



GRADING EDUCATION:

Making New York's Schools More Accountable

March 2007

Accountability • Innovation • Choice

Phone: (518) 383-2598 • Fax: (518) 383-2841 • 4 Chelsea Park, 2nd Floor, Clifton Park, NY 12065 • info@nyfera.org • www.nyfera.org

© The Foundation for Education Reform & Accountability, March 2007

NOTE: Nothing written herein is to be necessarily construed as reflecting the views of the Foundation for Education Reform & Accountability or as an attempt to aid or hinder the passage of any bill or resolution before the New York State Legislature or the United States Congress.



GRADING EDUCATION: Making New York’s Schools More Accountable

Table of Contents

EXECUTIVE SUMMARY.....	1
HISTORY OF EDUCATION ACCOUNTABILITY IN NEW YORK.....	5
ACCOUNTABILITY VIA <i>NO CHILD LEFT BEHIND</i>	15
MAKE ACCOUNTABILITY MEASUREMENTS TRANSPARENT.....	23
MEASURE THE RIGHT STUFF, AND MEASURE IT WELL.....	27
TIMELY FEEDBACK.....	37
HOW VALID ARE THE RESULTS?.....	41
THE SPITZER REFORM PROPOSALS.....	45
TRUE ACCOUNTABILITY: REAL INCENTIVES, REAL CONSEQUENCES.....	47

EXECUTIVE SUMMARY

Governor Eliot Spitzer outlined a sweeping vision for reforming education in New York State in an historic speech at Chancellor's Hall in the State Education Building in Albany on January 26, 2007.

While committing to push through a record increase in state funding for New York's public schools, Governor Spitzer promised a renewed focus on reform and accountability. In his words:

[F]or New York to become the economic engine it once was, to create jobs, opportunity and prosperity, we must change the way we educate our children....My vision for education reform is built on a single premise: to be effective, new funding must be tied to a comprehensive agenda of reform and accountability....[T]here will be no more excuses for failure. The debate will no longer be about money, but about performance; the goal will no longer be adequacy but excellence; and the timetable will no longer be tomorrow but today.

The Governor also has pledged to raise the cap on public charter schools from 100 to 250 schools, create a state tax deduction of \$1,000 for private tuition expenses, and expand state support to nonpublic schools – fulfilling campaign pledges.

The Governor's vision for increased accountability is on target, but will require a dramatic restructuring in the state's educational accountability system to work.

For more than a generation, the New York State Board of Regents and a series of State Education Commissioners have built an extensive accountability system, centered on subject-by-subject state learning standards, an extensive system of assessments, public reporting, and a graduated system of consequences for poorly performing schools. This effort has not been without much controversy, from affluent suburban parents who oppose the basic notion of performance testing, to parents whose kids are trapped in failing inner-city schools and who fear that higher standards will keep their children from graduating on time, to others in the business community who fear the state is not moving quickly enough to raise expectations and sanction schools that do not measure up.

The Board of Regents and the Education Commissioner deserve much credit for not buckling to those who would dismantle their whole accountability effort.

Yet, the time for doing even more and doing it better has arrived. Governor Spitzer is right to demand much higher accountability given the dramatically higher level of resources he will be providing to our public schools.

The problem is that the current system simply is not designed nor equipped to deliver the level of accountability that Governor Spitzer has called for and that is needed. This report outlines the need to consider the following specific reforms:

1. Measuring the Right Stuff and Measuring It Well: The State needs to hold schools accountable for the gains that they make with students over time, rather than the existing “status” or snapshot approach currently used. The current approach compares students in a single grade each year, even though the demographics of each grade can change dramatically from year to year, especially in urban schools. Measuring gains or growth in student achievement of same-student cohorts is a better way to hold schools accountable.
2. Making Accountability Transparent and Understandable: The current approach of sorting student performance into four levels (Level 1, Level 2, Level 3, and Level 4) is counter-intuitive for much of the public. People are much more accustomed to letter grades, a numeric 0-100 grading system, or at least an approach in which becoming #1 is considered the highest aspiration. On balance, the letter-grade approach for labeling schools, already used in Florida statewide and now being started in New York City, is a better approach. A meaningful accountability system must be understandable to the public at large in order to work
3. Fairer and More Accurate Labeling: Presently, because the measurement of academic performance is based on snapshots of different groups of students each year, rather than tracking same students over time, good schools are mislabeled as bad schools and *vice versa*. Even when the state implements a growth model, mislabeling will still occur if the state and federal government do not change the subgroup “triggers” for failing to meet yearly progress targets. Presently, an individual K-8 school, for example, can be labeled a “school in need of improvement” if it misses the mark in any of more than 50 separate subpopulations (reflecting the number of assessments, subjects, racial, economic, and academic groupings). Labeling a school that misses virtually all of its markers the same as a school that misses one or two simply makes no sense and substantially undermines the overall system’s credibility.
4. Eliminating Conflicts of Interest: The existing approach of allowing district schools to grade their own exams presents unavoidable conflicts of interest. To remove this conflict of interest and increase consistency throughout the scoring process, the state should prohibit teachers, schools, and districts from scoring their own exams, and should seriously consider centralizing and/or contracting-out the scoring of non-multiple choice questions.
5. Ensuring Timely Reporting: The State Education Department released the results of state exams in reading and math in grades 3 through 8 for school year 2005-06 the following school year in September and October. This delay was longer than in any other state last year, and is inexcusable. After much criticism, the Department has proposed releasing results in May and June of the same school year. The Department’s effort to shrink the turnaround time even further should be encouraged, and the Department should be granted whatever resources are needed to make this happen.
6. Providing Real Incentives and Real Consequences: Ultimately, the reporting of state test results does not advance accountability unless consequences result from strong or weak

performance. Once the state's assessment and reporting system are restructured, schools that do well should receive additional funding as an incentive and schools that do poorly should face serious consequences, which could include loss of funding, expanded choice options for parents, and/or, as the Governor has proposed, dismissal of the responsible school superintendent, school board, or principal.

The parameters laid out by Governor Spitzer for overhauling New York's educational accountability system are promising. As New York moves towards higher accountability, state legislators and other politicians should be mindful that, while they are right to call for an improved system, they simply do not have the expertise to decide technical questions of vertical integration of state assessments, the proper type of growth model to be used to measure advances in student achievement, or what performance benchmarks for schools are appropriately high. The New York State Board of Regents and the State Education Department need to be invested with the responsibility and resources to develop the state's new accountability system within broad parameters and timeframes agreed upon by the Governor and the State Legislature.

Governor Spitzer is right to link supplying any additional state financial resources to schools with additional accountability. The state has much work ahead to make this vision a practical reality.

HISTORY OF EDUCATION ACCOUNTABILITY IN NEW YORK

Diane Ravitch, Assistant Secretary in the U.S. Department of Education in the early 1990s, aptly summarized the distinction between commonplace testing in schools and the use of tests as a fundamental component of an educational accountability system:

Testing has long been a staple in American public education. Schools and colleges administered tests of various kinds in the nineteenth century and used them to limit promotion to the next grade and for college admission. But the contemporary idea of accountability, that is, holding not only students but also teachers, principals, schools, and even school districts accountable for student performance, is a recent invention. The idea of measuring the quality of education by the academic performance of students is not one with a long pedigree.¹

National Trends Toward Accountability

Scholars credit the beginning of a shift in American education thought toward accountability to the 1966 publication of the landmark report *Equality of Educational Opportunity* (commonly known as the Coleman Report for its lead author, James Coleman). This report, which examined how resources affected student achievement, was among the first to change its research focus from inputs to outcomes. Ravitch notes:

Before the Coleman Report, education reform had focused solely on the issue of resources, on the assumption that more generous provisions for teachers' salaries, facilities, textbooks, and supplies would fix whatever ailed the nation's schools. After the Coleman report, reformers advanced a broader array of proposals, many of which sought to changes in performance rather than (or in addition to) increases in resources.²

The shift of reform rhetoric away from calling for changes in inputs and instead insisting on a focus on outcomes was accentuated in 1970 with the establishment by the U.S. Education Department of the National Assessment of Educational Progress (NAEP) exams and the resulting wide availability of test result trends for American students. (After 1992, NAEP results from individual participating states were reported, thus offering a new measure of comparability.)

The first attempt by many states to respond to the growing focus on outcomes was to establish a system of minimum competency testing. These tests focused on the lower end of the student achievement scale. In the 1970s and early 1980s, minimum competency testing systems

¹ Ravitch, Diane, "Testing and Accountability, Historically Considered," in *School Accountability* (Stanford, CA: Hoover Institution Press, 2002), Evers, Williamson M., and Walberg, Herbert J., ed., p.9. Ravitch currently is a research professor at New York University and the Brown Chair of Education Policy at the Brookings Institution.

² Ravitch, "Testing and Accountability, Historically Considered," in *School Accountability*, p.14.

spread rapidly throughout the states: while in 1973 only 3 states had such exams, by 1983 the number of states with some form of minimum competency exams grew to 34.³

The growing trend to measure student progress toward specified academic outcomes was followed by a call for policymakers to “do something” to address the dismal results being revealed, a mandate that was crystallized in April 1983 with the publication of *A Nation at Risk: The Imperative for Educational Reform*, by the National Commission on Excellence in Education. This landmark report did not mince words: “...the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a Nation and a people.”⁴

In 1989, the nation’s governors met in an unprecedented summit with President George H.W. Bush to address concerns about the states’ educational performance. The result was a commitment to a set of national educational goals, dubbed *Goals 2000*. While rife with little-worth platitudes such as “...American students will become top world performers in science,” the policy statements in *Goals 2000* largely fixed educational policymakers’ sights on the important and growing trend of measuring student progress toward identified content goals, or standards-based reform. Education economist Eric Hanushek and political scientist Margaret Raymond summed up what they called a “sea change in policy perspective” this way:

*If one is interested in outcomes, one should focus on outcomes. As simple as this principle might be, it has not been recognized previously.*⁵

The Federal *No Child Left Behind Act*

The national movement toward standards-based accountability culminated in January 2002, when President George W. Bush signed into law what was billed as “the most sweeping education reform in a generation”: the *No Child Left Behind Act*.⁶ Rooted in standards-based reform, *No Child Left Behind* (NCLB) initiated the most dramatic federal mandate on states to establish learning standards, assess academic achievement, collect data, report progress, and establish consequences for falling short of academic proficiency goals. Even states that already had outcome-based assessment systems were required to submit new plans to the U.S. Education Department, revised to meet the annual testing, more complete data collection, and low-performance sanctions mandated under NCLB.

³ Robert L. Linn, “Assessments and Accountability,” CSE Technical Report 490, Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, November 1998, p.6

⁴ <http://www.ed.gov/pubs/NatAtRisk/risk.html>.

⁵ Hanushek, Eric A., and Raymond, Margaret E., “Sorting Out Accountability Systems,” in *School Accountability*, (Stanford, CA: Hoover Institution Press, 2002), p.81.

⁶ “Fact Sheet: No Child Left Behind Act,” The White House, Office of the Press Secretary, January 8, 2002. <http://www.whitehouse.gov/news/releases/2002/01/20020108.html>.

Testing and Educational Accountability in New York

New York reflected – and sometimes led – the nationwide trend toward standards-based and outcome-based education accountability systems. Shortly after the publication of the Commission on Excellence in Education’s *A Nation at Risk*, for example, New York State became one of the first states to link test scores more directly to school improvement and accountability efforts.⁷

History

In 1984, the year after release of the landmark *A Nation at Risk* report, the New York State Board of Regents issued the “Regents Action Plan to Improve Elementary and Secondary Education Results in New York.” Among the mandates in this plan was a requirement for every school district in the state to annually issue a report that summarized numerous educational indicators. Among the achievement measures required to be reported were the following:

- The total number of students taking the test and the percent scoring above the state minimum level for the Grade 3 and Grade 6 reading and math exams, the Grade 5 writing exam, and the Grade 6 social studies exam; and,
- The number of students tested and percent of students passing the Regents Competency Test and Regents Examinations.

These reports, called Comprehensive Assessment Reports (CAR), also required data to be collected and reported on numerous other indicators, such as attendance rates, average class size, racial/ethnic distribution, and pupil-to-staff ratios.

While the CAR system was a significant first step in developing a statewide accountability system, its flaws were quickly recognized. The Public Policy Institute of New York State noted:

Despite its name, CAR is not comprehensive, nor does it adequately assess all features of school performance. Moreover, a related focus on minimum competencies makes CAR irrelevant to 90 percent of the state’s districts, while tending to obscure both improvements and declines in performance among troubled schools.⁸

Pervasive low test scores led the state’s then-Commissioner of Education, Thomas Sobol, to declare in his 1988 annual report to the governor and state legislature that there was “a

⁷ “Separate and Unequal: The Reading Gap in New York’s Elementary Schools,” The Business Council of New York State, 1997, Section 1. <http://www.bcnys.org/ppi/sep1.htm>.

⁸ “Measuring Up In Our Schools: Improving Assessment and Accountability,” The Public Policy Institute of New York State, December 1988, p.3.

dramatic pattern of school failure among a substantial portion of our school population... The students whom our system fails are disproportionately poor and minority.”⁹

Beginning in 1989, Commissioner Sobol and the Board of Regents initiated a process to identify and target for intervention the state’s worst-performing schools, called “Schools Under Registration Review,” or SURR. Under this process, schools identified with a pattern of exceptionally low academic performance are visited by teams of designated education experts that prepare a report to the district recommending improvement, and performance targets are set that the schools are expected to achieve in a specified period of time. Each district with an identified SURR school is required to submit to the state a plan of corrective action, which is to be reported upon and revised annually.¹⁰

In March of 1991, the Board of Regents approved a policy statement called *A New Compact for Learning*. It provided a rationale for systemic educational reform and was structured as an agreement between parents, students, business, labor, education, and government representatives to raise academic standards and student achievement. The Regents stated:

For all the other changes around us, the American school today is more as it was in 1900 or 1950 than it is different. And what worked in the 1900s will not work in the 2000s. Either we will make now the fundamental changes needed in the ways we raise and educate our children, or we can begin the slide into a darker and less prosperous time. We must change the system so that we may achieve the results we need. And we are running out of time: either we will make the changes that a new century and a new era require, or we will sink into mediocrity.¹¹

More than a year later, in September 1992, the Board of Regents adopted a follow-up policy statement entitled “A Strategic Plan to Implement A New Compact for Learning: The State’s Role.” This was followed by the distribution of drafts of the types of measurements proposed to be used for judging the achievement of New Compact goals and proposed guidelines for the development of a standards-based state curriculum.

In Fall 1995, the Board of Regents published an overall strategy for raising academic goals for all New York students that included three elements: setting more rigorous learning standards and revising the state’s assessment system to measure progress toward these new benchmarks; “building the capacity of schools to support student learning”; and, developing an accountability system that included public reporting on how well students and schools were performing.¹²

⁹ New York State Education Department, *A Report to the Governor and the Legislature on the Educational Status of the State’s Schools*, February 1988.

¹⁰ New York State Education Department, *Registration Review Process*, 2006. www.oms.nysed.gov/budget/infolinks/surr2006.htm.

¹¹ <http://www.regents.nysed.gov/pol9193.html>.

¹² New York State Education Department, “Timeline on Implementing New York State’s Reform Initiative for Raising Learning Standards for All Students,” 2005. <http://www.oms.nysed.gov/budget/infolinks/chronology.htm>.

In 1996, the Board of Regents adopted a set of 28 broad learning standards for seven content areas: Mathematics, Science and Technology; English Language Arts; Social Studies; Languages Other than English; Health, Physical Education, and Family Consumer Sciences; the Arts; and Career Development and Occupational Studies. With only one exception, these 28 learning standards have remained unchanged since 1996 (a standard in mathematics was revised in January 2005).¹³

Following the adoption of the learning standards, the State Education Department (SED) issued a series of core curricula, designed in part to provide specificity to the learning standards. These curricula outlined the state's expectations of what students must know and be able to do, thus also forming the content basis for the state's subject-area exams.

In December 1996, the state issued its first-ever "report cards" for individual schools and presented test results and other performance indicators for every school in the state. Because of institutional delays in compiling and publishing the information, the report cards reflected performance statistics for the 1994-95 school year, a school year that ended almost 1½ years prior. These annual report cards would be issued for a decade, largely retaining the same format and content throughout those years.

Tests in New York: Elementary & Middle School

Federal aid first enabled SED to develop Pupil Evaluation Program (PEP) tests to provide a measurement of basic reading, writing, and math skills in grades three, six, and nine. These tests were first implemented in the 1965-66 school year.¹⁴

Beginning in 1979, New York State administered more comprehensive reading tests in grades three and six. These "Degrees of Reading Power" (DRP) exams, developed by the College Board and Touchstone Applied Science in consultation with the State Education Department, were basic reading skills tests. Pupils scoring below established cut scores on the DRP exams were assigned to remedial programs.

These exams were supplemented by a science exam in the fourth grade, a writing exam in fifth grade, and a social studies exam in sixth grade. Exams in 8th-grade in social studies and science were added for the 1988-89 school year.

In 1999, SED replaced the grade three and grade six DRP reading exam and the grade 5 writing exam with one 4th-grade and one 8th-grade English language arts exam. Mathematics exams, too, were switched to the fourth and eighth grades (and the social studies test was moved to the fifth grade). The Regents said the switch, done in part to coordinate with the grades in

¹³ <http://www.emsc.nysed.gov/3-8/home.html>.

¹⁴ <http://www.nysl.nysed.gov/edocs/education/sedhist.htm#exam>.

which federal NAEP exams are administered, would help identify whether increases were attributable to better teaching and learning or to manipulation of the test.¹⁵

New exam schedules again were required as of school year 2005-06 to comply with *No Child Left Behind*. Under New York's NCLB plan, all students in grades 3 through 8 are now required to be assessed in English language arts and mathematics (such a system already was in place in NYC). While 4th-grade and 8th-grade science exams also are administered, they were part of the state's NCLB accountability measures only in school years 2005-06 and 2006-07, and will not be for future school years. A 5th-grade social studies exam also is administered.

Tests in New York: High School

The first high school Regents examinations were instituted in 1876 after policymakers noted that an 8th grade exam administered for the previous 10 years had “a salutary influence... stimulating both teachers and pupils to thoroughness in the acquisition of the elementary branches.”¹⁶

New York now offers 15 Regents tests in 5 major subject areas, including: English, science (living environment; chemistry; earth science; physics), social studies (global history and geography; U.S. history and government), mathematics (“A” and “B”¹⁷); and foreign language (French; German; Hebrew; Italian; Latin; Spanish). In order to graduate, students must receive credit for five exams, which include Comprehensive English, Mathematics A, Global History and Geography; U.S. History and Government and one of the four science exams.

For many years, New York offered two types of high school diplomas: a Regents diploma, which was earned by passing the more rigorous Regents Exams; or a “local diploma” for students who couldn't pass these exams. These students most often took the far less rigorous Regents Competency Tests (RCT) in writing, U.S. history and government, global studies, reading, math, and science. The Board of Regents finally phased-out the RCT (except for special education pupils), requiring that students who first entered Grade 9 in September 1999 and thereafter would be required to take the standard Regents exams. This policy contributed to the Regents' repeated hesitation to increase the passing score on the Regents exams required for graduation from 55 to 65 (see below).¹⁸

¹⁵ Anemona Hartocollis, “Stiffer Standards for 4th and 8th Grades,” *New York Times*, February 11, 1998.

¹⁶ “History of Regents Examinations: 1865 to Present,” New York State Education Department summary, November 24, 1987, p.1.

¹⁷ 2006-07 is the last year that Math A and Math B – exams covering an amalgamation of subject areas within mathematics and geometry – will be offered. In 2005, the Board of Regents approved a realignment of math exams so that by 2010-11 there will be three math exams offered: Algebra; Geometry; and Integrated Algebra & Trigonometry.

¹⁸ For more on raising requirements for high school graduation in New York, see: Elizabeth DeBray, “Richard Mills and the New York State Board of Regents, 1995-2001,” *Journal of Cases in Educational Leadership*, Spring 2004.

Testing Benchmarks: Elementary-Middle Schools

As of the 1995-96 school year, the New York State Board of Regents had set a standard that at least 65 percent of students in both the third and sixth grades (the only grades in which elementary-level tests were administered at the time) in each school should achieve minimum competency skills in reading. Even with most schools not achieving this standard – including a vast majority of low-income and high-minority schools – for the 1996-97 school year the Regents increased their expectations to a 90 percent pass rate.

A decade after this proclamation, the reality in New York is a dramatic shortfall from this goal: in school year 2004-05,¹⁹ only 12 percent of public schools statewide administering the 4th-grade English language arts exams posted pass rates of 90 percent or better; 1.4 percent of public schools posted 90 percent or better pass rates on the 8th-grade exam.²⁰

High expectations for student performance are not the problem, of course. The inability of New York’s public education system to meet these expectations is.

Testing Benchmarks: High School Regents Exams

New York’s education policymakers have historically strived to increase high school graduation standards as well.

For years, students needed to score only a 55 (a combination of weighted point totals, not a percent of correct answers) to pass a high school Regents exam in any subject. In 1996, the Board of Regents adopted a plan to increase this required passing score to 65, and eliminate the lower-standard “local diploma” altogether. The higher benchmark was to first apply to students entering the 2001-02 school year as 9th-graders.

Facing what it projected to be a publicly embarrassing shortfall in the number of students who would successfully reach this higher standard, the Board of Regents has postponed the implementation of the higher passing score time and time again. In 2005 – nearly a decade after it first proposed increasing the passing score to 65 – the Board of Regents enacted a four-year phase-in schedule of the new benchmarks as follows:

- For students entering grade 9 in 2005: General education students must have at least two scores at 65 or above on the five required Regents exams and all scores at 55 or above.
- For students entering grade 9 in 2006: General education students must have at least three scores at 65 or above on the five required Regents exams and all scores at 55 or above.

¹⁹ 2004-05 was the final school year in which the tests referenced by the Regents were administered. From 2005-06 forward, new exams were administered in grades 3-8 in accordance with NCLB.

²⁰ Information and Reporting Services, State Education Department, *Media File Grade 4 & Grade 8 English Language Arts* (May 18, 2005), <http://www.emsc.nysed.gov/irts/press-release/20050518/home.htm>.

- For students entering grade 9 in 2007: General education students must have at least four scores at 65 or above on the five required Regents exams and all scores at 55 or above.
- For students entering grade 9 in 2008: General education students must pass all five required Regents exams at a score of 65 or above.²¹

The Regents also approved an appeals process through which students who score within three points of 65 would be eligible to appeal their scores in various instances to receive credit toward fulfilling the graduation requirements.

In an editorial at the time this new phased-in schedule was adopted, the *New York Post* lamented the continuing delay to implement a hard-and-fast passing score: “Retreating yet again from high standards in New York’s public schools, the state Board of Regents...delayed raising high-school graduation requirements for at least three years... [T]his is just the latest in a series of delays the Regent’s have OK’d.”²²

Prospects for Strong School Accountability Systems

The drive to enact a strong system of accountability in public education both in New York State and across the nation began and remains largely divided into two camps: one side calling for more and more resources (usually public funding) to provide the inputs deemed necessary to achieve desired outcomes; and the other side calling for accurate and easily digestible measures of academic progress with a structure of punishments for failure and rewards and incentives for improvement. Too often, the first group has resisted the reforms of the second and the second group has conceded to the wishes of the first.

Williamson Evers, a member of the National Educational Research Policy and Priorities Board and research fellow at Stanford University’s Hoover Institution, and Herbert Walberg, University Scholar at the University of Chicago, summed-up the need for strong school accountability systems this way:

If schools were doing well or even passably well, policymakers might be deservedly reluctant to insist on substantially greater accountability. Since schools are not doing well, however, the burden of proof in explaining the state of affairs should be on the present system. Yet the American K-12 system is distinctively unaccountable compared with other aspects of American life and compared with education systems in other countries.²³

²¹ James A. Kadamus, *School Executive’s Bulletin*, Office of Elementary, Middle, Secondary, and Continuing Education, New York State Education Department, Summer 2005. <http://www.emsc.nysed.gov/SB/sum05site/schlbltnjau05.html>.

²² “The Regents Regress,” editorial, *New York Post*, June 23, 2005.

²³ Evers, Williamson M., and Walberg, Herbert J., ed., “Introduction and Overview” in *School Accountability* (Stanford, CA: Hoover Institution Press, 2002), p.2.

The few states that have bucked what often was intense political pressure and imposed relatively strong outcome-based accountability systems – including North Carolina, Massachusetts, Texas, and Virginia – have experienced measurable and often dramatic gains in student performance, increases that have been especially significant for minority student populations.²⁴

Chester E. Finn, President of the Washington D.C.-based Thomas B. Fordham Foundation, an organization that has spent years comparing and grading state accountability systems, noted:

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 appropriate sanctions for those who fail?²⁵

Using these guidelines, a 2004 comparative analysis by the Fordham Foundation of state learning standards, content of state tests, alignment of tests to standards, rigor of state exams, test trustworthiness and openness, and state accountability policies, rated New York’s accountability system as “solid,” or a 4.1 on a 1-to-5 scale. New York was one of eight states to receive a “solid” rating, and topped the national average rating of 3.7.²⁶

²⁴ See Ravitch, Diane, “Testing and Accountability, Historically Considered,” in *School Accountability*, p.19, for example.

²⁵ In *Grading the Systems: The Guide to State Standards, Tests, and Accountability Policies*, Cross, Richard W., Rebarber, Theodor, and Torres, Justin, ed. (Thomas B. Fordham Foundation: Washington D.C., January 2004), p.i.

²⁶ *Grading the Systems: The Guide to State Standards, Tests, and Accountability Policies*, , p.9.

ACCOUNTABILITY VIA *NO CHILD LEFT BEHIND*

The landmark federal *No Child Left Behind* Act, signed into law on January 8, 2002, initiated the most sweeping reform of public education in decades. The Act was an attempt to ensure that all children meet basic academic standards by 2014 through the strategic creation and application of statewide accountability goals, and incorporating a focus on closing any existing state racial achievement gaps. *No Child Left Behind* (NCLB) is widely recognized as the largest and most comprehensive federal education data-gathering mandate in history.

NCLB makes state education departments responsible for designing and securing federal approval for accountability plans, and for ensuring that schools in turn are held responsible for their students' results, increasing parental choice for students in underperforming schools, and publicly reporting test data analysis. New York was one of the first five states to have its accountability plan approved by the U.S. Education Department.

As the Act approaches its federal deadline for renewal by Congress in 2007, much attention has been given to areas of the law that need improvement. Criticism of NCLB has come from stakeholders in the traditional public education establishment that would like to see far less reliance on and consequences from standardized testing. Other groups who are supporters of greater accountability also have criticized NCLB as containing too many loopholes and weak provisions that fail to effectively require solutions that directly benefit students in low-performing schools.

***No Child Left Behind* Accountability Basics in New York**

Under *No Child Left Behind*, each state is responsible for designing and implementing a federally approved accountability system that tracks how well schools and districts make progress each year toward achievement benchmarks. This mandated continuous improvement, known as Adequate Yearly Progress (AYP), is designed to meet NCLB's ultimate goal of universal proficiency by 2014. If a school successfully meets AYP goals, the school has attained a basic minimum level of performance that is keeping the school on track toward projected proficiency for all students. In New York State, AYP is currently measured using data from state English language arts, math, and science exams, and for high schools, graduation rates as well.

New York uses three measurements, or "performance indicators," to determine AYP. In the elementary- and middle-school years, students are tested annually in grades 3 through 8 in both English language arts and math. In grades 4 and 8, an additional exam is administered in science.²⁷ At the secondary level, the three performance indicators used to determine AYP are performance on Regents exams (the state's basic commencement competency tests) in English language arts and math, plus graduation rates. Alternative tests, such as the less-challenging

²⁷ In the 2007-08 school year, 4th and 8th grade science will no longer be used to calculate AYP. Attendance rates will instead be used as the third performance indicator.

Regents Competency Tests and the New York State Alternate Assessment, are offered to students with disabilities and factored into AYP measures.

To ensure that schools focus on the progress of each student, NCLB requires that schools meet AYP requirements on an average school-wide basis and within numerous subgroups of students. Schools also must ensure that student participation rates in the tests meet or exceed 95 percent for each subgroup²⁸ to ensure that lower performing students are not “counseled out” of taking the exams and the test results accurately reflect the true student population at the school.

The identification of subgroups and the measurement of academic performance within each subgroup is an important component of NCLB, and one that is designed to provide data that help identify and close chronic academic racial achievement gaps, for example. Accountability groups required to be measured under New York State’s NCLB accountability plan include the following:

- All Students
- Students with Disabilities
- Economically Disadvantaged Students
- Limited English Proficient Students
- American Indian/Alaskan Students
- Asian Students
- Black Students
- Hispanic Students
- White Students

Test data and participation rates for these groups of students are used to determine if schools have met the minimum standards set by the state to make AYP and remain in good standing.²⁹

Consequences for Failure to Make Adequate Yearly Progress

The failure of any school to make AYP for two consecutive years is used as a determinant that the school is not on track to achieve universal proficiency by the 2014 school year, and thus are labeled “in need of improvement.” These schools are publicly identified by the State Education Department and are required to develop and submit a plan outlining a series of reforms designed to lead to improve academics at the school. Schools that continue to fail are identified as follows:

²⁸ New York State has declared that subgroups at each school must have 40 or more students for these reporting regulations to apply.

²⁹ Extensive discussion of New York State’s accountability system, including how AYP is measured, can be found at: www.emsc.nysed.gov/irts/accountability/2005-06/accountability-rules-Nov2006_files/frame.htm.

Number of Years Failing to Make AYP in any Accountability Area	School Status Under NCLB
After 2 Consecutive Years	School In Need of Improvement -Year 1
After 3 Consecutive Years	School In Need of Improvement -Year 2
After 4 Consecutive Years	Corrective Action
After 5 Consecutive Years	Planning for Restructuring
After 6 Consecutive Years	Restructuring

As the years pass, provisions of NCLB are triggered that initiate a series of mandated school choice options and district interventions. During the first year of identification as “in need of improvement” (after a school’s second consecutive year of missing a target AYP), NCLB requires the district to offer students the option of transferring to a non-failing public school. After the second year of a school being labeled “in need of improvement” (three consecutive years of failing to meet AYP), students are offered free supplemental educational services, such as tutoring, in addition to the school-choice option.

District interventions begin when a school is identified as requiring “corrective action,” which occurs after failing to meet AYP for four consecutive years. These interventions can include recruiting new personnel, changing curricular offerings, and restructuring governance. During the “planning for restructuring” phase, districts with failing schools are required to submit to the state education department a plan for alternative governance at the school. The districts are required to carry out this plan if and when a school fails for the sixth consecutive year.

It is also possible for entire districts to be identified as “in need of improvement” based on district-wide results.

School Choice under NCLB

The original draft of the *No Child Left Behind* Act dramatically expanded school choice, allowing students trapped in failing schools the option of transferring to both public *and* private non-failing schools.³⁰ The final NCLB Act was passed with a much more limited and weaker school choice component, however.

Under the adopted version of NCLB, students in failing schools are able to transfer only to other public schools within the same district, and then only if seats are determined to be available. Particularly in urban areas, non-failing district schools often are overcrowded, and district schools that do have vacancies typically are failing just as badly as students’ resident schools. In New York, hundreds of thousands of public school students attending schools classified as “in need of improvement” and eligible to transfer to non-failing schools literally have no choice of quality schools to transfer.

³⁰ See: Andrew Rudalevige, “No Child Left Behind: Forging a Congressional Compromise,” in Peterson, Paul E., and West, Martin R., eds. *No Child Left Behind?* (Washington D.C.: Brookings Institution Press, 2003), p.23.

With strong school-choice provisions left out of the final bill, architects of NCLB relegated options for creating new capacity to accommodate students transferring out of failing schools to a “non-regulatory guidance” policy. Suggestions here include:

- Encouraging the creation of new charter schools within the district;
- Redrawing district attendance zones;
- Initiating inter-district choice programs with neighboring districts;
- Establishing programs through which local private schools can absorb some of the transferring students.³¹

Supplemental Education Services

Additional intervention mandated by NCLB is the provision of Supplemental Educational Services (SES) for students in schools failing to meet AYP for three or more consecutive years. Students and their parents are able to choose from a list of state-approved providers, including private and nonprofit providers, to receive tutoring services designed to fill the gaps in a child’s education left by the underperforming district school. The district is responsible for paying for these services using federal Title I funds, and may retain for its own use any unused Title I funds.

The provision of SES has been inconsistent in New York, however, and its impact has been largely negligible. According to data submitted by school districts to the State Education Department, in every major city other than New York City, fewer than one-third of students eligible for SES services request them (see table below). Because districts are allowed to retain Title I funds not spent on SES, and because reporting that sizable shares of students require supplemental instruction reflects poorly on district schools’ ability to do their job adequately, there is a built-in disincentive for districts to ensure that all eligible students pursue and receive these tutoring services.

STUDENTS ELIGIBLE FOR AND RECEIVING SUPPLEMENTAL EDUCATIONAL SERVICES (SES)

2004-05

CITY	# of Students Eligible for SES	# of Students Receiving SES	% of Eligible Students Receiving SES
New York City	218,174	97,596	45 %
Syracuse	3,166	1,028	32 %
Buffalo	11,922	2,801	23 %
Rochester	12,277	2,320	19 %
Albany	1,007	61	6 %
Yonkers	4,083	69	2 %

Source: *Annual Report of SES Provided in the State of New York*, New York State Education Department. Data is for the 2004-05 school year as reported by school districts.

³¹ “Public School Choice,” *No Child Left Behind*, Non-Regulatory Guidance, U.S. Department of Education, Draft, February 6, 2004.

Leaving Kids Behind

Despite the obvious intentions of NCLB to identify underperforming schools and to institute a structure that provides comprehensive remediation for them, there are no real consequences for failure given the way NCLB is being implemented in New York. There are an estimated 725,000 students currently enrolled in failing schools,³² and most are receiving little or no support allowing them to receive a better education.

Year after year, New York releases its lists of failing schools, but in many cases these designations have absolutely no impact on improving the quality of education individual students receive. As currently designed, NCLB presumes that traditionally designed public schools will continue to deliver educational services and, even without real sanctions for failure, improve simply to avoid negative labeling.

The more serious reforms – forcing a school to change its curricular offerings or to hire a new leader – are not guarantees of success by themselves. These reforms also come only after years of unchecked underperformance, meaning that an entire class of students could proceed through every grade level offered by a school without ever being touched by an improvement measure.

Fairness of the “Failing School” Label

Schools are identified as “in need of improvement” when students in an accountability group fail to make adequate yearly progress in the same subject area over two or more consecutive school years. Failure to make AYP in a particular subject and grade level can occur 1) if test results fall short of the annual performance target, or 2) if fewer than 95 percent of students are tested. When all accountability subgroups and assessment measures are considered, there are a multitude of ways through which a single school can be classified into improvement status.

Assume a K-8 school has enough students in each subgroup area to be responsible for AYP in all nine possible student groups. The school must make AYP in English language arts, mathematics, and science for each of these nine groups. In addition, the school must meet the 95 percent participation rate for each of the nine groups in all three subject areas. Therefore, this particular K-8 school would need to fulfill each of 54 different ways to make AYP (see chart on the following page). If after two consecutive years the school has failed to make AYP in any subgroup reporting category in a given subject, even if the subgroup that fails is different from one year to the next, the school is identified as “in need of improvement.”

³² Campaign for Fiscal Equity, Inc., “Education Advocates Disclose Number of NYC & NYS Students in Schools that Do Not Meet NCLB or State Standards,” January 24, 2007.

**MULTIPLE WAYS FOR A K-8 SCHOOL
TO GET LABELED A “FAILING SCHOOL” UNDER
NO CHILD LEFT BEHIND**

2006-07 School Year

	Combined (all-grade) AYP Targets		
	ELA	Math	Science
All Students Academics			
Participation (95% min.)			
Students with Disabilities Academics			
Participation (95% min.)			
Limited English Proficient Academics			
Participation (95% min.)			
Economically Disadvantaged Academics			
Participation (95% min.)			
American Indian/Alaskan Academics			
Participation (95% min.)			
Asian Students Academics			
Participation (95% min.)			
Black Students Academics			
Participation (95% min.)			
Hispanic Students Academics			
Participation (95% min.)			
White Students Academics			
Participation (95% min.)			

Failure to hit the annual yearly progress target in any one of these three subject areas in two consecutive years earns a school the “failing” label

The attempt to ensure that each demographic subgroup in a school is progressing under each measure of progress has resulted, critics argue, in an unfair application of the “failing school” label. A school that fails to make AYP in one subject and one subgroup is treated the same way as a school that fails to make AYP in all three subjects and multiple subgroups, for example. Critics also have justifiably argued that less diverse schools are provided a structural advantage under NCLB because they simply have fewer subgroups for which AYP targets must be met. A school with a homogenous and/or wealthier student population by design has fewer opportunities to fail to make AYP than a school with large populations of minority and economically disadvantaged students.³³

Exacerbating the question of fairness is that once a school is labeled “in need of improvement,” the State Education department fails to clearly and publicly identify whether such a designation is made because one subgroup of students failed to make AYP, an entire class fell short, or for some other reason.

NCLB’s fundamental goal that *all* students should perform at a proficient level is worthy, and as a result a school that fails to hit its AYP target for *any* student subgroup legitimately is dinged for failing its mission to provide an adequate education. Yet there is legitimacy to strive for a measure that can distinguish between a school that is miserably failing all or most of its students, and a school that has a smaller, presumably more “fixable” problem. Designing an accountability system that allows differentiation between the two types of schools likely will be even more complicated than the current administration of NCLB already is, however, and caution should be exercised to ensure that greater inefficiencies do not arise.

Holding schools accountable for the progress of their demographic subgroups has revealed how many districts with a reputation for excellence are in fact failing a segment of its student population. However, the same reporting mechanism has told the public nothing new about schools that for years have been unable to succeed with students under any circumstances. For those schools that have been chronically failing, the use of subgroup reporting has made little impact.

Reauthorization

NCLB comes up for reauthorization by the U.S. Congress in 2007. It presents an opportunity to fine-tune issues surrounding testing and other accountability requirements. *Newsweek* writer Jonathan Alter, calling NCLB “flawed but landmark” and believing that there were funding promises associated with NCLB that were not upheld, stated:

Once...Congress addresses the money problem and works out some testing kinks, the real fault of NCLB will become clear: it doesn’t go far enough.³⁴

³³ For more discussion on the complications of accountability systems, see: Paul E. Peterson and Martin R. West, ed., “No Child Left Behind? The Politics and Practice of School Accountability,” *The Brookings Institution*. Washington, DC, 2003.

³⁴ Jonathan Alter, “Stop Pandering on Education,” *Newsweek*, Feb. 12, 2007 issue.

The *New York Times* also notes the significant effect NCLB has had on state education policies:

Spurred by President Bush's No Child Left Behind law, educators across the nation are putting extraordinary effort into improving the achievement of minority students, who lag so sharply that by 12th grade, the average black or Hispanic student can read and do arithmetic only as well as the average eighth-grade white student... [E]ven critics acknowledge that the requirement that schools release scores categorized by students' race and ethnic group has obliged educators to work harder to narrow the achievement gap.³⁵

³⁵ Sam Dillon, "School Law Spurs Effort to End the Minority Gap," *New York Times*, May 27, 2005.

MAKE ACCOUNTABILITY MEASUREMENTS TRANSPARENT

The State Education Department (SED) determines how many students in a school achieve proficiency on state exams using a scale that sorts students into four ranked groups labeled appropriately, if not unimaginatively, Level 1, Level 2, Level 3, and Level 4. This scaled grouping of scores is used to determine *No Child Left Behind* accountability status.³⁶ Students scoring in Level 1 and Level 2 are deemed to have failed to meet proficiency standards.

The percent of students in a particular grade scoring in Level 3 and Level 4 on a state test in any given subject are deemed to have “passed” the test, having met or exceeded proficiency standards. However, SED does not total or present on the annual school report cards this subtotal of students passing the test – calculations for the test results for each grade and each subject must be made by any individual wishing to know this basic information. Further, the annual report cards, the most comprehensive resource individual school performance distributed by the state, do not calculate a single school-wide passing rate. Each grade and each subject are presented separately. Because of the reporting method chosen by the state, the only academic performance comparisons to other schools that can legitimately be made are between a single subject in a single grade (e.g., 3rd grade English language arts).

New York’s Level 1-4 grading system is a unique reporting system that differs greatly from most every other education performance reporting measure. More people understand the familiar letter-grade system of A, B, C, D, and F, or a numerical score of 1-100 with rational break-points, for example. The inability to compare one school as a whole with another hinders the usefulness of New York’s approach. Accountability systems that use the more transparent and intuitive system of letter-grades, and one which determines a school-wide performance level that can be compared from school to school would prove to be much more valuable to the general public.

Such a system is already in place in Florida, for example. Florida implemented an A-through-F scale in 1999 as part of its statewide A+ Accountability Plan reform initiative. To determine an overall grade for a public school in Florida, the state considers test results that measure individual students’ annual gains, overall school performance, and the performance of the lowest performing students.³⁷ Florida’s accountability system then rewards with additional state financial aid all “A” level schools and all schools that improve from one letter grade to the next. Schools that are consistently graded as “D” and “F” are subject to a set of improvement initiatives, including the mandatory development and implementation of a comprehensive improvement plan. For a time, students in “F” schools were eligible for school-choice vouchers

³⁶ Every student receives an exam scale score which is then converted to a simple performance level of one, two, three, or four. The number of students in performance levels two, three, and four are used to calculate the Performance Indices used to determine whether a school has met AYP targets.

³⁷ For details on the Florida A+ Accountability System: Paul E. Peterson, “Reforming Education in Florida: A Study Prepared by the Koret Task Force on K-12 Education,” *Hoover Institution*, 2006. Online at: <http://www.hoover.org/publications/books/3895672.html>.

that could be used to transfer to a private school. This accountability reform was recently ruled by the Florida courts as unconstitutional, however, and as such has been discontinued.³⁸

Since the practice of assigning grades to schools began in 1999, student achievement in Florida has been on the rise. Not only did low-scoring A, B, C, and D schools improve within their own “comparison” groups, but “F schools made gains that were even larger than would have been expected simply given how low their previous scores were.”³⁹

Similarly, the New York City Department of Education is in the process of implementing its “Children First” reform initiative that in the 2007-08 school year will incorporate progress reports that use an A, B, C, D, and F rating system for individual schools. The City’s letter-grade reporting system will rate schools on three quantitative measures: 1) progress, measured by value-added student growth; 2) performance, measured by average achievement on state exams; and, 3) school environment, using statistics such as attendance rates and community satisfaction. Under the City’s plan, a series of rewards and consequences will be used to hold schools accountable for achievement, as reflected by this new grading system. Schools with persistently low grades (defined as schools chronically earning “C” ratings, or “D”- and “F”-level schools) will face serious consequences including leadership restructuring or closure. High performing schools can receive extra per-pupil funding and staff bonuses.⁴⁰

Both Florida and New York City have adopted accountability systems that report data in clear, concise, easy-to-understand, and easy-to-compare ways. This stands in stark contrast to the muddled, multi-layered approach currently used by New York State.

Also, both Florida and New York City’s accountability plans include measurements of growth in individual student performance. In contrast, New York State measures student performance only as a snapshot in time, with no measurement of how long a student was in attendance at a given school or whether large gains or drops in individual student performance occurred from one year to the next. The data generated from a snapshot is useful in many instances, but it makes sense for an accountability plan to also measure how individual student performance is affected directly as a result of the school’s effort from year to year.

Transparency & Understandability

A welcome change to New York’s education accountability system would be to make the results of student performance testing more transparent. Currently, no single performance indicator summarizes the academic achievement of schools in a way that is easily understood by parents, students, civic and business leaders, or policymakers. Instead, education results in New York are reported using multiple measures that may be useful to analysts but that are not

³⁸ Transfers to other public schools out of failing schools are still possible under the Opportunity Scholarships program, which makes portable per pupil public education aid for children in “F”-rated schools.

³⁹ Jay P. Greene, *An Evaluation of the Florida A-Plus Accountability and School Choice Program*, Center for Civic Innovation, Manhattan Institute (February 2001), p.11. http://www.manhattan-institute.org/html/cr_aplus.htm.

⁴⁰ Details on Children First can be found online at: <http://schools.nyc.gov/Offices/ChildrenFirst/default.htm>.

particularly helpful in painting a clear picture of the success or failure of the state's public schools.

New York State also should strive to ensure that data that reveal academic problems can be traced to individual classroom teachers. Reliable measures, including longitudinal value-added and growth data, can reveal which teachers achieve better results with their students. This data could help provide guidance in designing interventions for teachers in need of mentorship and further training, offer evidence that some teachers aren't performing up to expectations, and provide support for rewarding outstanding teachers.

MEASURE THE RIGHT STUFF, AND MEASURE IT WELL

A key component of any well-designed accountability plan is to ensure that the data being gathered is appropriate and robust enough to measure exactly what stakeholders desire. This is not a particularly easy assignment due in part to the various types of stakeholders and the various measurements each seeks. Parents, students and school administrators want to know which schools are doing a particularly good or bad job of educating their children. School administrators and teachers want to know which classroom teachers show the most (or least) progress with their students. And policymakers want to know whether the education delivery structure as a whole is achieving stated objectives.

Designing Accountability Systems

Accountability systems may be designed in a number of ways. According to the Center for Assessment:

[S]chool performance might be described in four main ways: a) status, or performance at a point in time without reference to previous performance; b) improvement of successive groups (e.g., grade 3 in 2005 compared to grade 3 in 2004), c) student longitudinal growth (e.g., students' performance in grade 4 in 2005 compared to the performance of the same students in grade 3 in 2004), and d) change in rate of change (either of improvement or growth).⁴¹

When considering what type of accountability system to put in place, it is critical to first answer the question: What is it you really want to know? If the answer comes in multiple parts, multiple measurement approaches may be warranted.

Criterion-Referenced Assessments vs. Norm-Referenced Assessments

Because universal student proficiency on state learning standards is the mandated end goal of state assessment systems under *No Child Left Behind*, assessments used must be criterion-referenced. The New York State Education Department notes that such tests “compare a student’s performance to a previously established criterion [meeting state standards] rather than to other students from a normative sample.”⁴² According to the Council of Chief State School Officers, “A criterion referenced growth target emphasizes sufficiency over capacity and

⁴¹ Brian Gong, Marianne Perie, and Jenn Dunn, *Using Student Longitudinal Growth Measures for School Accountability Under No Child Left Behind: An Update To Inform Design Decisions*, Center for Assessment, September 18, 2006, p. 2. <http://www.ccsso.org/content/pdfs/BGongGrowthUpdate091806BG.doc>.

⁴² State Education Department, Office of Vocational and Educational Services for Individuals with Disabilities, *Guide for Determining Eligibility and Special Education Programs and/or Services for Preschool Students with Disabilities* (January 2003), <http://www.vesid.nysed.gov/specialed/publications/preschool/guide/indeval.htm#Fore>.

establishes the amounts of growth needed to have all students reach a set achievement score by a set date. This is the method used under NCLB.”⁴³

By comparison, norm-referenced tests measure student performance against the averaged performance of various comparison groups, such as students of the same grade, age, gender, or racial/ethnic group, or economic class. Norm-referenced assessments typically allow a school to rank the performance of its students in comparison to other students nationwide. The Stanford 10 series, Terra Nova, and the Iowa Test of Basic Skills are examples of such tests. Students tested at the beginning of the school year may fall into the 63rd percentile nationwide in mathematics, for example; by the end of the year, after another administration of the nationally-normed exam, students may fall into the 71st percentile. This measurement shows that this group of students is learning faster than its peers across the country.

What norm-referenced tests don't do is to measure student progress toward a set learning standard. If it is intended that traditionally low-performing student groups are to eventually meet the same performance targets as higher-achieving student groups, larger gains are required for these lower-performing students if the achievement gap is to be overcome. A norm-referenced test doesn't measure the progress toward such goals, and instead references only average performance of a broad group of peer students at any given point in time.

The state exams offered at all grade levels in New York are tied to specified learning standards and are thus criterion-referenced. Many of the state's public charter schools, however, supplement the state's tests with beginning-of-year and end-of-year norm-referenced tests, allowing these schools to document progress relative to students across the nation.

Status Models and Improvement Models

The U.S. Department of Education (USED) required state NCLB accountability plans to be based on “status models,” or snapshots of student performance at the time of test administration that are then compared to predetermined targets, or “Annual Measurable Objectives” (AMOs), set by the state. Meeting these AMOs is a part of the determination of a school's adequate yearly progress.

For schools failing to meet the AMOs under these state status models, student performance also is calculated on an “improvement model” which measures year-to-year performance on an individual assessment (such as 8th grade math in 2005 compared to 8th grade math in 2006). Schools making significant progress – that is, the total number of students testing below proficiency level on a particular assessment decreases by 10 percent from one year to the next – are categorized as entering “safe harbor” zone, which prevents such schools from being labeled as “In Need of Improvement.” The state accountability models required by USED thus combine measurements of how many students score proficient (a status model) and the performance over time of different cohorts on a particular exam (an improvement model).

⁴³ Pete Goldschmidt, et. al., *Policymakers Guide to Growth Models for School Accountability: How do Accountability Models Differ?*, Council of Chief State School Officers (October 2005), p. 10, <http://www.ccsso.org/publications/details.cfm?PublicationID=287>.

While NCLB is revolutionary in that it required states to devise and implement a new accountability system with a goal of universal proficiency by 2014, the required status models and improvement models being used have substantial drawbacks.

First, status models only measure performance for a single point in time (when the assessments are administered) and fail to indicate how much students are gaining academically from their experiences at a particular school. Status models also ignore outside factors that may be contributing to student achievement and all elements of success or failure is attributed the school in which the students are currently enrolled.

Second, status models fail to measure the significance of gains and increases towards proficiency targets. For example, a school may have only 10 percent of its students scoring proficient on 4th grade English language arts exam in 2005 and then show 45 percent scoring proficient in 2006. Despite making major gains from one year to the next on a particular assessment, if the stated target was to have 50 percent of students proficient, a status model would not note the significant gain, but only that the school fell short. Also, all schools failing to meet performance targets are recognized the same, regardless of whether they are one point or, say, 15 points away from the target.

Third, status models don't account for students who could potentially start a school year above grade level, fail to learn anything during that year, and then still know enough to score proficient on state assessment. Status models thus do not penalize a school for failing to teach a high-achieving student anything new.

Finally, status models tend to be biased against school districts populated by lower-income families. Wealthier schools and districts tend to have academically stronger students and are less likely to have students that fail to meet proficiency targets, regardless of how much or how little individual students may progress over time. The inverse is true of low-income schools and districts, which tend to be populated with greater proportions of students that fail to meet proficiency standards. The snapshot of average performance provided by status accountability models almost certainly always show lower-income districts underperforming wealthier districts, regardless of how much learning is actually taking place.

Improvement models also have structural weaknesses. The most obvious weakness is that this type of accountability measure tries to compare two different sets of data: "How did this year's 6th-grade class do compared to last year's 6th-grade class?" for example. Problems arise because a comparison is attempted between two totally different sets of students without accounting for differences in starting academic proficiency or other characteristics. Some researchers have illustrated the problems of improvement models by using the shorthand terminology of "good class, bad class," where one year's students tend to be a naturally high-performing crop and the next year's class tends to be low-performing.⁴⁴ Yet, NCLB measures

⁴⁴ Gong, Perie, and Dunn, *Using Student Longitudinal Growth Measures for School Accountability Under No Child Left Behind: An Update To Inform Design Decisions*, p. 3.

rely on this type of model as well, striving to find continuing increases from one year's grade-level class to the next.

Growth Models

Instead of simply taking a snapshot in time (status model) or measuring the change in performance between one year's class of students and the next (improvement model), states and USED are considering more sophisticated accountability plans known as "growth models" that track individual student achievement over time. Such models account for long-term gains or losses in performance on an individual student and/or class level. Growth models would allow for a class of students' performance on a 6th-grade math assessment to be compared to the same cohort's performance in 5th grade the previous year and even in 4th grade the year before that, for example. Growth models identify if, on average, students are improving or doing worse than they did in the past.

Growth models require a series of assessments that produce comparable results for each grade. In order for student progress to be tracked over multiple grade levels, the content must be measured on a "vertical scale," as explained by educational accountability researchers Robert W. Lissitz and Huynh Huynh:

[S]tudents are expected to show performance improvements at each year, and these improvements should be reflected in a steady increase in their ability to do mathematics. The tests for grades 7 and 8 should be linked so that scores are directly comparable along a common continuum or dimension... The content must have some sense of commonality across grades in order to be meaningfully equated across grade levels. These scales are often considered developmental, in the sense that they encourage the examination of changes in a student's score across grades that indicate the improvement in that student's competency level. Sometimes the equating is only for adjacent grades and sometimes equating is across the whole school experience.⁴⁵

New York's new English language arts and mathematics assessments, developed by CTB/McGraw-Hill, are based on a vertical scale and have overlapping questions. According to the New York State Education Department:

Each test will contain 'overlapping items' in the multiple-choice section of the tests. The content of these items, however, will be based upon the content taught in the lower grade. For example, the overlap items in the grade 5 mathematics test will include a few questions from content areas that were covered in grade 4. These will be the overlap items that will also be covered on the grade 4 test. The grade 5 test will also have a few

⁴⁵ Robert W. Lissitz and Huynh Huynh, "Vertical Equating for State Assessments: Issues and Solutions in Determination of Adequate Yearly Progress and School Accountability," *Practical Assessment, Research & Evaluation* (April 2003), <http://pareonline.net/getvn.asp?v=8&n=10>.

items that will be identical to the grade 6 test. These items will be from content areas covered in grade 5, so they should be achievable for grade 6 students as well. The purpose of the overlapping items is to try to achieve a vertical scale which is often seen as helpful in assessing the growth of individual students from one grade level to the next. With some constraints, a test that is placed on a vertical scale can be used to directly compare a student's score from year to year.⁴⁶

Statewide identification (ID) systems, which assign each student in the state with a unique number and create a profile for each student that is stored in a central database, are required to track student performance over time. This allows the tracking of each student's performance at any school throughout the state he or she attends, and enables schools (and/or the state) to create a trajectory of expected performance for each student. New York has recently initiated a statewide ID system, and this data-collection methodology now could be used for the implementation of a growth model of accountability.⁴⁷

The ability of growth models to track individual student and cohort performance over time make them appealing to states and USED because they provide the truest measure of how much learning is taking place, and provides the greatest detail about where such gains (or losses) occur. A recent study by the Center for Assessment notes the following:

Most states are developing data tracking systems so they have both the required amounts of testing and the usable data to implement growth systems for school accountability. Thus, now is a good time to consider growth models because there is greater conceptual clarity about how they might be incorporated into school accountability systems and it is becoming more practically possible to do so.⁴⁸

Recognizing the potential of growth models for state accountability plans under NCLB, USED announced a growth model pilot program in November 2005 under which it would approve up to ten proposals. States submitting a proposal and receiving approved plans include Arkansas, Delaware, Florida, North Carolina, and Tennessee.⁴⁹

USED identified the following principles for what it considered high-quality growth models:

⁴⁶ State Education Department, *Teleconference Questions and Answers* (October 20, 2004) <http://www.emsc.nysed.gov/3-8/faq.htm>.

⁴⁷ State Education Department, *State Education Department Unveils New System to Track and Report Student Data, System Will Provide New Tool to Improve Student Achievement*, September 7, 2006. <http://www.emsc.nysed.gov/irts/press-release/20060907/GROW-Reports-Release.htm>.

⁴⁸ Gong, Perie, and Dunn, *Using Student Longitudinal Growth Measures for School Accountability Under No Child Left Behind: An Update To Inform Design Decisions*, p. 3.

⁴⁹ U.S. Department of Education, *Growth Models: Ensuring Grade-Level Proficiency for All Students by 2014*, November 9, 2006. <http://www.ed.gov/admins/lead/account/growthmodel/proficiency.html>.

1. Ensure that all students are proficient by 2014 and set annual goals to ensure that the achievement gap is closing for all groups of students;
2. Set expectations for annual achievement based upon meeting grade-level proficiency, not based on student background or school characteristics;
3. Hold schools accountable for student achievement in reading/language arts and mathematics;
4. Ensure that all students in tested grades are included in the assessment and accountability system, hold schools and districts accountable for the performance of each student subgroup, and include all schools and districts;
5. Include assessments in each of grades three through eight and high school in both reading/language arts and mathematics, must have been operational for more than one year, and must receive approval through the NCLB peer review process for the 2005-06 school year. The assessment system must also produce comparable results from grade to grade and year to year.
6. Track student progress as part of the State data system; and
7. Include student participation rates and student achievement on a separate academic indicator in the state accountability system.⁵⁰

While each of the five states' approved plan is different, each contain the basic component specified by USED of setting individual yearly growth targets that students must meet over a number of years to be classified as "on track to be proficient."⁵¹

Value-Added Models

One type of growth model is the "value-added model," which has the ability to use a student's detailed background information and achievement data to predict growth and then isolates the primary reason for a student's academic progress (or lack thereof). Value-added models may use statistical models to remove non-academic factors in measuring the cause of a student's academic progress, which may include a particular learning program, teacher, school, or other factor. Various value-added models have been developed and are being used by several states, including Colorado, Iowa, Ohio, Pennsylvania, and Tennessee.⁵²

⁵⁰ U.S. Department of Education, *Secretary Spellings Approves Additional Growth Model Pilots for 2006-2007* (November 9, 2006), <http://www.ed.gov/news/pressreleases/2006/11/11092006a.html>.

⁵¹ For details on the various approaches taken by states approved to use growth models, see "State Applications, Decisions Letters and Additional Information", U.S. Department of Education, online at <http://www.ed.gov/admins/lead/account/growthmodel/index.html>.

⁵² For an analysis comparing value-added models, see Pete Goldschmidt, et. al., *Policymakers Guide to Growth Models for School Accountability: How do Accountability Models Differ?*, Council of Chief State School Officers (October 2005), <http://www.ccsso.org/publications/details.cfm?PublicationID=287>.

Tennessee

In 1992, Tennessee implemented the Tennessee Value Added Assessment System (TVAAS). Under this system, students in grades 3-8 are tested annually in five subject areas: reading, math, language, social studies and science. A customized version of the TerraNova exam is used. A complex statistical model then uses students' prior results as a control to predict future performance, measured by a normative national sample and local averages. TVAAS then analyzes actual student performance to expected performance, and the model allows for missing data if students do not take a particular test and is designed to determine the effect of schools, teachers, and other variables on achievement. The data also is used as part of the process for evaluating teachers, but teacher-level data is not publicly released.⁵³

Tennessee's value-added model was developed by Dr. William Sanders, a former professor at the University of Tennessee. The ability to accurately measure individual student progress, to develop expected levels of future performance, and to estimate teacher and school effectiveness based on student performance data made the model appealing to numerous educators and policymakers. In 2000, Dr. Sanders teamed up with a private software company to make the accountability system he developed for Tennessee more widely available.⁵⁴

Colorado, Iowa, Ohio, and Pennsylvania now all have contracts to use the Sanders value-added model.⁵⁵ These states have tinkered with the model to accommodate their state tests and their particular needs. Pennsylvania, for example, does not link teacher data to student performance and therefore does not measure teacher impact on learning.

Value-Added in New York

Since 2003, the New York State School Boards Association has advocated for a value-added accountability model for the state's public schools. After hosting a summit on the topic in 2004, it announced the first phase of a pilot value-added program in 2005 with the Capital Region BOCES to serve 15 school districts.⁵⁶ Phase two of the initiative was launched in 2006,

⁵³ Pete Goldschmidt, et. al., *Policymakers Guide to Growth Models for School Accountability: How do Accountability Models Differ?*, Council of Chief State School Officers (October 2005), p. 16, <http://www.ccsso.org/publications/details.cfm?PublicationID=287> and Evergreen Freedom Foundation, *School Directors' Handbook* (2001), p. VA15-16, <http://www.effwa.org/studies/sdh.php>.

⁵⁴ Evergreen Freedom Foundation, *School Directors' Handbook* (2001), p. VA15-16, <http://www.effwa.org/studies/sdh.php>.

⁵⁵ Lynn Olson, "Value Added Models Gain in Popularity," *Education Week* (November 17, 2004), <http://www.edweek.org/ew/articles/2004/11/17/12value.h24.html> and The Center for Greater Philadelphia, *Value-Added Assessment and Student Progress*, http://www.cgph.upenn.edu/ope_nation.html.

⁵⁶ Barbara Bradley, "NYSSBA, NYSCOSS Host Value-Added Summit," *On Board Online* (August 16, 2004), http://www.nyssba.org/scriptcontent/va_custom/va_cm/eventpage.cfm?article=2374&event=print, and Brian M. Butry, "Value-Added Pilot Project Begins this Fall," *On Board Online*, August 22, 2005, http://www.nyssba.org/scriptcontent/va_custom/va_cm/eventpage.cfm?article=3971&event=print.

which expanded and offered the program to any district in the state.⁵⁷ In 2004, the New York State Commission on Education Reform (also known as the Zarb Commission) issued a report calling for the implementation of a value-added model.⁵⁸

As part of his first executive budget, unveiled in January 2007, New York Governor Eliot Spitzer proposed enhancing the state's accountability system by implementing a growth accountability model within a year and a value-added model within three years.

Specifically, Governor Spitzer's plan calls for the establishment of an interim accountability system by the start of the 2008-09 school year that uses existing state assessments and is based on a growth model of measuring individual students and cohorts over time. The New York State Board of Regents and the Education Commissioner are called upon to establish an enhanced accountability system with revised or new assessments by the 2010-11 school year that "includes a value-added assessment model that employs a scale-score approach to measure growth of students in all levels" and "includes an evaluation of the specific effects of a particular school, program, district or teacher on the academic progress of individual students over time."⁵⁹

New York City

In June 2005, New York City's Chancellor of Education, Joel Klein, announced a plan he called "Gains Analysis," a value-added component to be added to the New York City schools' accountability system. Gains Analysis, implemented in 2006, tracks individual student achievement in English language arts and math assessments over time. The system allows for data from similar schools to be compared and allows administrators to identify the gains in student learning being caused by individual schools.⁶⁰ The value-added measurement system racks data for all of New York City's 1.1 million students in 1,400 schools.

In March 2007, New York City announced an \$80 million, 5-year deal with IBM to further implement its value-added accountability system. The new system, called the Achievement Reporting and Innovation System, will combine existing demographic and school-based data on students with data from annual state exams and new data from diagnostic exams to be administered every six weeks or so that measure whether students have mastered specific skills. The planned accessibility to data and transparency of results is unprecedented:

⁵⁷ Brian M. Butry, "Interest Rises in BOCES Value-Added Project," *On Board Online*, August 21, 2006. http://www.nyssba.org/scriptcontent/va_custom/va_cm/eventpage.cfm?article=5547&event=print.

⁵⁸ New York State Commission on Education Reform, *Ensuring Children an Opportunity for a Sound Basic Education*, March 29, 2004. <http://education.cornell.edu/rsp/ZarbFinal.pdf>.

⁵⁹ *2007-08 Executive Budget* (S.2107/A4307), § 319, January 31, 2007, p. 4-5. <http://assembly.state.ny.us/leg/?bn=S02107&sh=t>.

⁶⁰ New York City Department of Education, *City Schools*, June 2005, <http://schools.nyc.gov/Administration/mediarelations/Newsletters/june2005.htm#Article2>; and

Teachers will be able to see an entire classroom of results at once. Principals will be able to see an entire school. Parents eventually will have access to their own kids' data plus summary facts about their child's school, the results of parent, student and teacher surveys and details about how their school scored on annual reviews. Much of the data will be folded into letter grades that soon will be assigned to all 1,400 city school.⁶¹

The apparent trend of New York to enact growth and value-added accountability systems is promising. Care should be taken to ensure that such a model effectively measures all necessary data about student learning and incorporates provisions that ensure the adults in the education system – teachers, building administrators, district bureaucrats – are properly held accountable for that learning.

⁶¹ Erin Einhorn, "Big Brother Is Looming Over City's Schools," New York Daily News, March 6, 2007.

TIMELY FEEDBACK

New York's process for developing state exams is a meticulous one, involving testing and education practitioners in the development of questions, peer review of the draft, and field-testing of the exams. After painstakingly developing detailed learning standards and matching curricula in various subject areas, the State Education Department also strives to ensure that statewide exams are aligned with these learning standards and accurately measure whether New York students are academically meeting those standards. Unfortunately, the state has undermined the value of this effort by delaying the public release of test-result data to the point where there is little that can be done with it.

Making the Tests

New York State undertakes a lengthy and detailed process to develop new state assessments. Committees comprised of teachers, State Education Department (SED) staff, and representatives from the contracted test-development company are formed and charged with first producing guidelines on the format and types of test questions to be used, and then developing questions that assess content contained in the state standards. The test-development company then reviews and edits the questions developed by the committees, and sends them back to different review panels made up of teachers and SED staff. The review panels further examine and revise the test questions for alignment to state standards, appropriateness for grade level and content, and clarity and format of questions. SED, in concert with test-development company, selects the sets of questions to be used for field testing.

Selected questions are then field-tested at a sample of schools statewide, and students' answers provided are reviewed by committees of teachers, SED staff, and testing company representatives to form scoring guides. Field-test results also are used to help analyze and ensure the validity and reliability of the questions used. A second field test then is administered at selected schools throughout the state.

Approximately one year before the first administration of an assessment, sample draft tests and sample student answers, based on the results of SED's field-testing, are sent to all schools as an example of what the test will be like. The Commissioner of Education gives his final approval of the exam, and the assessment is then ready to be administered statewide.⁶²

Taking and Scoring Tests

Schools are shipped state assessments a week before the exam date and are required to take inventory to confirm that all required materials have been received, to store them in a locked

⁶² State Education Department, *New York State Testing Program: Introduction to the Grades 3-8 Testing Program in English Language Arts and Mathematics* (October 2005), <http://www.emsc.nysed.gov/3-8/intro.pdf>

safe or vault, and to check daily to ensure that the exams have not been tampered with. Detailed instructions are provided to schools that outline the conditions for administering the exams.

Following the administration of an exam, schools have the option of scoring written responses in-house, in coordination with other schools across the district, or across the region. Detailed guidelines are provided to test scorers in an attempt to ensure that exams are graded fairly and according to a uniform process. After the tests have been scored and checked, answer sheets are sent to regional scanning centers.

SED conducts an annual audit of the scoring of exams as a check on the integrity of the scoring process. This audit includes approximately 10 percent of the state's schools, all randomly selected.⁶³

Reporting Test Results: Too Long To Wait

The delay in publicly reporting results from New York State exams is the longest in the nation.⁶⁴

In school year 2005-06, the first year in which NCLB-required assessments in English language arts and mathematics were implemented for grades 3 through 8, New York's turnaround time for scoring and reporting results was eight months, and spanned the gap between two school years. The state English language arts assessment was administered to students in January 2006, but SED didn't make student-level results available online to schools until mid-September. For mathematics, the assessment was administered in March 2006, and results weren't posted until November.

An analysis by the *New York Sun* noted:

New York State has the worst record in the nation in terms of the length of the gap between the time that children take the tests and the point at which the scores become available is indeed unique. It puts New York State at the bottom of all of the 50 states in getting results back to teachers and parents...

The experience of other large states shows just how far behind the curve New York is. California administers its tests in March and May, yet is able to return all results in May and June. Florida tests in June, and results are ready before the opening of school in late August. Texas tests children early in the school year, in October, and provides results by December, in time to use the information to help shape the instructional program for each child for the remainder of the school year.

⁶³ State Education Department, *New York State Testing Program: Grades 3-8 English Language Arts Tests, School Administrator's Manual for Public Schools*, 2007. <http://www.emsc.nysed.gov/3-8/sam/ela07p.pdf>.

⁶⁴ Andrew Wolf, "An Uproar Erupts at State's Delay in Grading Exams," *New York Sun* (June 20, 2006), <http://www.nysun.com/article/34697>.

Nearby states also do better. New Jersey administers tests in March and April and delivers results in June, as does Pennsylvania. Connecticut tests in March and delivers results in July. The Commonwealth of Virginia has the results of its exams just three weeks after they are administered in May and June.⁶⁵

This 8-month turnaround time – a gap four times the national average – rendered the data largely useless for the purposes of identifying students in need of intervention, making decisions about the retention of students or initiation of summer school, revising the curriculum for the upcoming school year to address deficiencies, offering needed staff development to teachers whose classes were populated with a disproportionate number of low-scoring students, and more.

The State Education Department claimed that New York’s lengthy delay was due to “norming.” Under this process, the state tallies the results of all exams and then determines a cut score above which students’ performance is considered to show proficiency.

Robert Tobias, a New York University professor, indicated that this extensive delay could have been reduced if SED had completed some of the required norming processes prior to the exams being administered and if the process had been scheduled with a greater sense of urgency.⁶⁶

Promises of Improvement

For the 2006-07 school year, the State Education Department claims it will cut the time it takes to make test results public about in half: results for the English Language arts exam, administered in January 2007, will be posted on May 22; results for the math exam, administered in March, are scheduled to be made public on June 14.⁶⁷ While a predicted improvement – one that has yet to be verified – these delays are still twice the national average.

⁶⁵ “An Uproar Erupts at State’s Delay in Grading Exams,” *New York Sun*, June 20, 2006.

⁶⁶ “An Uproar Erupts at State’s Delay in Grading Exams,” *New York Sun*, June 20, 2006

⁶⁷ State Education Department, *Important Dates for Grades 3-8 Tests in 2006-07, English Language Arts Tests* (January 24, 2007), <http://www.emsc.nysed.gov/3-8/sam/eladates-07.htm> and State Education Department, *Important Dates for Grades 3-8 Tests in 2006-07, Mathematics Tests* (February 21, 2007), <http://www.emsc.nysed.gov/3-8/sam/mathdates-07.htm>.

HOW VALID ARE THE RESULTS?

When state accountability systems include high-stakes consequences – for students, schools, teachers, administrators, etc. – the incentive to cheat increases. One recent news article noted, “...with the implementation of the *No Child Left Behind Act*, the incentives for teachers and administrators to manipulate the results from high-stakes tests will only grow, especially as schools begin to feel the consequences of low scores.”⁶⁸

Ensuring that the processes employed by New York State for shipping, storing, administering, scoring, and reporting state assessments are valid is an important and integral part of maintaining a useful and valuable accountability system.

Like many other states, New York State’s current assessment system is plagued by conflicts of interest, opportunities for dishonesty, and instances where the quality of the testing system could be improved substantially. Some of these issues include the following:

- The State Education Department (SED) assembles committees of teachers each year that develop rubrics for scoring the statewide exams. These rubrics identify the quality of responses expected to earn a particular score on each non-multiple-choice question. The membership of these committees varies every year, however, creating the potential for inconsistent rubrics with varying degrees of difficulty from one year to the next. A committee that includes a larger portion of teachers from high-performing schools, for example, may develop a more difficult scoring rubric than one made up of teachers from traditionally low-performing schools because its members have different expectations for what students should be able to do.

Additionally, there are no specified qualifications for teachers who are chosen to serve on the committees. A committee working on the 7th grade English language arts assessment, for example, is not necessarily comprised of 7th grade English teachers that have proven their expertise in the subject. Teachers eligible to develop scoring rubrics ought to at least be required to teach the grade and subject of the assessment and have a track record of excellence.

- New York allows schools and teachers to score the non-multiple choice portion of their own students’ exams. A website essay by a Texas elementary school teacher noted: “Law schools do not give the bar exam, nor do driving schools give drivers’ license exams. And imagine the public reaction if we decided to eliminate track meets and award regional and state championships according to times and distances submitted by coaches.”⁶⁹ This is the same as New York’s public schools grading their own assessments and submitting the results to the state, a clear conflict of interest that could easily be prohibited.

⁶⁸ Brian A. Jacob and Steven D. Levitt, “To Catch a Cheat,” *Education Next* (Winter 2004), p. 70, <http://www.hoover.org/publications/ednext/3345096.html>.

⁶⁹ Jerry Jesness, “A Test on Test Security,” <http://speakout.com/activism/opinions/2856-1.html>.

- The grading of essay portions of assessments can be very subjective, even despite the fact that scoring rubrics are distributed by SED. The potential exists for teachers to “read into” the written answers of some of their students’ essays and search for a couple of extra points, particularly if a student is on the borderline of passing. Instead, a system could be put in place that scans essay answers and sends them electronically to a centralized location which is staffed by trained scorers. This would create more consistency in the scoring of student assessments, relieve teachers entirely from having to score assessments, and potentially increase the turnaround time for the reporting of scores.
- Large portions of the state assessments currently administered in New York are multiple-choice, with students filling-in the “bubbles” on answer sheets. These bubble-sheets are scored electronically at regional centers. In order to limit the amount of time that completed answer sheets are stored at the school and could potentially be illegally altered, schools should be required to deliver them directly to the scoring centers the day the exams are administered. Currently, schools and districts are allowed two weeks before these answer sheets are due to the scanning center.
- During the proctoring of exams, New York State guidelines say that teachers are not permitted to provide assistance to students beyond enabling them to understand the procedures for taking the test. Teachers are prohibited from assisting students to understand questions or to coach students by providing clues to the answers or leading questions that reposition a students reasoning.

Still, research shows that teachers who administer exams to their own students are about 50 percent more likely to cheat.⁷⁰ Simply prohibiting altogether any teacher from proctoring any exam for his or her own students or subjects would reduce the likelihood of cheating. Mandating a minimum of two proctors in each testing room, thus creating multiple witnesses to the administration of exams, also would have a chilling effect on cheating.

- Fear of retribution, losing one’s job, and negative portrayal of the teaching profession from media attention to a cheating scandal likely serve as disincentives for teachers and administrators to report cheating on tests by others in their building. Over a six-year period (1998-2003), 21 teachers were tagged for cheating on state assessments in New York;⁷¹ it is unclear how large or small a portion of the entire population of cheaters this truly is. The State Education Department (SED) could strengthen whistle-blower protections, launch new education efforts and hot-line reporting initiatives, and undertake comprehensive investigation practices.

Currently, New York law classifies cheating on state assessments as a misdemeanor. Punishment for guilty teachers and school administrators, however, is left to local school districts to be determined.⁷² According to SED, punishments have ranged from simple

⁷⁰ Jacob and Levitt, “To Catch a Cheat,” <http://www.hoover.org/publications/ednext/3345096.html>.

⁷¹ Michael Gormley, “Punish Cheating Teachers, Education Advocates Urge,” Associated Press, (Nov. 10, 2003).

admonishments to termination. New York has yet to strictly enforce this law and charge all teachers and administrators found guilty of cheating with a misdemeanor, although reminders have gone out of superintendents' obligation to ensure the security of state tests and to report suspected cheating.

As could be expected, the unions representing teachers has reacted strongly to calls for increased efforts to identify and punish test-scorers who cheat. "I don't think a teacher should call the state and say so-and-so cheated," said Dennis Tomkins of the New York State United Teachers. Tomkins also told a newspaper reporter that he did not see the benefit of identifying cheaters.⁷³

While state records show that the State Education Department, some school administrators, and school boards acted swiftly to pursue reports of cheating – including coaching students during tests, sending students back to their desks to correct wrong answers, giving answers at the head of the class, and writing answers into students' tests – the documents also note that "some teachers and a local union official refused to report incidents and colleagues to administrators."⁷⁴

⁷² New York Education Law, § 225.

⁷³ Michael Gormley, "Schools Asked to Nab Test Cheats," Times Union, November 10, 2003.

⁷⁴ Michael Gormley, "Schools Asked to Nab Test Cheats," Times Union, November 10, 2003.

THE SPITZER REFORM PROPOSALS

In a speech delivered at Chancellor's Hall in the historic New York State Education Department building in Albany, New York Governor Eliot Spitzer laid out an ambitious education reform plan less than one month after taking office. The plan became a key component of his first state budget proposal offered less than a week later. Gov. Spitzer proclaimed:

[F]or New York to become the economic engine it once was, to create jobs, opportunity and prosperity, we must change the way we educate our children.... My vision for education reform is built on a single premise: to be effective, new funding must be tied to a comprehensive agenda of reform and accountability.... [T]here will be no more excuses for failure. The debate will no longer be about money, but about performance; the goal will no longer be adequacy but excellence; and the timetable will no longer be tomorrow but today.⁷⁵

The fundamental components of Gov. Spitzer's proposal included the following:

- **Conditions on Funding Increases.** After a reconstitution of the state education funding formula, districts that are slated to receive an annual increase of 10 percent or \$15 million would be required to develop a "Contract for Excellence" that indicates how they will spend the new funding on measures that, according to the Governor, "have been demonstrated to improve student performance," including:
 - class size reduction;
 - longer school days and school years;
 - after-school education programs; and/or
 - restructuring of the school day.

The State would establish "a menu of approved strategies and initiatives that have been demonstrated to improve learning outcomes based on hard data and professional research" from which districts could choose.

- **Teacher Quality Initiatives.** Alternative certification routes for teachers would be expanded, and the effectiveness of current teacher education programs would be measured. Importantly, the granting of tenure to teachers would be revised from a process based simply on the number of years an individual holds a teaching job to a process that involves "the review of the supervisor, an evaluation by professional colleagues, and an examination of data as well as qualitative information about how a teacher's students perform over multiple years."
- **Defined Goals of Academic Progress.** Under these "Contracts for Excellence," districts would be required to set clear and discrete goals for academic success from the chosen

⁷⁵ See http://www.ny.gov/governor/keydocs/0129071_speech.html for the complete text of Governor Spitzer's speech, which was delivered on January 26, 2007.

changes. According to Gov. Spitzer, “That means telling parents, as well as the State, how many more children will read and do math at grade level, how many more students will graduate from high school with Regents diplomas, and how many of them will go on to college.” These plans would sunset after three or four years, allowing a genuine analysis of the effect of any changes.

- **Management Overhaul for Continued Failure.** The state would require districts to sign contracts with superintendents that include a performance clause under which the superintendent would be dismissed after four years of substantially deficient performance. After further failure of the district, the Commissioner of Education would seek the removal of the districts’ boards of education, too. School building principals and district superintendents also would be the subject of “report cards” that offered comparative data on the academic progress or lack thereof of the schools within their scope of responsibility, including performance on state assessments and graduation rates.
- **Close Bad Schools.** The Commissioner of Education would be required to develop regulations that lead to the closure of persistently bad schools, “perhaps as many as five percent of all the schools in the state.”
- **Praise for Success.** State funds would be available for rewarding the faculty at schools that show significant academic progress, and for holding up their schools as models of success.
- **Revamped Accountability System.** By July 2008, the New York Board of Regents would establish improvement targets for individual schools and school districts using measures that include state assessments and graduation rates. By that same time, individual student progress report would be available that give parents information on their children’s educational progress over multiple years of testing. School districts also would be required to report and publish basic data on school-by-school resource allocations. By July 2010, the Regents would establish an accountability system through which the educational attainment of all children can be measured based upon cumulative achievement and growth.

Governor Spitzer summarized his rationale for proposing such a radical departure from the traditional “spend lots more, expect no more” approach practiced in New York for years toward enhanced accountability this way:

[W]e need performance accountability, because unless we have meaningful consequences for good and bad performance, we will never be able to change the status quo that is failing too many of our children... Money can no longer be an excuse for failing our children anywhere in the state. Now, if children fail, adults must be held accountable. And accountability means consequences, both good and bad, for the performance of schools and school districts.⁷⁶

⁷⁶ http://www.ny.gov/governor/keydocs/0129071_speech.html.

TRUE ACCOUNTABILITY: REAL INCENTIVES, REAL CONSEQUENCES

Chester E. Finn, president of the Thomas B. Fordham Foundation and the John M. Olin Fellow at the Manhattan Institute, advised that policymakers seeking to map a “clear path to a sound accountability system” should first establish some basic key criteria, including applying the following three questions:

1. Which accountability strategy focuses most directly on academic achievement?
2. Which is the most apt to work effectively (i.e., to produce the desired results)?
and,
3. Which is the most amenable to implementation?⁷⁷

Education economist Eric Hanushek adds a further test:

[T]hough accountability systems create direct incentives for students, they produce only indirect ones for schools and teachers. Accountability systems work well only if they provide a direct link between outcomes and the behavior of each person in question.⁷⁸

Adopting the Right Measurement Systems

Governor Eliot Spitzer is right to call for a transformation in the type of accountability model New York uses to judge the performance of its public schools. Gains students achieve over time, measured through a growth or value-added approach rather than the existing “status” or snapshot model, are better indicators of the success or failure of a given district, school, or teacher. The demographics and identity of students change in each grade from year to year, and measuring growth on an individual basis or within same-student cohorts is the only reliable way to structure an accountability model.

New York City should be praised for the recent advancement of efforts in this regard, and New York State would be wise to look to Tennessee and other states that have revamped their accountability processes to incorporate value-added measurements and calculations of growth in academic achievement.

Setting the Right Standards for Achievement

New York’s education policymakers need to ensure that standards for achievement on accountability assessments are set sufficiently high so that passing a mandated state test truly means something.

⁷⁷ Finn, Chester E., Jr., “Real Accountability in K-12 Education,” in *School Accountability* (Stanford, CA: Hoover Institution Press, 2002), Evers, Williamson M., and Walberg, Herbert J., ed., p.30.

⁷⁸ Hanushek and Raymond, “Sorting Out Accountability Systems,” in *School Accountability*, p.93.

At the elementary and middle-school levels, passing a state English Language Arts test should mean that students can read and write proficiently. Passing a high-school Regents exam should mean that students possess and can demonstrate knowledge that is sufficient as a basis for either higher, collegiate-level learning or professional application in the work-world.

Currently in New York, many believe this is not the case. According to the Fordham Foundation:

[T]he cut score for proficiency [is] 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.⁷⁹

An uproar occurred when analysts pointed out that the required passing score of 55 on Regents exams really wasn’t equivalent to getting 55 percent of the answers correct, as many believed. A recent article in the *New York Post* noted:

When does 23 equal 55? When it comes to grading the Math A Regents exam. Never before has the state required students to answer so few questions correctly on the mandatory test to eke out a passing grade of 55...than it has this year. Students who sat for the exam...needed just 23 out of 84 points, or about 27 percent – to earn a 55. Only last year, a score of 26 – or 31 percent – was required to reach the benchmark.⁸⁰

“It’s fooling the kids and it’s fooling the parents,” summarized one a math teacher who wished to remain unidentified in the story.

Dumbed-down achievement goals do not serve anyone well.

Holding Everyone Accountable

New York’s accountability system must be designed to measure and affect every element of the educational system. Hanushek notes:

The incentives that derive from the design and use of accountability systems work only to the extent that they motivate students, teachers, and schools to examine their performance and make changes to improve, if necessary. Without consequences, incentives disappear.... Relying solely on rewards may not be sufficient to overcome the inertia of habit.⁸¹

⁷⁹ *Grading the Systems: The Guide to State Standards, Tests, and Accountability Policies*, Cross, Richard W., Rebarber, Theodor, and Torres, Justin, ed. (Thomas B. Fordham Foundation: Washington D.C., January 2004), p.26.

⁸⁰ David Andreatta, “‘Sickening’ E-ZPass Regents Math Exam,” *New York Post*, January 30, 2006.

⁸¹ Hanushek and Raymond, “Sorting Out Accountability Systems,” in *School Accountability*, p.99.

Data must be collected, analyzed, and presented in such a manner that classroom instructors can use the information about student performance to make improvements in the way they teach, principals can hold their teachers accountable for classroom performance, superintendents can hold principals accountable for the management of schools, boards of education can hold superintendents accountable for the performance of their school districts, policymakers can hold school boards accountable for their approach to education, and parents can hold everyone accountable for each stage of their children's education.

Responding to criticism by some of new state testing requirements, then Deputy Commissioner of Education James Kadamus supported the position that state tests had a broader role than simply passing judgment on a school's performance. "This is to change instruction," Kademus said, stating that state tests specifically include components to test "subskills," the results from which were expected to be used directly to tailor each grade's curriculum.⁸² That is, schools are expected to *do something* based on the data provided through the state's testing programs.

Individual student data must be collected and compiled so that progress over time can be measured and same-student cohort analysis can be meaningfully performed. The data also must be traceable to individual teachers for accountability at that level.

Governor Spitzer's proposed innovative approach to require performance clauses in superintendents' contracts is an appropriate response to the goal of holding all players responsible. Given the transient nature of superintendents in office, however, it is unclear whether imposing sanctions only after four years of continuous low performance will have the desired effect. Consider the following:

The Institute of Educational Leadership has portrayed the urban superintendence as a merry-go-round with an average tenure of less than three years...while other studies indicate that the average tenure of superintendents is at least five years.⁸³

If long-serving superintendents truly are a rarity, performance clauses that kick-in after an individual's typical tenure draws to a close will have little impact. Shorter periods before performance clauses apply, financial-based performance clauses, and other innovations ought to be explored.

⁸² Susan Saulny, "Delayed Test Results Trouble New York," *New York Times*, August 12, 2005. http://www.nytimes.com/2005/08/12/nyregion/12school.html?ex=1281499200&en=00a0db274edfad3c&ei=5090&pa_rter=rssuserland&emc=rss.

⁸³ Larry Lashway, "The Superintendent in an Age of Accountability," ERIC Digest 161 – September 2002, Clearinghouse on Educational Policy and Management, University of Oregon College of Education, <http://eric.uoregon.edu/publications/digests/digest161.html>, March 2, 2007.

Making Accountability More Transparent and More Useful

New York's current approach of sorting student performance into the generic groupings of Level 1, Level 2, Level 3, and Level 4 clouds rather than clarifies the accountability picture. The letter-grade approach for labeling schools, already used in Florida statewide and now being started in New York City, is a better approach. A meaningful accountability system must be understandable to the public at large in order to work

Performance data also must be made public in a timely fashion if it is have any real value. If the state's accountability system exposes problems, stakeholders should be provided with reasonable opportunities to fix them. That means providing test results data to principals and teachers within the same school year in which tests are administered. Last year, the State Education Department released results of state exams in reading and math for school year 2005-06 in the following school year. This delay was longer than in any other state, and is inexcusable.

Strengthening the Validity of the Testing & Reporting Process

New York State's current approach of allowing district schools to grade their own exams creates unnecessary conflicts of interest. The state should prohibit districts from scoring their own exams, and should consider contracting out the scoring of non-multiple choice questions.

NCLB Reforms

New York can enact reforms to the way it currently administers its *No Child Left Behind* compliance programs to enhance accountability and support increased student achievement. Included among such reforms could include the following:

- Requiring failing schools to use the entire amount of Title I funds available for supplemental education services (SES) on such services. Currently, schools are required to spend a portion of the federal Title I funds only for students who request SES. School-provided information explaining the SES option is insufficient at best, often uninspired and designed to discourage participation, and nonexistent at worst. Because schools may retain for other uses any Title I funds that remain after paying for SES, schools have a disincentive to effectively market the availability of these services. If schools are forced to use all available Title I funds for SES or forfeit them, the disincentive would disappear. More students receiving more remediation, tutoring, and other supplemental education services would likely lead to greater student academic achievement, especially among currently low-performing students.
- Waiving collective bargaining work rules that prevent longer school days, longer school years, and other structural changes in failing public schools.

- NCLB requires all teachers to be “highly qualified.” While requirements vary depending on when a teacher enters the field and what grade level and subject is taught, generally in New York teachers must simply meet routine state certification requirements, including earning an education degree from an accredited institution and passing the appropriate state teaching exams.⁸⁴ The intention of this section of NCLB is clear: to ensure that 100 percent of classes are led by highly skilled professionals with sufficient knowledge in the subject area being taught. Unfortunately, the simple use of state certification and licensing tests as proxies for subject competency is not valid by itself. Requiring each classroom instructor to take and pass any state exam required of the students they teach would be a better measurement of the basic premise that classroom teachers are, at a minimum, proficient in the subjects they are teaching. Elementary-level teachers could be required to score in the top level, Level 4, on the state English language arts and math exams, and high school teachers could be required to pass the relevant Regents exams with, say, a score of 85 or higher.

Market-Based Solutions

While standards-based accountability systems are pretty good at identifying problems, such as failing schools, they lack the ability to do much about fixing those problems. There is a critical need to ensure that part of New York’s school accountability system includes real remedies for identified problems.

Some of the most effective methods to address deficiencies in the educational system are market-based solutions. Such reforms also are among the quickest. Too often, students linger for years in underperforming schools that have not been forced to change in any significant, measurable way, and new generations of students enroll in these same schools following in the same footprints of failure.

Approaches such as using publicly funded school-choice vouchers for students trapped in failing schools, education tax credits to empower parents with expanded choices of education providers, and establishing corporate scholarship programs through tax or other incentives that can underwrite a students educational expenses all are some of the more radical solutions that can and should be explored to address the most serious problems in public education.

* * * * *

⁸⁴ Further details about New York’s requirements for “highly qualified teachers” are available online at: www.highered.nysed.gov/pdf/nclb052006.pdf

The Fordham Foundation's Chester Finn sounds a note of caution to any policymaker attempting to establish, reform, or maintain a good accountability system:

...the central dilemma of school accountability [is that] *none* of the approaches is idiot-proof. None is immune to bad ideas, distorted priorities, inept management, and old-fashioned laziness. Every one of them hinges on the sagacity, competence, integrity, and determination of those running it – no matter whether that's a governor or a parent. ...[T]he prospect of success is brightest in the...intersection of standards based, top-down accountability and market-style, bottom-up accountability.... [I]t's superior to the available alternatives and worth trying to perfect.⁸⁵

⁸⁵ Finn, Chester E., "Real Accountability in K-12 Education," in *School Accountability*, p.39.