational outcomes—from increased funding, to improved teacher accreditation, to new forms of school governance, to the investigation of new pedagogical approaches—will be difficult to support in the long run without a sensible way of gauging their effectiveness.

That said, we also believe that bad high-stakes testing programs, precisely because of their power to reshape schools, are worse than no testing at all.

High-stakes accountability costs money whether it is done well or poorly. It is disruptive whether it is done well or poorly. It will be with us for a while whether it is done well or poorly. The incremental costs of doing accountability well (whether measured in dollars and time or in sweat and agita) are relatively small. Further, there are real costs to doing accountability poorly, including the millions of teacher and student hours and tens of millions of dollars wasted each year on tests that are redundant, disconnected from the curriculum, and yield no useful information to any teacher or parent struggling to help students to learn.

The point of *Testing the Testers* is to highlight the differences between programs done well and badly, to underscore that there is a choice, and to insist that those who design accountability systems be held accountable for their impact on the next several generations of students. We hope you'll re-join us next year to see what progress has been made.

Appendix I – Detailed Scores: 1-25

Note: Percentages are rounded up and so appear to total more than 100

State	Rank	1a. Granular Standards		1b. Overlap	1c. Test Choice	2a. Well-written and scored	2b. Multiple Item Types		2c. Pre-Validation and Review	2d. Curves and Established	2e. Curves Consistent	3a. Public Contract	3b. Public Test Specs	3c. Security and Due Process	3d. Timely Scoring	3e. Annual Test Release	3f. Disaggregated State Reporting	4a. Value Add	4b. Multiple Measures	4c. Stakes and Re-takes	4d. Useful Student data	4e. Useful school Data	4f. Support Programs	4g. Flexibility	4h. Data Warehouse	Weighted total	Unweighted Score (of 44)
	Weight	6%	8%	6%	4%	5%	2%	2%	3%	2%	5%	6%	6%	8%	5%	5%	5%	4%	5%	4%	5%	4%	4%				
NY	1	2	2	1	2	2	2	2	2	0	2	2	2	1	2	1	2	2	2	2	2	2	2	1	88.5	38	
MA	2	2	2	0	2	2	2	2	2	1	2	2	2	2	1	2	0	2	2	2	2	2	2	2	85.7	38	
TX	3	2	2	0	2	2	2	2	2	0	2	2	2	2	1	2	2	2	1	2	2	2	2	1	84.3	37	
NC	4	2	2	0	2	2	2	2	2	2	1	2	2	1	1	2	2	2	2	2	2	2	2	1	84.0	38	
VA	5	2	2	2	2	2	2	2	2	1	2	2	2	1	2	1	0	0	2	2	2	2	2	1	81.7	36	
LA	6	2	2	0	2	2	2	2	2	2	2	2	2	1	1	0	2	2	2	2	2	2	2	1	81.0	37	
FL	6	2	2	0	2	2	2	2	2	2	2	2	2	1	1	2	0	2	2	2	2	2	2	1	81.0	37	
AZ	8	2	2	0	2	2	2	2	2	1	2	2	2	1	1	0	2	2	2	2	2	2	2	1	80.2	36	
OK	8	2	2	0	2	2	2	2	2	1	2	2	2	1	1	0	2	2	2	2	2	2	2	1	80.2	36	
CA	10	2	2	1	2	2	2	2	2	1	2	2	2	1	1	2	0	2	2	2	2	2	0	1	79.7	35	
SC	11	2	2	0	2	2	2	0	2	1	2	2	2	0	1	2	2	2	2	2	2	2	2	1	79.5	35	
MS	12	2	2	1	2	2	2	2	2	2	2	2	2	1	1	0	0	2	2	2	2	2	2	1	78.7	36	
PA	13	2	2	1	2	2	2	2	2	2	2	2	1	1	1	0	2	2	2	2	2	2	2	1	78.0	36	
UT	14	2	2	0	0*	2	2	2	2	2	2	2	2	1	1	0	2	2	2	2	2	2	2	1	77.2	35	
MN	15	2	2	0	2	2	2	2	2	2	2	2	2	1	1	1	0	2	2	2	2	1	2	2	77.1	36	
CO	16	2	2	0	2	2	2	0	2	1	2	2	2	1	2	0	0	2	2	2	2	2	2	2	76.7	34	
TN	17	2	2	0	2	2	2	2	2	0	2	2	1	1	1	2	2	2	2	1	2	0*	1	76.5	33		
NV	17	2	2	0	2	2	2	0	2	2	1	2	2	1	1	0	2	2	2	2	2	2	2	1	76.5	34	
IL	19	2	2	0	2	2	2	2	2	2	2	2	2	1	1	1	0	2	0	2	2	2	2	2	76.2	35	
ME	20	2	2	0	2	2	2	2	2	2	1	2	1	2	1	0	2	0	2	2	2	2	2	1	76.0	34	
OR	20	2	2	0	2	2	2	2	2	0	2	2	2	1	1	0	2	0	2	2	2	2	2	1	76.0	33	
OH	22	2	1	0	2	2	2	2	2	2	1	2	2	2	1	0	2	2	1	2	2	2	2	1	75.8	35	
KY	23	2	2	0	2	2	2	2	2	0	2	2	1	1	1	0	2	0	2	2	2	2	2	2	74.7	33	
WA	24	2	2	0	2	2	2	2	2	2	2	2	1	1	1	0	2	0	2	2	2	2	2	1	74.5	34	
AK	25	2	2	0	2	2	2	2	2	0	2	2	2	1	1	0	0	2	2	2	2	2	2	1	74.2	33	

* No information was available

Appendix I – Detailed Scores: 26-50

Note: Percentages are rounded up and so appear to total more than 100

Note: Percentages are rounded up and so appear to total more than 100

State	Rank	1a. Granular Standards	1b. Overlap	1c. Test Choice	2a. Well-written and scored	2b. Multiple Item Types	2c. Pre-Validation and Review	2d. Curves and Established	2e. Curves Consistent	3a. Public Contract	3b. Public Test Specs	3c. Security and Due Process	3d. Timely Scoring	3e. Annual Test Release	3f. Disaggregated State Reporting	4a. Value Add	4b. Multiple Measures	4c. Stakes and Re-takes	4d. Useful Student data	4e. Useful school Data	4f. Support Programs	4g. Flexibility	4h. Data Warehouse	Weighted total	Unweighted Score (of 44)
	Weight	6%	8%	6%	4%	5%	2%	2%	3%	2%	5%	6%	6%	8%	5%	5%	5%	4%	5%	4%	5%	4%	4%		
MI	26	0	2	0	2	2	2	2	2	1	2	2	1	1	1	0	2	2	2	2	2	2	2	73.0	34
ID	27	2	2	0	2	2	2	2	2	1	0	2	1	2	2	1	2	0	2	2	0	2	1	72.6	32
NJ	28	2	2	0	2	2	2	0	2	1	2	2	1	1	1	0	2	2	1	2	2	2	1	72.3	32
AR	29	2	0*	0	2	2	2	2	2	1	1	2	2	2	1	0	2	0	2	2	2	2	2	72.0	33
CT	30	2	2	0	2	2	2	2	2	1	2	2	1	1	1	1	0	0	2	2	2	2	1	71.1	32
NE	31	2	2	2	2	2	2	2	2	0	1	2	0*	0	1	0	2	0	2	2	2	2	1	70.0	31
VT	32	0	2	0	2	2	2	2	2	1	2	1	1	1	2	2	2	0	1	2	2	2	1	69.6	32
AL	33	0	2	0	2	2	2	2	2	0	2	2	1	1	2	0	2	2	2	0	2	1	1	67.5	30
MO	34	0	2	0	2	2	2	2	2	2	0	2	2	1	1	0	2	0	2	1	2	2	2	67.0	31
MD	35	2	0*	0	2	2	2	2	2	0	0	2	1	1	2	0	2	2	2	2	2	2	1	66.2	31
DE	36	0	2	0	1	2	2	0	2	1	0	2	2	1	1	2	0	2	2	2	2	2	1	65.6	29
NM	37	2	0*	0	2	2	2	2	2	0	1	2	2	0	1	0	2	2	2	2	2	2	1	65.5	31
NH	38	0	2	0	2	2	2	2	2	2	0	1	1	2	1	0	2	0	2	2	2	2	1	64.7	30
DC	39	2	1	0	2	2	2	2	2	0*	1	2	2	0	1	0	0	1	2	2	2	2	1	62.5	29
GA	40	2	2	0	2	2	2	2	2	0	0	2	1	2	1	0	0	2	2	0	0	2	1	61.7	27
KS	41	2	0*	0	2	2	2	2	2	1	1	2	1	1	1	0	2	0	1	2	1	2	1	58.2	28
IN	42	2	0*	0	2	2	2	2	2	1	0	2	1	1	1	0	0	2	1	2	2	2	1	56.8	28
HI	43	0	2	0	2	2	2	0	2	2	1	1	1	0	1	0	2	0	2	1	2	2	1	55.5	26
WY	44	0	0*	0	2	2	2	2	2	0*	0	1	2	1	1	0	2	0	2	2	2	2	1	54.5	26
ND	45	0	2	0	2	2	2	0	2	1	1	2	0	0	2	0	2	0	2	2	1	2	0	54.3	25
WI	46	0	2	0	2	2	2	n/a	n/a	2	0	2	0	1	1	0	0	0	2	2	2	2	1	53.2	23
WV	47	2	0*	0	2	2	0	2	2	0	0	2	0	0	1	0	2	0	2	2	2	2	1	52.2	24
SD	48	2	2	0	2	2	2	n/a	n/a	n/a	0	0	1	0	1	0	2	0	0*	1	2	2	1	49.8	20
RI	49	0	2	0	2	2	2	2	n/a	0	0	1	0	0	1	1	2	0	0	2	2	2	1	48.5	22
IA		0	2	1	0*	2	0*	2	2	0	0*	0*	0	0	1	0	2	0	0	2	2	2	1	42.2	19
MT	50	0	0*	0	2	0	1	2	2	0	0	1	0	0	1	0	0	0	2	2	0	2	1	29.0	16

* No information was available

Appendix II

INDICATORS, RUBRICS, AND EXPLANATIONS FOR TESTING THE TESTERS 2003

Criterion #1: Tests are aligned to academic content knowledge and skills as specified by the states' curriculum standards (20% of Rank)

1a) Are standards granular enough that a small number of test items can reasonably measure a student's mastery of that granule? (6%)

2 points if there are specific statements that can be measured with 3 or fewer items (e.g., "Add whole numbers"). 0 points if standards are broader than that.

Standards that can't be measured with a small number of items are either too broad or too vague. Because reading and writing skills are necessary in all content areas, we gave added weight to English standards over other content areas.

1b) Is there substantial overlap between the published curriculum standards and those that are actually tested? (8%)

2 points if two-thirds or more of the published standards are tested, even if they are not all tested in one year. 1 point if one-third to two-thirds of the published standards are tested. 0 points if fewer than one-third of the published standards are tested, or if this information is not available.

A poor match between curriculum and tested standards creates needless uncertainty for teachers and students, and can reward schools for focusing on gamesmanship and test-preparation at the expense of actual teaching and learning.

1c) Do states allow schools or students to choose from among a number of tests? (6%)

2 points if the state uses end-of-course tests and permits schools or students to choose from among them and from tests offered by other states or organizations. 1 point if students or schools may choose from a number of tests, either state or local, to meet state requirements, but all are created by the state and may or may not be end-of-course. 0 points if the state mandates that all students must take the same set of tests.

Accountability frameworks that require all students to take the same test and, effectively, the same courses tip the balance too far against individual and school control over education, especially in high school. End-of-course tests provide reasonable curricular flexibility while still supporting accountability. There is no reason why students should be limited to tests produced by their own state departments of education: appropriate tests from other states or from independent organizations could be accepted as well.

Criterion #2: The tests are capable of determining that those standards have been met (20% of Rank)

2a) Are the items well written and the tests scored accurately and completely? (4%)

2 points if the items are well written and the tests are scored accurately and completely. 1 point if the items are well written, but the tests are not scored accurately. 0 points if the items are not well written, the tests are not scored accurately and completely, or this information is not available.

Accountability programs will not foster confidence unless the basics are done well, but not all states meet these baseline expectations.

2b) Do the tests include multiple item types (e.g., multiple choice, open ended, computation, performance activities, etc.)? (5%)

2 points if the tests include multiple item types, or a single item type other than multiple-choice. 0 points if the tests include only one item type, or if this information is not available.

Multiple-choice questions, or variations such as a true/false format, are cheaper to score and report, but tests that use them exclusively may not be able to capture the full quality of a student's reasoning. They also encourage the kind of superficial drill-and-kill test preparation that is rightly derided.

2c) Are items validated before the test is assembled, and does someone other than the test developer review items before the test is constructed? (2%)

2 points if items are validated, and are reviewed by someone other than the test developer, before test construction. 1 point if one or the other, but not both, are done before the test is constructed. 0 points if items are neither validated nor reviewed by an outside party, or if this information is not available.

These are basic psychometric best-practices and quality-control measures.

2d) Were the achievement cutoff points (and scoring curve, if applicable) established before the first live administration of the test? (2%)

2 points if the cutoff points (and scoring curve, if applicable) were established before the first live administration of the test. 0 points if the cutoff points (and scoring curve, if applicable) were not pre-established. The score is reported as "not applicable" if the tests are currently being field-tested or if there is not sufficient data to establish cutoff points. [Note: Slight adjustment of cutoff points after the first live administration of the test is acceptable.]

Accountability policies have a strong political component, for example the desire of politicians to appear "tough on schools" or the fear of alienating large blocks of voters by denying diplomas to many of their children. Pass/fail cutoffs and curves (for tests that have them) should be set beforehand to the extent that is psychometrically valid. Setting them after tests are scored opens the process to inappropriate manipulation.

2e) Are the cutoff points (and scoring curves, if applicable) for various grades and subjects consistent enough across subjects and years to enable comparison?(3%)

2 points if the cutoff points (and scoring curves, if applicable) are consistent enough to enable comparison. 0 points if they are not consistent enough to enable comparison, or this information is not available. "Not applicable" if the tests are currently being field-tested or if there is not sufficient data to establish cutoff points.

A grade of B+ or a score of 72, for example, should represent roughly the same level of achievement in science as in math, or for a sixth-grader as for a ninth-grader.

Criterion #3: The policies and procedures surrounding the tests are open, and open to ongoing improvement (30% of Rank)

3a) Are contract terms with the testing companies readily available for public inspection? (2%)

2 points if full contracts are available online or through a telephone, E-mail, or fax request. 1 point if testing company name, duration of contract, and dollar amount are available online or through a telephone, E-mail, or fax request. 0 points if contracts and contract terms are not available or must be requested via written letter.

In recent years a series of significant administrative and scoring failures by testing companies and state agencies affected tens of thousands of students across the country. Families have the right to know that the testing companies are also accountable.

3b) Are detailed test specifications readily available? (5%)

2 points if the specifications are readily available, for example, on the state's Web site. 1 point if detailed test specifications are not readily available and must be requested. 0 points if detailed test specifications do not exist, are not available whatsoever to the public, or the state claims that test specifications for commercially-available tests are proprietary information.

Detailed, accessible test specifications are the starting point for any independent review of an accountability program: What is the test trying to measure, and how?

3c) Is there a reasonable level of security surrounding the test and scoring procedures, and are there due process guidelines for students and educators accused of cheating? (6%)

2 points if the test and scoring procedures are reasonably secure, and there are clear and implemented guidelines to deal with students and educators accused of cheating. 1 point if the test and scoring procedures are reasonably secure, but there are no published guidelines to deal with students and educators accused of cheating. 0 points if the test and scoring procedures are not reasonably secure, and there are no published security guidelines, or if this information is not available.

Security and due process go hand-in-hand. In a high-stakes world, weak security (or inherently insecure systems like California's or Colorado's use of the same test form year after year) encourages cheating by educators or, less often, by students. As stakes rise, states must make it more difficult to cheat, while building reasonable procedures for those accused of cheating to defend themselves.

3d) Are complete test scores released to the public in a timely manner? (6%)

Tests that contain only multiple-choice items: 2 points if all scores and subscores are released to the public within one month of the test date. 1 point if scores are released to the public within three months of the test date or if partial scores or subscores are released within one month. 0 points if neither is the case.

Tests that contain open-ended items: 2 points if all scores and subscores are released to the public within three months of the test date. 1 point if scores are released to the public within five months of the test date or if partial scores or subscores are released within three months. 0 points if neither is the case.

Parents and other stakeholders are supposed to be the ultimate consumers of accountability data. Unnecessary delay or failure to provide it will ultimately undermine public confidence in any accountability system, no matter how well-designed. However, it takes extra time to hand-score essay questions, and states should not be penalized for that.

3e) Is the test released every year? (8%)

2 points if the entire test is released every year. 1 point if the entire test is released every few years, or if only sample items from the test are released every year. 0 points if neither of these is the case.

Releasing tests each year has numerous benefits. It improves test security, since new tests are necessarily generated every year. It improves parent and educator understanding of how curriculum standards are actually tested. It encourages public scrutiny of the tests themselves, leading to more robust instruments. It is incrementally more expensive than non-disclosure, but this is a small part of the overall cost of the testing program that adds tremendous value.

3f) Does the state publish detailed information on the disparate performance of different groups? (5%)

2 points if the state releases disaggregated item-response data. 1 point if the state releases disaggregated statistics for the overall test. 0 points if the state does not release disaggregated information.

When data are released in aggregate for all students, it's impossible to compare schools fairly by looking at the relative performance of students with similar characteristics. Disaggregated reporting on the overall test is now required by NCLB.

Criterion #4: Accountability systems affect education in a way that is consistent with the goals of the state (30% of Rank)

4a) Does the state track and judge schools by value-added analysis? (5%)

2 points if the state tracks and judges schools by value-added analysis. 1 point if the state tracks but does not judge schools by value-added analysis. 0 points if the state does not track or judge schools by value-added analysis or if this information is not available.

Value-added analysis measures the effect that teachers and/or schools have on increasing student performance from year to year. Comparisons based on absolute scores rather than the level of relative improvement over-reward teachers or schools whose students would have done well regardless of what went on in the classroom.

4b) Does the state judge schools by multiple indicators, such as graduation, promotion, attendance, and violence rates? (5%)

2 points if the state judges schools based on multiple indicators. 0 points if the state judges schools by test scores only or if this information is not available.

It's not hard to raise test scores if raising test scores is all you want to do. Ignoring all other aspects of school quality risks making schools into grim places for both teachers and students and rewards the driving of marginal students out of school.

4c) Do tests have stakes for students, and do students have opportunities to retake the test if necessary? (4%)

2 points if the state requires a student to pass a test to be promoted or to receive a diploma, and students are given multiple opportunities to retake unpassed portions of a test. 1 point if the state requires a student who does not pass to attend summer school, but does not require a student to pass a test to receive a diploma. 0 points if stakes are not involved or if a student is not given any opportunities to retake a test with associated stakes.

If there are no stakes for students then there is little incentive for them to take these tests seriously, significantly undermining the reliability of the results. At the same time, it is unfair to make important decisions about students' lives based on a single poor performance.

4d) Is test data regarding individual students distributed to educators and families in useful detail? (5%)

2 points if educators and families receive test score reports from the state that are useful in identifying areas of excellence and weakness for each student. 1 point if only educators receive such score reports from the state, regardless of whether districts are responsible for forwarding the reports to families. 0 points if scores are reported without any useful detail or if this information is not available.

It is crucial that educators and families receive test data in a way that enables them to identify areas of improvement so that students can be helped. Otherwise, testing is largely pointless.

4e) Is school-level performance data shared with the public along with explanations and contextual detail appropriate for a general audience? (4%)

2 points if performance data includes contextual detail that lucidly conveys how students performed and identifies trends and plans for improvement. 1 point if performance data includes only limited contextual detail. 0 points if performance data are not shared with the public, or if no contextual detail is provided.

Citizens are the ultimate audience for performance data. It should be presented in such a way that interested parents and community members can make sense and use of it.

4f) Are support programs in place to assist students and schools in overcoming their deficiencies? Are there consequences based on the performance of these programs? (5%)

2 points if there are both support programs and consequences. 1 point if there is a support program, but no consequences. 0 points if there are no support programs, or if this information is not available.

It's not enough to tell someone they're having trouble, you have to help them as well. And those efforts at improvement must themselves have consequences for success or failure if they're to be taken seriously by educators.

4g) Does the state give districts and schools the latitude to meet performance standards in reasonably flexible ways? (4%)

2 points if there has been a significant effort to relax input regulations as accountability is implemented. 1 point if there has been some effort to relax regulations. 0 points if the state has not relaxed its regulations, or if this information is not available.

It's unfair to penalize schools for doing a poor job while mandating at each step of the way

how they must approach the problem. Measuring the outputs of education should permit more freedom from state mandates on the inputs.

4h) Does the state maintain publicly available data warehouses to evaluate educational progress over time? (4%)

2 points if the state has a multi-year data warehouse that includes item-response data and student performance data broken out demographically. 1 point if the state collects and makes available multi-year student performance data, but not item-response data. 0 points if the state makes no data available or if data are available for one year only.

A tremendous amount of time, effort, and money is being devoted to accountability systems and a tremendous amount of data is being generated as a result. The answers to many difficult and long-standing questions about what works for which kids reside here, and states should collect and save the information in ways that encourage the greatest possible number of public-policy analyses. The ability to look at performance data over time, down to the item level, and correlate it with teacher and school characteristics, academic scope-and-sequence and numerous other factors would be immensely valuable, and can be accomplished while protecting the privacy of the individuals involved.

Appendix III — Changes from 2002

<u>2002 Indicator</u>	<u>2003 Indicator</u>	<u>Changes</u>
Total points possible: 50	Total points possible: 44	Some indicators were dropped or consolidated this year
<p><i>1a. Are standards granular enough that a small number of test items can reasonably measure a student's mastery of that granule?</i></p> <p>2 points if there are specific statements that can be measured with single items (e.g., "Add two-digit numbers"). 1 point if there is a combination of broad and specific statements that can be measured with 2–5 items per statement. 0 points if standards are broader than that.</p>	<p><i>1a. Are standards granular enough that a small number of test items can reasonably measure a student's mastery of that granule?</i></p> <p>2 points if there are specific statements that can be measured with 3 or fewer items (e.g., "Add whole numbers"). 0 points if standards are broader than that.</p>	From 2-1-0 to 2-0. All of 2002's states that had been in the 1-pt range had to be re-examined and placed in 2 or 0.
<p><i>1b. Is there substantial overlap between the published curriculum standards and those that are actually tested?</i></p> <p>2 points if two-thirds or more of the published standards are tested, even if they are not all tested in one year. 1 point if one-third to two-thirds of the published standards are tested. 0 points if fewer than one-third of the published standards are tested, or if this information is not available.</p>	<p><i>1b. Is there substantial overlap between the published curriculum standards and those that are actually tested?</i></p> <p>2 points if two-thirds or more of the published standards are tested, even if they are not all tested in one year. 1 point if one-third to two-thirds of the published standards are tested. 0 points if fewer than one-third of the published standards are tested, or if this information is not available.</p>	Weighting was decreased from 10% to 8%
<p><i>1c. Do states allow schools to choose from among equated tests and standards?</i></p> <p>2 points if the state uses end-of-course tests and permits schools or students to choose from several rigorous curricula, each tied tightly to its own test, with the tests equated to one another for comparability. 1 point if students or schools may choose from a number of tests,</p>	<p><i>1c. Do states allow schools or students to choose from among a number of tests?</i></p> <p>2 points if the state uses end-of-course tests and permits schools or students to choose from among them and from tests offered by other states or organizations. 1 point if students or schools may choose from a number of tests, either state or local, to meet state</p>	<p>Emphasis shifted from equating to flexibility</p> <p>Weighting was increased from 4% to 6%</p>

2002 Indicator

either state or local, to meet state requirements but the tests are not equated and may or may not be end-of-course. 0 points if states mandate that all students must take the same set of tests.

2a. Are the items well written and the tests scored accurately and completely?

2 points if the items are well written and the tests are scored accurately and completely. 1 point if the items are well written but the tests are not scored accurately. 0 points if the items are not well written, the tests are not scored accurately and completely, or if this information is not available.

2b. Do the tests include multiple item types (e.g., multiple choice, open ended, computation, performance activities, etc.)?

2 points if the tests include three or more items types. 1 point if the tests include two or more item types. 0 points if the tests include only one item type, or if this information is not available.

2c. Does someone other than the test developer review items before test construction?

2 points for yes. 0 points for no, or if this information is not available.

2003 Indicator

requirements, but all are created by the state and may or may not be end-of-course. 0 points if the state mandates that all students must take the same set of tests.

2a. Are the items well written and the tests scored accurately and completely?

2 points if the items are well written and the tests are scored accurately and completely. 1 point if the items are well written, but the tests are not scored accurately. 0 points if the items are not well written, the tests are not scored accurately and completely, or this information is not available.

2b. Do the tests include multiple item types (e.g., multiple choice, open ended, computation, performance activities, etc.)?

2 points if the tests include multiple item types, or a single item type other than multiple-choice. 0 points if the tests include only one item type, or if this information is not available.

2c. Are items validated before the test is assembled, and does someone other than the test developer review items before the test is constructed?

2 points if items are validated, and are reviewed by someone other than the test developer, before test construction. 1 point if one or the other, but not both, are done before the test is constructed. 0 points if items are neither validated nor reviewed by an outside party, or if this information is not available.

Changes

Weighting was decreased from 5% to 3.8%

Rubric change

Weighting decreased from 5% to 4.5%

Two of 2002's criteria (2c and 2d) were combined into 2003's 2c.

Combined weighting decreased from 4% to 1.5%

2002 Indicator	2003 Indicator	Changes
<p><i>2d. Are the items validated before the test is assembled? 2 points if items are reviewed, field tested, and then evaluated. 1 point if items are reviewed but not field tested. 0 points if items are not reviewed or field tested, or if this information is not available.</i></p>		<p>Two of 2002's criteria (2c and 2d) were combined into 2003's 2c.</p>
<p><i>2e. Are the scoring curve and achievement cutoff points established before the test is given? 2 points if the curve and cutoff points are established before the test is given. 1 point if one or the other, but not both, are established before the test is given. 0 points if neither is established before the test is given, or if the information is not available.</i></p>	<p><i>2d. Were the achievement cutoff points (and scoring curve, if applicable) established before the first live administration of the test? 2 points if the cutoff points (and scoring curve, if applicable) were established before the first live administration of the test. 0 points if the cutoff points (and scoring curve, if applicable) were not pre-established. The score is reported as "not applicable" if the tests are currently being field-tested or if there is not sufficient data to establish cutoff points. [Note: Slight adjustment of cutoff points after the first live administration of the test is acceptable.]</i></p>	<p>2003 is a wording change from 2002 in order to clarify. The point of the question is the same. Weighting increased from 2% to 2.3%</p>
<p><i>2f. Are the scoring curves and cutoff points consistent from subject to subject and from year to year? 2 points if the curves and cutoff points are consistent. 1 point if either one is consistent across years and subjects. 0 points if they are not consistent, or if this information is not available.</i></p>	<p><i>2e. Are the cutoff points (and scoring curves, if applicable) for various grades and subjects consistent enough across subjects and years to enable comparison? 2 points if the cutoff points (and scoring curves, if applicable) are consistent enough to enable comparison. 0 points if they are not consistent enough to enable comparison, or this information is not available. "Not applicable" if the tests are currently being field-tested or if there is not sufficient data to establish cutoff points.</i></p>	<p>2003 is a wording change from 2002 in order to clarify. The point of the question is the same. Weighting decreased from 4% to 3%</p>

<u>2002 Indicator</u>	<u>2003 Indicator</u>	<u>Changes</u>
<p><i>3a. Are contract terms with the testing companies readily available for public inspection?</i></p> <p>2 points if contracts are readily available. 1 point if testing company name and duration of contract are readily available. 0 points if contract terms are not readily available, even if the testing company name is provided, or if the name of the testing company is not readily available.</p>	<p><i>3a. Are contract terms with the testing companies readily available for public inspection?</i></p> <p>2 points if full contracts are available online or through a telephone, E-mail, or fax request. 1 point if testing company name, duration of contract, and dollar amount are available online or through a telephone, E-mail, or fax request. 0 points if contracts and contract terms are not available or must be requested via written letter.</p>	<p>In 2003, the wording was changed to clarify.</p> <p>A significant change was made in giving 0 pts to states requiring a written request for contracts/terms.</p> <p>Weighting decreased from 3% to 1.5%</p>
<p><i>3b. Do detailed test specifications exist and are they easily available?</i></p> <p>2 points if the specifications are readily available, for example, on the state's website. 1 point if the specifications exist but must be requested. 0 points if no specifications exist, if no specifications is available to the public, or if the state claims that test specifications exist for commercially available tests.</p>	<p><i>3b. Are detailed test specifications readily available?</i></p> <p>2 points if the specifications are readily available, for example, on the state's Web site. 1 point if detailed test specifications are not readily available and must be requested. 0 points if detailed test specifications do not exist, are not available whatsoever to the public, or the state claims that test specifications for commercially available tests are proprietary information.</p>	<p>Weighting increased from 3% to 4.5%</p>
<p><i>3c. Are all students tested, and are all scores included in the statistical profile of a school?</i></p> <p>2 points if all students are tested, even with an Alternate Assessment, and all scores are included in the profile of a school. 1 point if guidelines exist for determining which students should be tested, and all scores are to be included in the profile of a school, but aren't. 0</p>		<p>This question was removed from 2003</p>

<i>2002 Indicator</i>	<i>2003 Indicator</i>	<i>Changes</i>
<p>points if not all students must be tested, or if certain scores may be excluded from the statistical profile of a school.</p>		
<p><i>3d. Is there a reasonable level of security around the test and scoring procedures, and due process guidelines for people accused of cheating?</i> 2 points if there are clear and implemented guidelines to deal with students, teachers, and administrators accused of cheating, and if the test and scoring procedures are reasonably secure. 1 point if the test and scoring procedures are reasonably secure, but no guidelines are published to deal with people accused of cheating. 0 points if there are no published security guidelines, or if this information is not available.</p>	<p><i>3c. Is there a reasonable level of security around the test and scoring procedures, and are there due process guidelines for students and educators accused of cheating?</i> 2 points if the test and scoring procedures are reasonably secure, and there are clear and implemented guidelines to deal with students and educators accused of cheating. 1 point if the test and scoring procedures are reasonably secure, but there are no published guidelines to deal with students and educators accused of cheating. 0 points if the test and scoring procedures are not reasonably secure, and there are no published security guidelines, or if this information is not available.</p>	<p>No change</p>
<p><i>3e. Are complete test scores released to the public in a timely manner?</i> 2 points if all scores and subscores are released to the public within one month of the test date. 1 point if scores are released to the public within three months of the test date or if partial scores or subscores are released within one month. 0 points if neither is the case.</p>	<p><i>3d. Are complete test scores released to the public in a timely manner?</i> Tests that contain only multiple-choice items: 2 points if all scores and subscores are released to the public within one month of the test date. 1 point if scores are released to the public within three months of the test date or if partial scores or subscores are released within one month. 0 points if neither is the case.</p> <p>Tests that contain open-ended items: 2 points if all scores and subscores are released to the public within three months of</p>	<p>Scoring was changed in 2003 to allow more time for tests with open-ended answers.</p> <p>Weighting increased from 4.5% to 6%</p>

<u>2002 Indicator</u>	<u>2003 Indicator</u>	<u>Changes</u>
	the test date. 1 point if scores are released to the public within five months of the test date or if partial scores or subscores are released within three months. 0 points if neither is the case.	
<p><i>3f. Is the test released every year?</i> 2 points if the entire test is released every year. 1 point if the entire test is released every few years, or if sample items from the test are released every year. 0 points if neither of these is the case.</p>	<p><i>3e. Is the test released every year?</i> 2 points if the entire test is released every year. 1 point if the entire test is released every few years, or if only sample items from the test are released every year. 0 points if neither of these is the case.</p>	Weighting increased from 6% to 7.5%
<p><i>3g. Does the state publish detailed information on the disparate performance of different groups?</i> 2 points if the state releases disaggregated response data. 1 point if the state releases disaggregated statistics for the test overall. 0 points if the state does not release disaggregated information.</p>	<p><i>3f. Does the state publish detailed information on the disparate performance of different groups?</i> 2 points if the state releases disaggregated item-response data. 1 point if the state releases disaggregated statistics for the overall test. 0 points if the state does not release disaggregated information.</p>	Weighting increased from 3% to 4.5%
<p><i>4a. Does the state report indicators of school quality other than test scores, such as dropout, teacher experience, or crime rates?</i> 2 points if states report multiple criteria and employ value-added analysis. 1 point if the state uses multiple indicators or value-add. 0 points if the states report only test scores without value-add, or if this information is not available.</p>		This question was dropped from 2003.

<i>2002 Indicator</i>	<i>2003 Indicator</i>	<i>Changes</i>
<p><i>4b. Does the state judge schools by multiple measures, as above?</i> 2 points if the state judges schools based on multiple criteria and employ value-added analysis. 1 point if the state judges schools based on multiple indicators. 0 points if the states judge schools by test scores only, or if this information is not available.</p>	<p><i>4b. Does the state judge schools by multiple indicators, such as graduation, promotion, attendance, and violence rates?</i> 2 points if the state judges schools based on multiple indicators. 0 points if the state judges schools by test scores only or if this information is not available.</p>	<p>Because value-add was separated into its own question in 2003, it's easier for a state to receive 2 pts for this question in 2003.</p>
<p><i>4c. Do tests have stakes for students, and do students have opportunities to re-take the test if necessary?</i> 2 points if there are stakes for students and they are given multiple opportunities to retake the test. 1 point if there are stakes for students and they are given only one opportunity to retake the test. 0 points if they are unable to retake the test, or if stakes are not involved.</p>	<p><i>4a. Does the state track and judge schools by value-added analysis?</i> 2 points if the state tracks and judges schools by value-added analysis. 1 point if the state tracks but does not judge schools by value-added analysis. 0 points if the state does not track or judge schools by value-added analysis or if this information is not available.</p> <p><i>4c. Do tests have stakes for students, and do students have opportunities to re-take the test if necessary?</i> 2 points if the state requires a student to pass a test to be promoted or to receive a diploma, and students are given multiple opportunities to retake unpassed portions of a test. 1 point if the state requires a student who does not pass to attend summer school, but does not require a student to pass a test to receive a diploma. 0 points if stakes are not involved or if a student is not given any opportunities to retake a test with associated stakes.</p>	<p>Although 2003 4a resembles 2002 4b, the focus solely on value-add makes this a separate question.</p> <p>Rubric change</p> <p>Weighting increased from 3% to 3.5%</p>

2002 Indicator	2003 Indicator	Changes
<p><i>4d. Is test data distributed to educators in useful detail that can be linked to other databases?</i> 2 points if school administrators receive data by student and by teacher in such a way that it can tie into their school data systems electronically. 1 point if school administrators receive data by student and by teacher, but no structure exists to track this information electronically. 0 points if school administrators do not receive data by both student and teacher, or if this information is not available.</p>	<p><i>4d. Is test data regarding individual students distributed to educators and families in useful detail?</i> 2 points if educators and families receive test score reports from the state that are useful in identifying areas of excellence and weakness for each student. 1 point if only educators receive such score reports from the state, regardless of whether districts are responsible for forwarding the reports to families. 0 points if scores are reported without any useful detail or if this information is not available.</p>	<p>Wording and rubric change</p> <p>Weighting increased from 3% to 5.3%</p>
<p><i>4e. Are support programs in place to assist failing teachers, students, and schools in overcoming their deficiencies? Are there consequences based on the performance of these programs?</i> 2 points if there are both support programs and consequences. 1 point if there is either a support program or consequences but not both. 0 points if there is neither, or if this information is not available.</p>	<p><i>4f. Are support programs in place to assist students and schools in overcoming their deficiencies? Are there consequences based on the performance of these programs?</i> 2 points if there are both support programs and consequences. 1 point if there is a support program, but no consequences. 0 points if there are no support programs, or if this information is not available.</p>	<p>Weighting increased from 3% to 5.3%</p>
<p><i>4f. Does the state give districts and schools the latitude to meet performance standards in reasonably flexible ways?</i> 2 points if there has been a significant effort to relax regulations as accountability is implemented. 1 points if there has been some effort to relax regulations. 0 points if the state has not relaxed its regulations, or if this information is not available.</p>	<p><i>4g. Does the state give districts and schools the latitude to meet performance standards in reasonably flexible ways?</i> 2 points if there has been a significant effort to relax regulations as accountability is implemented. 1 points if there has been some effort to relax regulations. 0 points if the state has not relaxed its regulations, or if this information is not available.</p>	<p>Weighting decreased from 4.5% to 3.5%</p>

2002 Indicator	2003 Indicator	Changes
<p><i>4g. Is performance data shared with the public along with explanations and contextual detail appropriate for a general audience?</i> 2 points if information is shared with the public with contextual detail allowing readers to get a clear sense of how students performed. 1 point if limited contextual detail is provided. 0 points if contextual detail is not provided.</p>	<p><i>4e. Is school-level performance data shared with the public along with explanations and contextual detail appropriate for a general audience?</i> 2 points if performance data includes contextual detail that lucidly conveys how students performed and identifies trends and plans for improvement. 1 point if performance data includes only limited contextual detail. 0 points if performance data are not shared with the public, or if no contextual detail is provided.</p>	<p>Wording change. Weighting increased from 3% to 3.5%</p>
<p><i>4h. Does the state maintain publicly-available data-warehouses to evaluate educational progress over time, with privacy protected?</i> 2 points if the state has a multi-year data warehouse, including item-response data and student-performance data broken out demographically. 1 point if the state collects and makes available multi-year student performance but not item-response data. 0 points if the state makes no data available or if it is available for only one year.</p>	<p><i>4h. Does the state maintain publicly available data warehouses to evaluate educational progress over time?</i> 2 points if the state has a multi-year data warehouse that includes item-response data and student performance data broken out demographically. 1 point if the state collects and makes available multi-year student performance data, but not item-response data. 0 points if the state makes no data available or if data are available for one year only.</p>	<p>Weighting increased from 3% to 3.5%</p>

Appendix IV

The Question of Proficiency

In this study we have purposely avoided the question of whether each state's standards are appropriate, let alone rigorous. We have tried to avoid prescribing content in favor of prescribing systems by which any given set of content can be taught, learned, and assessed. There are many organizations for whom the content of standards is a daily concern and who feel that it is pointless to judge judgment on accountability systems if you are not also judging the value of what they hold students and schools accountable for. To them, good policies supporting the learning of weak standards are irrelevant at best. Perhaps. But that is a bit like saying there is no point in owning a good car until you are certain you will only be driving it to the best possible destination. We believe that notions of what students ought to learn will inevitably shift (sometimes quite rapidly) but that bad systems are remarkably enduring. Fortunately, there is no reason why work on one must wait upon the other.

There is another aspect of rigor, however, somewhat less remarked but which ultimately plays a greater role in determining the number of failing students and schools. It is the one that is also increasingly in play as states struggle with the Adequate Yearly Progress (AYP) provisions of NCLB. We are speaking, of course, of the variability of the proficiency bar from state to state. Regardless of what exactly a state expects its fourth-graders to know it can decide that proficiency consists of mastering, say, fifty percent of that content instead of, say, eighty percent.

Back in the olden days before NCLB, governors often found it expedient to “get tough on schools”, crime having been gotten tough on during the previous campaign. The simple way to demonstrate this toughness was to have your appointees at the Board or Department of Education set proficiency levels rather high, such that the majority of schools would be labeled as failing. You could then announce your new policy initiatives designed to shake up the complacent education bureaucracy and return to the lost rigors of the Three ‘Rs. Indeed, many governors found this an attractive path. ¹

When NCLB passed, however, both the bully pulpit and the big stick shifted from the state house to the White House. The Federal Department of Education decided, shockingly, to hold states to the letter of their proficiency levels, such that 100% of students in each state would become so proficient within twelve years, with significant milestones and onerous consequences along the way for schools and states that lagged behind. Now, it would not be the state ranking school districts, but the Department of Education ranking states. Those high proficiency bars (either admirably, unrealistically, or self-servingly high, depending on your viewpoint) were now a liability rather than an asset, and a financial liability to boot since NCLB ties much of Title I funding to a state's successful achievement of AYP.

Unsurprisingly, there has been a rush by states to re-define “proficient.” Unsurprisingly, the re-definition is almost always downward so that, suddenly, schools don't seem quite as underperforming as they once did. Michigan, for example, went from 1,513 failing schools last year to 216 this year not by having more successful students but by redefining a proficient school from one where 75% of students are proficient to one where 38% of students are. Well, that was easy.

We are not saying that high proficiency bars are a priori better than low ones. For example, setting the bar very high such that large numbers of schools are labeled as failing is likely to lead to a dispiriting emphasis on simply raising test scores, which can make schools into grim places that do little to instill the love of teaching and learning that is ultimately required for real academic success. Further, NCLB has fairly significant penalties for schools that fail to meet proficiency targets, penalties that can make life even more difficult for students and staff at those schools.

Nevertheless, we thought it would be interesting to have an overview of where states are setting the proficiency bar. This is a bit more complicated than simply comparing cut-off points since one state may require a high percentage of mastery of quite undemanding content while another requires a low success rate against a much more demanding set of standards. What is needed, then, is a common standard against which to gauge proficiency. As is so often the case, we turned to the National Assessment of Educational Progress (NAEP), the one indicator widely, if sometimes grudgingly, accepted as a common standard of academic achievement.

Since what we sought was a good-enough measure, a rough indicator, we decided to compare the states' claims of proficiency relative to one another using NAEP proficiency levels as a sort of go-between. We used scores from the 2000 Grade 8 Math NAEP test and the state's own proficiency levels on their most recent statewide 8th grade math test, generally from 2002. In some cases the states used slightly different data on which to base their proficiency claims; we've noted those discrepancies in the chart below. South Dakota lacked appropriate data for comparison and is not included here.

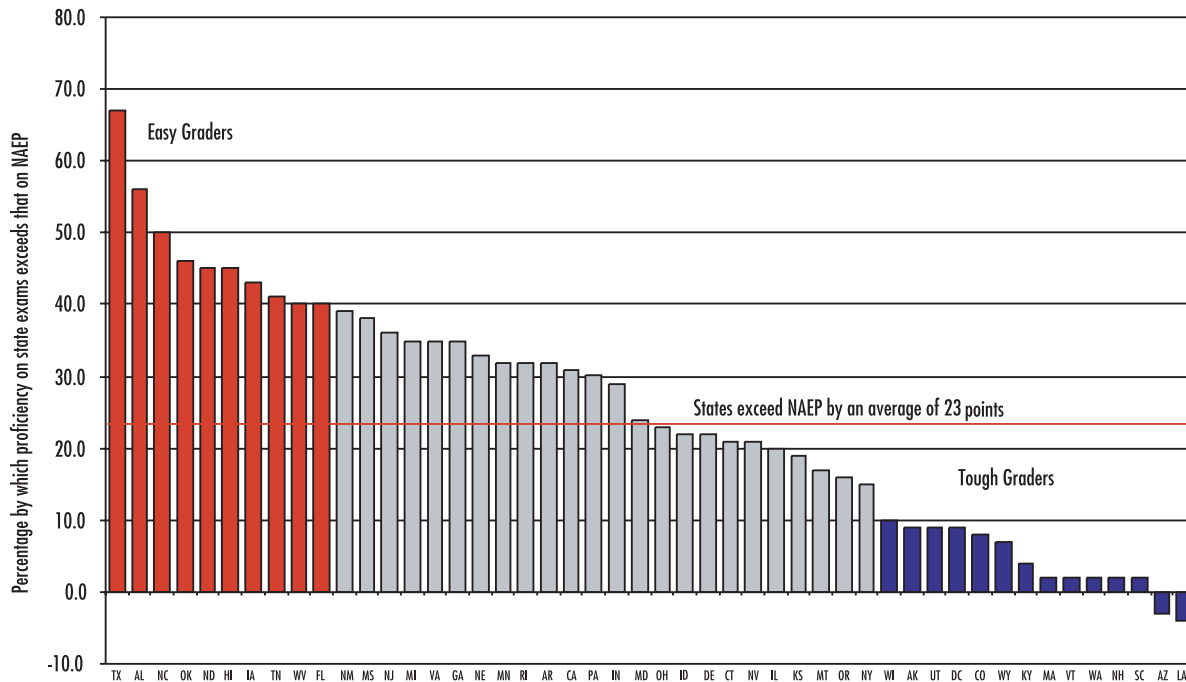
STATE	NAEP Proficiency	State Proficiency	DIFFERENCE	NOTES
TX	24	91	-67	
AL	16	72	-56	State proficiency is for Grade 6
NC	30	80	-50	
OK	19	65	-46	
HI	16	61	-45	
ND	31	76	-45	
IA	31	74	-43	Biennial average
TN	17	58	-41	
FL	17	57	-40	
WV	18	58	-40	
NM	13	52	-39	
MS	8	46	-38	
NJ	24	60	-36	
GA	19	54	-35	
MI	28	63	-35	State proficiency is for Grade 7
VA	26	61	-35	
NE	31	64	-33	State proficiency is combined grades 6-9
AR	14	46	-32	
MN	40	72	-32	
RI	24	56	-32	Broke down Math/Reading into "basic, concepts, etc." scores.
CA	18	49	-31	State proficiency is for Grade 7
PA	21	51	-30	2001 data
IN	31	60	-29	
MD	29	53	-24	
OH	31	54	-23	
DE	19	41	-22	
ID	27	49	-22	

STATE	NAEP Proficiency	State Proficiency	DIFFERENCE	NOTES
CT	34	55	-21	
NV	20	41	-21	State proficiency is for Grade 7
IL	27	47	-20	
KS	34	53	-19	State proficiency is for Grade 7
MT	37	54	-17	2001 data
OR	32	48	-16	
NY	26	41	-15	
WI	32	42	-10	
AK	30	39	-9	
DC	6	15	-9	
UT	26	35	-9	
CO	25	33	-8	
WY	25	32	7	
KY	21	25	-4	
MA	32	34	-2	
NH	25	27	-2	State proficiency is for Grade 6
SC	18	20	2	
VT	32	34	-2	Broke down Math/Reading into "basic, concepts, etc." scores.
WA	26	28	-2	State proficiency is for Grade 7
AZ	21	18	3	
LA	12	8	4	
MO	22	14	8	
ME	32	21	11	

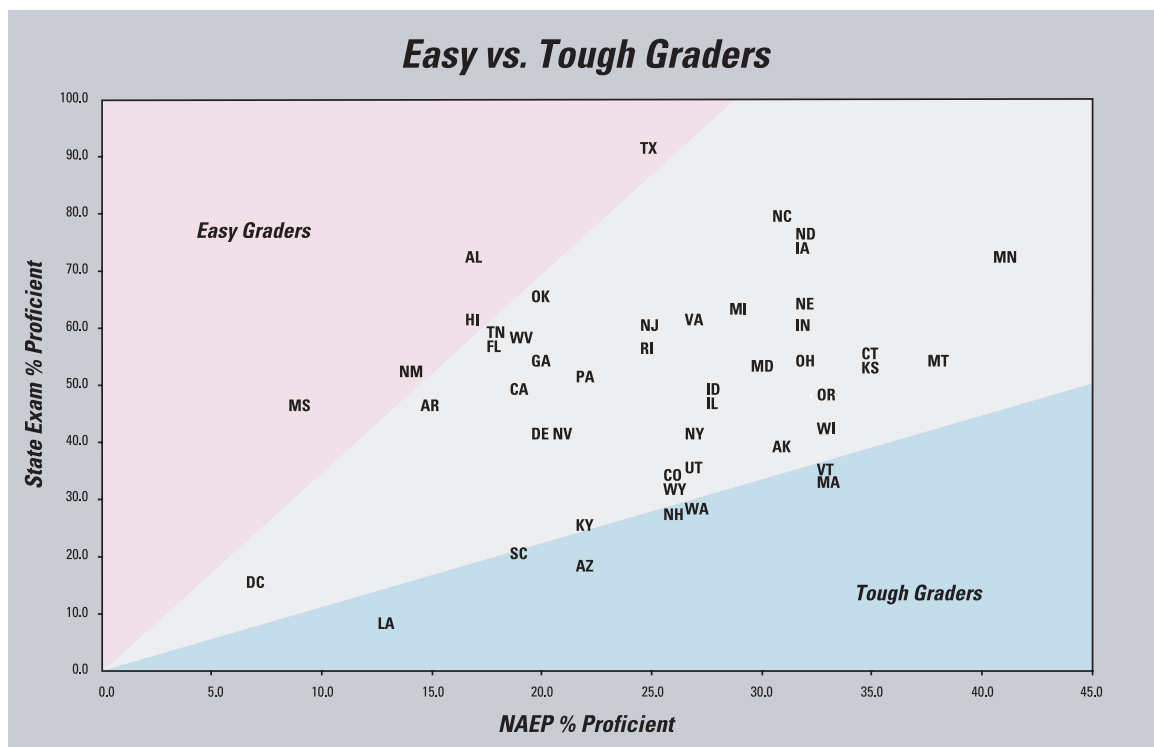
Most states label many more students proficient on their own tests than NAEP finds by its measures. For example, Nevada considers 41% of its eighth-graders proficient in math as judged by its test, while NAEP labels only 20% of Nevada eight-graders as proficient. This is not in itself very illuminating, since forty-six of the fifty states for which we had appropriate data were more generous to their eighth-graders than NAEP would have indicated, by an average of slightly more than twenty-three percentage points. NAEP may be a standard in the sense that it is widely deployed but not, for better or worse, in the sense that states are generally aligned to it.²

Given this nearly universal discrepancy we thought it more helpful to look at how states diverged from one another in their common divergence from NAEP. In other words, relative to one another which states seem to be setting the proficiency bar low and which high. The following graph illustrates the pattern.

States Have Varying Definitions of Proficiency



Divergence is not the only issue. It is worth noting the degree of proficiency claimed as well. For example, though Louisiana and Arizona are much “tougher” state graders than, say, Montana or Minnesota, neither the states nor NAEP makes much of a claim for student proficiencies there.



We do not have a position as to whether states should adjust their practice to NAEP or vice versa. Given, though, that NCLB requires states to use a second test to confirm AYP, and that most if not all states will use NAEP, it does seem prudent to begin a discussion of the discrepancies.