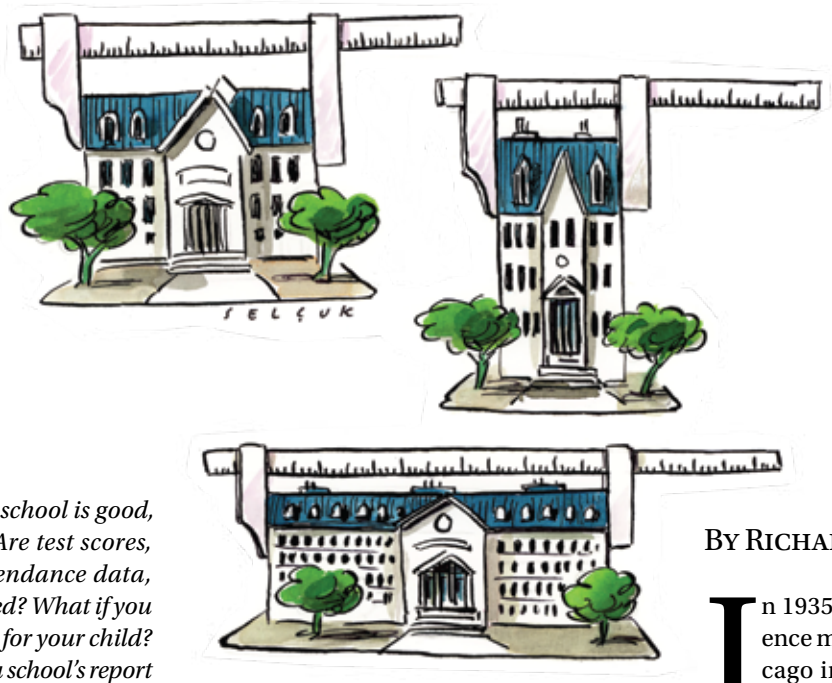
What's Wrong with Accountability by the Numbers?



ILLUSTRATED BY SELÇUK DEMIREL

How do you know if a school is good, bad, or in-between? Are test scores, graduation rates, attendance data, and the like all you need? What if you were selecting a school for your child? Would you just review a school's report card online, or would you schedule a visit so that you could get to know the principal, observe a few classes, and even interview some students? Would you contact some parents, check out the neighborhood, and look for nearby after-school activities? We hope that you would both pay attention to the data and *pay a visit to the school. And so we wonder: why would our education accountability system do anything less?*

In this article, Richard Rothstein explores the well-established problems—in education, health care, and other fields—with accountability systems that focus exclusively on quantitative data. Then, in the article that follows (see page 24), Rothstein and his colleagues, Rebecca Jacobsen and Tamara Wilder, propose a completely new approach to accountability that's inspired in part by England's system of inspecting schools and calls for a national assessment of a full range of cognitive and noncognitive skills.

Did they get it right? That's for you to decide. Their goal is to start a conversation. Since dissatisfaction with our current accountability system is widespread, it's time to ask: what are our goals for education and how can we help all schools meet them?
–EDITORS

Richard Rothstein is a research associate at the Economic Policy Institute, former national education columnist with the New York Times, *and author of several books, including* Class and Schools: Using Social, Economic, and Educational Reform to Close the Black-White Achievement Gap. *This article is adapted with permission from* Grading Education: Getting Accountability Right, *coauthored with Rebecca Jacobsen and Tamara Wilder, published in 2008 by the Economic Policy Institute and Teachers College Press.*

By Richard Rothstein

In 1935, a 19-year-old political science major at the University of Chicago interviewed Milwaukee city administrators for a term paper. He was puzzled that, when money became available to invest in parks, school board and public works officials could not agree on whether to hire more playground supervisors or improve physical maintenance of the parks themselves. He concluded that rational decision making was impossible because "improving parks" included multiple goals: school board members thought mostly of recreational opportunities for children, while public works administrators thought mostly of green space to reduce urban density.

The next year, the director of the International City Managers' Association hired the young graduate as a research assistant. Together they reviewed techniques for evaluating municipal services, including police, fire, public health, education, libraries, parks, and public works. Their 1938 book, *Measuring Municipal Activities*, concluded that quantitative measures of performance were mostly inappropriate because public services have goals that can't easily be defined in simple numerical terms. Public services have multiple purposes and, even if precise definitions for some purposes were possible, evaluating the services overall would require difficult judgments about which purposes were relatively more important. Also, it was never possible to quantify whether outcome differences between cities were attributable to differences in effort and competence of public employees, or to differences in the conditions—difficult to measure in any event—under which agencies worked.

The senior author, Clarence E. Ridley, directed the City Man-

agers' Association until retiring in 1956. His assistant, Herbert A. Simon, went on to win the Nobel Prize in economics for a lifetime of work demonstrating that weighing measurable costs and benefits in simple numerical terms does "not even remotely describe the processes that human beings use for making decisions in complex situations."[1]

The past few decades have seen growing enthusiasm among politicians and policymakers for quantitative accountability systems that might maximize public service efficiency. But they have rushed to develop measurement systems without giving great thought to issues that Ridley and Simon raised 70 years ago.

In Great Britain a quarter century ago, Margaret Thatcher attempted to rationalize public enterprises: where they could not be privatized, her government hoped to regulate them, using rewards and sanctions for numerically specified outcomes. Tony Blair later accelerated these efforts, while in the United States, the Clinton administration's Government Performance Results Act of 1993 proposed to "reinvent government" by requiring measurable outcomes for all government agencies.

Enthusiasm for holding schools accountable for student test scores is but part of this broader trend that has proceeded oblivious to the warnings of Herbert Simon and other notable social scientists. Scholars have often concluded that, when agents in other sectors are held accountable for improving production of a simple numerical output, performance on that easily measured output does improve. *But overall performance frequently deteriorates*. So economists, sociologists, and management theorists generally caution against accountability systems that rely exclusively, or even primarily, on numerical outcome measures.

In 1975, social scientist Donald T. Campbell formulated what he called his "law" of performance measurement:

> The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.[2]

Such corruption occurs primarily because of the problem Herbert Simon identified—an indicator that can be quantified often reflects only an aspect of the outcome of interest, so undue attention to this aspect will distort the balance of services being provided.

Examples of Campbell's law abound. Motorists stopped by police for trivial traffic violations may have experienced an accountability system in which police sergeants evaluate officers by whether they meet ticket quotas. Certainly, issuing citations for traffic violations is one measure of good policing, but when officers are disproportionately judged by this easily quantifiable outcome, they have incentives to focus on trivial offenses that meet a quota, rather than investigating more serious crimes where the payoff may be less certain. The numerical accountability system generates false arrests, and creates incentives for police officers to boost their measured productivity by disregarding suspects' rights. In New York City a few years ago, the use of quantifiable indicators to measure police productivity resulted in the publicized (and embarrassing, to the police) arrest of an 80-year-old man for feeding pigeons and of a pregnant woman for sitting down to rest on a subway stairway.[3]

The annual rankings of colleges by *U.S. News and World Report* offer another example of Campbell's law. The rankings are truly an accountability system; many colleges' boards of trustees consider the rankings when determining presidential compensation. In at least one case, a university president (at Arizona State) was offered a large bonus if the university's ranking moved up on his watch.[4]

U.S. News rankings are based on several factors, including the judgments of college presidents and other administrators about the quality of their peer institutions, and the selectiveness of a college, determined partly by the percentage of applicants who are admitted (a more selective college admits a smaller percentage of applicants). Thus, the rankings are a candidate for illustration of Campbell's law, because these factors would be quite reasonable if there were no stakes attached to measuring them. College presidents and other administrators are in the best position to know the strengths and weaknesses of institutions similar to their own, and asking them for their opinions about this should be a good way to find out about college quality. But once an accountability rating is based on these answers, presidents have incentives to dissemble by giving competing institutions poorer ratings and making their own institutions appear relatively superior.

Enthusiasm for holding schools accountable for student **test scores** is part of a broader trend that has proceeded oblivious to the warnings of notable social scientists.

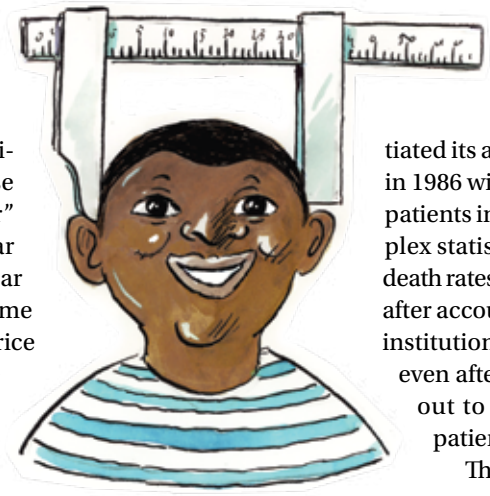
Likewise, the selectiveness of a college was once a reasonable factor to consider, since higher-quality colleges are likely to accept relatively fewer applicants because demand for admission is strong. But once this indicator became an accountability measure, colleges had an incentive to recruit applicants who were bound ultimately to be rejected. Colleges, for example, have sent promotional mailings to unqualified applicants and waived application fees in order to attract unsuccessful (and unsuspecting) applicants. The indicator nonetheless persists in the *U.S. News* rankings, although it now has questionable value.[5]

As a 1968 presidential candidate, Richard M. Nixon promised a "war" on crime. After his election, the FBI publicly reported crime statistics by city. It judged whether police departments were effective by the sum of crimes in seven categories: murder, forcible rape, robbery, aggravated assault, burglary, auto theft, and serious larceny (defined as theft resulting in a loss of at least $50). Many cities subsequently posted significant reductions in crime.[6] But the crime reductions were apparently realized by playing with crime classifications. The biggest reductions were in larcenies of $50 or more in value. Valuing larceny is a matter

of judgment, so police departments placed lower values on reported losses after the implementation of the accountability system.[7] Although the number of alleged $50 larcenies (which counted for accountability purposes) declined, the number of alleged $49 larcenies (which did not count) increased.

More Sophisticated Measures Help, But Not Enough

Probably the most obvious solution to the goal distortion that results from blunt measures is to create more sophisticated measures. But even carefully constructed quantitative measures fall short. In education, test-based accountability systems should (though often do not) adjust results for differences in student characteristics. A school with large numbers of low-income children, high residential mobility, great family stress, little literacy support at home, and serious health problems may be a better school, even if its test scores are lower, than another whose pupils don't have such challenges. Education policymakers sometimes try to adjust for these differences by comparing only "similar" schools—those, for example, with similar proportions of minority students, or similar proportions of students who are low income (eligible for the federal free and reduced-price lunch program). Such adjustments are worth making, but they don't really solve the problem. Stable working-class families, with incomes nearly double the poverty line, are eligible for the federal lunch program; schools with such students can easily get higher scores than schools with very poor students, yet the latter schools may be more effective.



Medicine faces similar problems; some patients are much sicker, and thus harder to cure, than others with the same disease. Patients' ages, other diseases, history of prior treatment, health habits (smoking, for example), diet, and home environment must all be taken into account. So before comparing outcome data, health care report cards must be "risk-adjusted" for the initial conditions of patients. Although risk adjustment in medicine is far more sophisticated than controls in education for minority status or lunch eligibility, health policy experts still consider the greatest flaw in medical accountability systems to be their inability to adjust performance comparisons adequately for patient characteristics.

For example, the Health Care Financing Administration (HCFA) initiated its accountability system for cardiac surgery in 1986 with its reports on death rates of Medicare patients in 5,500 U.S. hospitals. HCFA used a complex statistical model to identify hospitals whose death rates after surgery were greater than expected, after accounting for patient characteristics. Yet the institution labeled as having the worst death rate, even after sophisticated risk-adjustment, turned out to be a hospice caring for terminally ill patients.[8]

The following year, HCFA added even more

What Really Happens in the Private Sector?

When New York City Mayor Michael Bloomberg announced a 2007 teachers' union agreement to pay cash bonuses to teachers at schools where test scores increase, he said, "In the private sector, cash incentives are proven motivators for producing results. The most successful employees work harder, and everyone else tries to figure out how they can improve as well."[1] Eli Broad, whose foundation promotes incentive pay plans for teachers, added, "Virtually every other industry compensates employees based on how well they perform.... We know from experience across other industries and sectors that linking performance and pay is a powerful incentive."[2]

These claims misrepresent how private sector firms motivate employees. Although incentive pay systems are commonplace, they are almost never based exclusively or even primarily on quantitative output measurement for professionals. Indeed, while the share of private sector workers who get performance pay has been increasing, the share of workers who get such pay based on numerical output measures has been decreasing.[3] The business management literature nowadays is filled with warnings about incentives that rely heavily on quantitative rather than qualitative measures.

Because of the ease with which most employees game purely quantitative incentives, most private sector accountability systems blend quantitative and qualitative measures, with most emphasis on the latter. This method characterizes accountability of relatively low- and high-level employees. McDonald's, for example, does not evaluate its store managers by sales volume or profitability alone. Instead, a manager and his or her supervisor establish targets for easily quantifiable measures such as sales volume and costs, but also for product quality, service, cleanliness, and personnel training, because these factors may affect long-term profitability as well as the reputation (and thus, profitability) of other outlets.[4]

Certainly, supervisory evaluation of employees is less reliable than numerical output measurements such as storewide sales (or student test scores). Supervisory evaluation may be tainted by favoritism, bias, inflation and compression (narrowing the range of evaluations to avoid penalizing or rewarding too many employees), and even kickbacks or other forms of corruption.[5] Yet the widespread management use of subjective evaluations, despite these flaws, suggests that, as one personnel management review concludes, "it is better to imperfectly measure relevant dimensions than to perfectly measure irrelevant ones."[6]

–R.R.

See last page for endnotes for this excerpt.

patient characteristics to its statistical model. Although the agency now insisted that its model adequately adjusted for all critical variables, the ratings consistently resulted in higher adjusted mortality rates for low-income patients in urban hospitals than for affluent patients in suburban hospitals.[9] Campbell's law swung into action—when surveyed, physicians and hospitals began to admit that they were refusing to treat sicker patients.[10] Surgeons' ratings were not adversely affected by deaths of patients who had been denied surgery. Surveys of cardiologists found that most were declining to operate on patients who might benefit from surgery but were of greater risk.[11] Some hospitals, more skilled at selection, got higher ratings, while others did worse because they received a larger share of patients with more severe disease. In 1989, St. Vincent's Hospital in New York City was put on probation by the state after it placed low in the ranking of state hospitals for cardiac surgery. The following year, it ranked first in the state. St. Vincent's accomplished this feat by refusing to operate on tougher cases.[12]

Attempts to hold schools accountable for math and reading test scores have corrupted education by reducing the attention paid to other important curricular goals; by creating incentives to ignore students who are either above or far below the passing point on tests; by misidentifying failing and successful schools because of test unreliability; by converting instruction into test preparation that has little lasting value; and by gaming, which borders on (or may include) illegality.

As the examples provided demonstrate, each of these corruptions has parallels in other fields, often studied by social scientists and management theorists. But education policymakers have paid little attention to this expertise.[13] Instead, state and federal governments adopted test-based accountability as the tool for improving student achievement, duplicating the worst features of flawed accountability systems in other public and private services.

Some advocates of test-based accountability in education, confronted with evidence of goal distortion or excessive test preparation, have concluded that these problems stem only from the inadequacy of teachers. As one critic argues, good teachers "can and should" integrate subject matter so that raising math and reading scores need not result in diminished attention to other curricular areas.[14] But this expectation denies the intent and power of incentives that, if successful, *should* redirect attention and resources to those outputs that are rewarded. The consistency with which professionals and their institutions respond in this fashion in all fields should persuade us that this is not a problem with the ethos of teachers, but an inevitable consequence of any narrowly quantitative incentive system.

And yet, the fact that exclusively quantitative accountability systems result in goal distortion, gaming, and corruption in a wide variety of fields is not inconsistent with a conclusion that such systems nonetheless improve average performance in the narrow goals they measure. At the very least, they may direct attention to outliers that warrant further investigation. Several analyses by economists, management experts, and sociologists have concluded that narrowly quantitative incentive schemes have, at times, somewhat improved the average performance of medical care, job training, welfare, and private sector agents. The documentation of perverse consequences does not indicate that, in any particular case, the harm outweighed the benefits of such narrow quantitative accountability. But it does raise important questions.

How much gain in reading and math scores is necessary to offset the **goal distortion**—less art, music, physical education, science, etc.—that inevitably results from rewarding schools for score gains only in reading and math?

In education, how much gain in reading and math scores is necessary to offset the goal distortion—less art, music, physical education, science, history, character building—that inevitably results from rewarding teachers or schools for score gains only in reading and math? How much misidentification of high- or low-performing teachers or schools is tolerable in order to improve their average performance? How much curricular corruption and teaching to the test are we willing to endure when we engage in, as one frequently cited work in the business management literature puts it, "the folly of rewarding A while hoping for B"?[15]

Fortunately, no accountability at all is not the only alternative to the flawed approach of exclusive reliance on quantitative output measures. It is possible, indeed practical, to design an accountability system in education to ensure that schools and educators meet their responsibilities to deliver the broad range of outcomes that the American people demand, without relying exclusively on measures as imperfect as test scores. Such a system would be more expensive than our current regime of low-quality standardized tests, and would not give policymakers the comfortable, though false, precision that they want quantitative measures like test scores to provide.

Because Americans have broad goals for their children—from solid academics to responsible citizenship to good health—we require an equally broad accountability system, one that considers test scores, but also relies on human judgment. And, because schools cannot be solely responsible for youth development (or even for closing the achievement gap, which exists before kindergarten), this broad accountability system should include both schools and other institutions that support our children. For more on what such a system should look like, please see the next article. ☐

See last page for endnotes for this excerpt.

Grading Education

Test-Based Accountability Can't Work, But Testing Plus Careful School Inspections Can

By Richard Rothstein, Rebecca Jacobsen, and Tamara Wilder

Noble though its intent may be, the No Child Left Behind Act—the federal law that requires virtually all students to be proficient in reading and math by 2014—is an utter failure. Many critics have denounced it, as well as similar state accountability policies based exclusively on quantitative measures of a narrow set of school outcomes. Critics have described how accountability for math and reading scores has inaccurately identified good and bad schools, narrowed the curriculum (by creating perverse incentives for schools to ignore many important purposes of schools beyond improving math and reading test scores), caused teachers to focus on some students at the expense of others, and tempted educators to substitute gamesmanship for quality instruction.

Despite widespread dissatisfaction with No Child Left Behind (NCLB), Congress has been unable to devise a reasonable alternative and so, for now, NCLB remains on the books. There have been many proposals for tinkering with the law's provisions—extending the deadline for reaching proficiency, measuring progress by the change in scores of the same group of students from one year to the next (instead of comparing scores of this year's students with scores of those in the same grade in the previous year), adding a few other requirements (like graduation rates or parent satisfaction) to the accountability regime, or standardizing the definitions of proficiency among the states. Yet none of these proposals commands sufficient support because none addresses NCLB's most fundamental problem: although tests, properly interpreted, can contribute some important information about school quality, testing alone is a poor way to measure whether schools, or their students, perform adequately.

Perhaps the most important reason why NCLB, and similar testing systems in the states, got accountability so wrong is that we've wanted to do accountability on the cheap. Standardized tests that assess only low-level skills and that can be scored electronically cost very little to administer—although their hidden costs are enormous in the lost opportunities to develop young people's broader knowledge, traits, and skills.

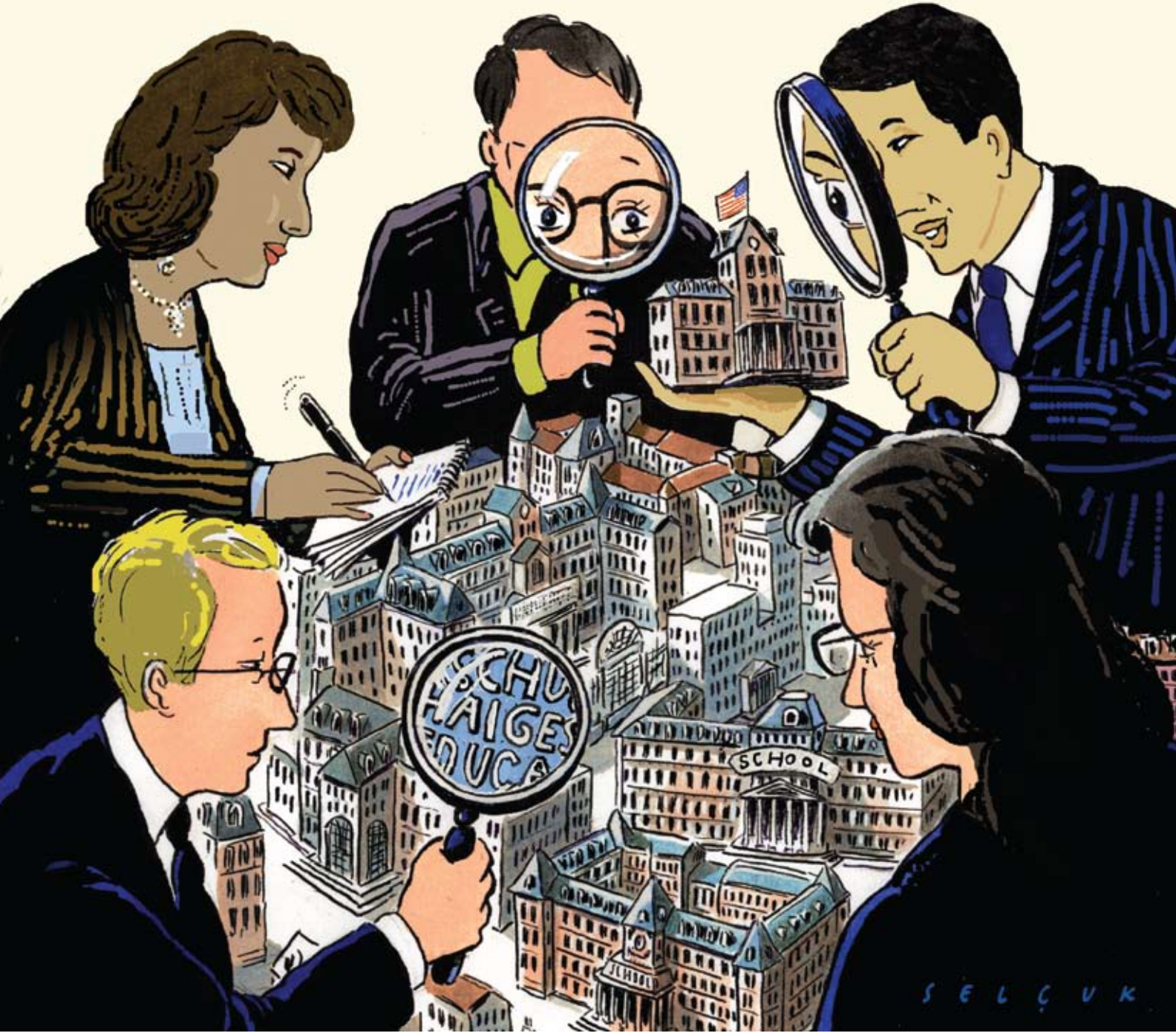
The fact is, schools have an important but not exclusive influence on student achievement; the gap in performance between schools with advantaged children and schools with disadvantaged children is due in large part to differences in the social and economic conditions from which the children come.[1] For this reason, schools can best improve youth outcomes if they are part of an integrated system of youth development and family support services that also includes, at a minimum, high-quality early childhood care, health services, and after-school and summer programs. An accountability system should be designed to ensure that *all* public institutions make appropriate contributions to youth development. When schools are integrated with supporting services, they can substantially narrow the achievement gap between disadvantaged and middle-class children.

A successful accountability system, such as the one we will propose in this article (and which we more fully explain in our book, *Grading Education: Getting Accountability Right*), will initially be more expensive. Our proposal calls for both a sophisticated national assessment of a broad range of outcomes and a corps of professional inspectors in each state who devote the time necessary to determine if schools and other institutions of youth development—early childhood programs, and health and social services clinics, for example—are following practices likely to lead to adult success. But while such accountability will be expensive, it is not prohibitively so. Our rough estimate indicates that such accountability could cost up to 1 percent of what we now spend on elementary and secondary education. If we want to do accountability right, and we should, this level of spending is worthwhile.

In the long run, trustworthy accountability is cost effective. Because narrow test-based accountability can neither accurately identify nor guide schools that need to improve, we now waste

Richard Rothstein is a research associate at the Economic Policy Institute, former national education columnist with the New York Times*, and author of several books, including* Class and Schools: Using Social, Economic, and Educational Reform to Close the Black-White Achievement Gap. *Rebecca Jacobsen is an assistant professor of teacher education and education policy at Michigan State University. Tamara Wilder is a postdoctoral fellow at the University of Michigan's Ford School of Public Policy. Adapted with permission from a book by Rothstein, Jacobsen, and Wilder,* Grading Education: Getting Accountability Right *(**www.epi.org/publications/entry/books_grading_education**), published in 2008 by the Economic Policy Institute and Teachers College Press.*

ILLUSTRATED BY SELÇUK DEMIREL



billions of dollars by continuing to operate low-quality schools. And we waste billions by forcing good schools to abandon high-quality programs to comply with the government's test obsession. We cannot know how much money could be saved by more intelligent accountability, but it is probably considerable.

Of course, no accountability system can be successful without first defining the outcomes that schools and other institutions of youth development should achieve. Before we put forth our vision for a new approach to accountability, let's take a moment to compare the goals that Americans have long valued with the goals that we are currently pursuing.

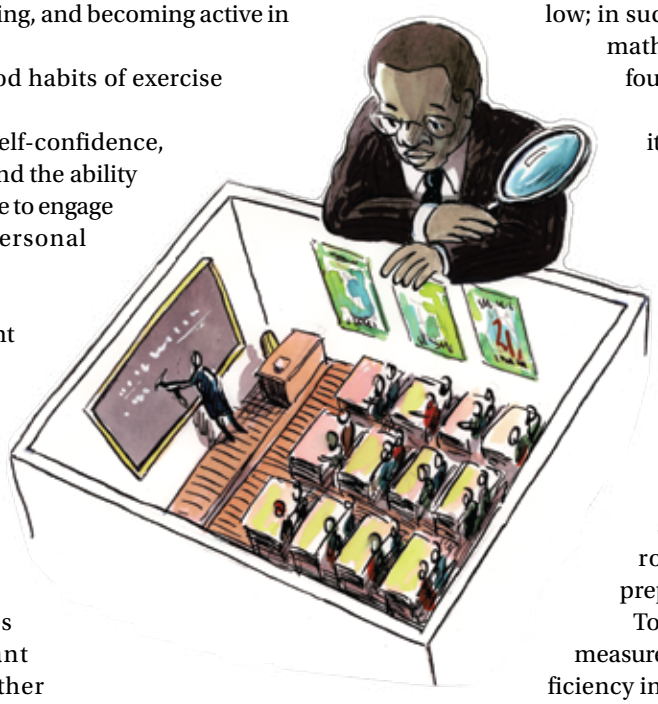
First Things First: Accountability for What?

From our nation's beginnings, Americans have mostly embraced a balanced curriculum to fulfill public education's mission. Looking back over 250 years, we reviewed a small sample of the many statements produced by policymakers and educators to define the range of knowledge, skills, and character traits that schools ought to develop in our youth. We were struck by how similar the goals of public education have remained during America's history. Although some differences of emphasis have emerged during different eras, our national leaders—from Benjamin Franklin to Horace Mann to various university presidents and school superintendents—seem consistently to have wanted public education to produce satisfactory outcomes in the following eight broad categories:

1. *Basic academic knowledge and skills*: basic skills in reading, writing, and math, and knowledge of science and history.
2. *Critical thinking and problem solving*: the ability to analyze information, apply ideas to new situations, and (more

recently) develop knowledge using computers.

3. *Appreciation of the arts and literature*: participation in and appreciation of musical, visual, and performing arts as well as a love of literature.
4. *Preparation for skilled employment*: workplace qualifications for students not pursuing college education.
5. *Social skills and work ethic*: communication skills, personal responsibility, and the ability to get along with others from varied backgrounds.
6. *Citizenship and community responsibility*: public ethics; knowledge of how government works; and participation by voting, volunteering, and becoming active in community life.
7. *Physical health*: good habits of exercise and nutrition.
8. *Emotional health*: self-confidence, respect for others, and the ability to resist peer pressure to engage in irresponsible personal behavior.



Having examined recent surveys of the public's goals for education and having conducted our own poll (in 2005) of the general public, school board members, and state legislators, we are fairly confident that these are, indeed, the outcomes that Americans still want from our schools and other youth institutions.

Unfortunately, today's obsession with reading and math scores means that almost all of these eight goals are ignored. Several surveys of school and district officials, principals, and teachers confirm that the public school curriculum has been dangerously narrowed. But the narrowing did not begin with No Child Left Behind; there was evidence of it throughout the last couple of decades as math and reading tests steadily gained importance. In a 1994–95 survey of Maryland teachers, two-thirds said that they had reduced the amount of time they spent on instruction in nontested subjects, especially art, music, and physical education.[2] In the 1990s, similar curricular shifts were also common in Texas (which, being George W. Bush's home state, provided the model for NCLB). In that state, and especially in schools serving disadvantaged minority students, teachers of art, history, and science were required to put their curricula aside to drill students in the basic math and reading skills that were tested by the state exam.[3]

A survey of school principals in North Carolina, after the state implemented a test-based accountability system in 1999, found that over 70 percent had redirected instruction from other subjects and from character development to reading, math, and writing, and that this response was greatest in the lowest-scoring schools.[4] A 2003 survey of school principals in Illinois, Maryland, New Mexico, and New York found that those in high-minority schools were more likely to have reduced time for history, civics, geography, the arts, and foreign languages to devote more time to math and reading.[5]

The most comprehensive investigations of test-driven curricular shifts have been conducted by the Center on Education Policy, which surveyed 349 representative school districts during the 2006–07 school year. It found that accountability does work: 62 percent of these districts had increased time devoted to reading and math. The increases were greatest in urban districts sanctioned under NCLB because their test scores were too low; in such districts, the increase in reading and math instruction totaled an average of over four hours a week.[6]

This is just what test-based accountability systems intend to accomplish. Students whose reading and math performance was lowest were getting a lot more instruction in these subjects. But increased time for test preparation in reading and math comes at the expense of time for something else. These districts cut an average of an hour or more per week from instruction in social studies, science, art and music, physical education, and recess. Most districts facing sanctions cut time from several of these subject areas to make room for more reading and math test preparation.

To make matters worse, even such drastic measures are unlikely to bring all students to proficiency in reading and mathematics. Inadequate schools are only one reason disadvantaged children perform poorly. They come to school under stress from high-crime neighborhoods and economically insecure households. Their low-cost daycare tends to park them before televisions, rather than provide opportunities for developmentally appropriate play. They switch schools more often because of inadequate housing and rents rising faster than parents' wages. They have greater health problems, some (like lead poisoning or iron-deficiency anemia) directly depressing cognitive ability, and some (like asthma and vision difficulties) causing more absenteeism or inattentiveness. Their households include fewer college-educated adults to provide more sophisticated intellectual environments, and their parents are less likely to expect academic success.[7] Nearly 15 percent of the black-white test-score gap can be traced to differences in housing mobility, and 25 percent to differences in child and maternal health.[8]

Yet contemporary test-based accountability policies expect that school improvement alone will raise all children to high levels of achievement, poised for college and professional success. Teachers are expected to repeat the mantra "all children can learn," a truth carrying the false implication that the level to which children learn has nothing to do with their starting points or with the out-of-school supports they receive. Policymakers and school administrators warn teachers that any mention of

children's socioeconomic disadvantages only "makes excuses" for teachers' own poor performance.

Of course, there are better and worse schools, and better and worse teachers. And of course, some disadvantaged children excel more than others. But our current federal and state test-based accountability policies have turned these obvious truths into the fantasy that teachers can wipe out socioeconomic differences among children simply by trying harder.

It is surprising that so many education policymakers have been seduced into thinking that simple quantitative measures like test scores can be used to hold schools accountable for achieving complex educational outcomes. After all, similar accountability systems have been attempted, and have been found lacking, in other sectors, both private and public, many times before. The corruptions and distortions resulting from test-based accountability are no different from those that have been widely reported in the business world, as well as in fields like health care, welfare, job training, law enforcement, and other government services. (For a quick review of the problems caused by quantitative measures in law enforcement, higher education, health care, and other sectors, see "What's Wrong with Accountability by the Numbers?" on page 20.)

Teachers are expected to repeat the mantra "all children can learn," a truth carrying the **false implication** that the level to which children learn has nothing to do with the out-of-school supports they receive.

The solution, as we briefly stated in the introduction, is not to abandon testing, but to supplement it with periodic inspections of both schools and other organizations that support our youth. Appreciating the arts, developing a strong work ethic, accepting responsibility as a citizen—these goals are as important as our academic goals, and our accountability system should treat them as such. Simply put, we must devise ways of holding schools and other youth development institutions accountable for achieving all eight of the goals that Americans have long valued. And, instead of setting fanciful targets that set up our institutions to fail, we must devise realistic targets that inspire continuous improvement.

Test Prep or True Learning: What's Behind Those Test Scores?

Other nations have also struggled with accountability for public education. Yet while Americans have relied upon test scores alone—and even worse, proficiency cut scores—to judge school quality, others have supplemented standardized testing with school inspection systems that attempt to assess whether students are developing a balanced set of cognitive and noncognitive knowledge and skills. While England, Scotland, Wales, Northern Ireland, the Netherlands, the Czech Republic, Belgium, Portugal, France, and New Zealand[9] all have some form of inspection system, Her Majesty's Inspectors in England offer us a particularly intriguing model because they hold schools and other social welfare institutions accountable for education and youth development.

Because the English inspection system continually undergoes revision, the following describes the English inspectorate as it existed until 2005, when a major revision commenced.

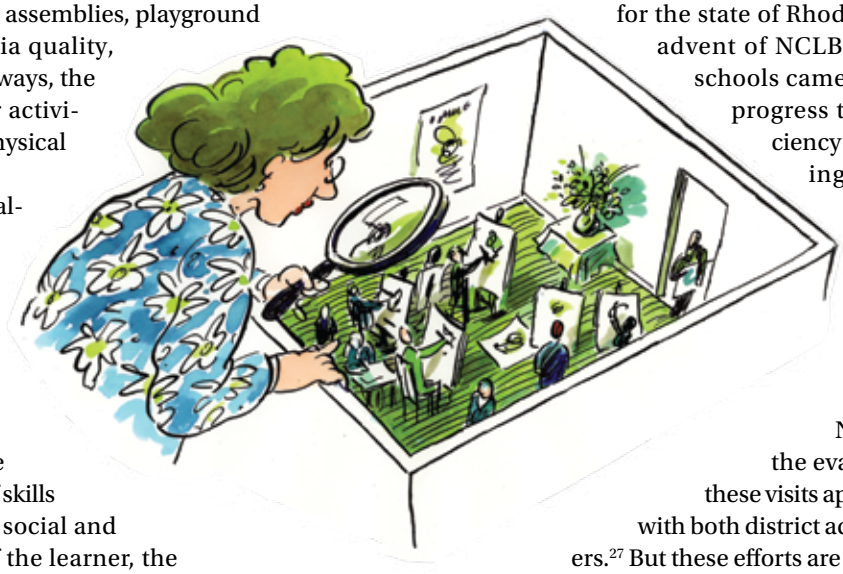
Accountability is overseen by an independent government department, the Office for Standards in Education (Ofsted). In the early part of this decade it had a corps of about 6,000 inspectors who visited schools and wrote reports on their quality. Most inspectors, usually retired school principals or teachers, were directly employed by a dozen or so firms with which Ofsted contracted to conduct the inspections. An elite group, about 200 of "Her Majesty's Inspectors" (HMIs), were employed directly by Ofsted and oversaw the entire process. Ofsted trained the contracted inspectors, required them to attend annual retrainings, and certified them prior to employment. Ofsted also assured the reliability of inspectors' judgments by having several inspectors judge the same educational activity and then comparing their ratings. Ofsted monitored the inspectors' work and removed those whose quality was inadequate—for example, those who never found lessons to be unsatisfactory.[10]

To ensure quality, the leader of each school inspection team underwent a higher level of training than the other team members, and an HMI sometimes also participated in each larger team of contracted inspectors. Ofsted also required each team to include one lay inspector, often a retiree from another profession, to give the inspections greater credibility with the public. Each inspection resulted in a report published on the Internet within three weeks; the report was mailed to every parent, with photocopies also made available to the public.[11] In the case of schools that persistently failed to pass inspection, local governments assumed control and, in the most serious cases, closed them.[12]

Until 2005, a typical full-time English inspector may have visited from 15 to 30 schools each year, and part-time inspectors (usually retired principals) may have visited seven or eight.[13] Because of this experience and their training, English inspectors were highly respected by teachers and principals, who were thus more likely to take inspectors' advice seriously and consider inspectors' evaluations legitimate. Ofsted inspectors were required to spend most of their time observing classroom teaching, interviewing students about their understanding, and examining random samples of student work.[14] Ofsted inspectors decided which students to interview and which classrooms to visit at any particular time.[15] Although they spent relatively little time meeting with administrators, Ofsted inspectors did require principals to accompany them on some classroom observations,

after which the inspectors asked the principals for their own evaluations of the lessons. In this way, the inspectors were able to make judgments (which became part of their reports) about the competence with which the principals supervised instruction.[16]

Ofsted's contracted inspectors observed every teacher in each school, evaluating pupil achievement in all academic as well as in noncognitive areas.[17] Ofsted inspectors rated everything they observed, including teaching skill, student participation, achievement, and academic progress, on a seven-point scale, with supporting paragraphs justifying the ratings. They also wrote reports on student assemblies, playground practice, school cafeteria quality, student behavior in hallways, the range of extracurricular activities, and the quality of physical facilities.[18]



Ofsted reports also evaluated how well schools teach not only academic knowledge and skills but personal development: "the extent to which learners enjoy their work, the acquisition of workplace skills, the development of skills which contribute to the social and economic well-being of the learner, the emotional development of learners, the behaviour of learners, the attendance of learners, the extent to which learners adopt safe practices and a healthy lifestyle, learners' spiritual, moral, social, and cultural development, [and] whether learners make a positive contribution to the community."[19]

Inspections used to be every six years, but then Ofsted changed them to every three years[20] and became more flexible about the frequency of inspections. As the system developed, schools with a history of very high ratings were visited less frequently, with smaller teams, and without every classroom and teacher visited. Schools with a history of poor ratings were visited more often and more intensively.[21]

In recent years, Ofsted added on inspections of early childhood care providers and vocational education programs, and evaluations of how well schools coordinate their own programs with such services. When possible, Ofsted conducts inspections of schools and other child and welfare services in the same community simultaneously.[22]

Ofsted has made no effort to produce fine rankings of schools by which the public could judge each school in comparison with all others. Rather, Ofsted has reported which of three categories schools fall into: those that pass inspection, those in need of fairly modest improvements, and those requiring serious intervention to correct deficiencies.

In addition to regular school inspections, the English system has also included special inspections to evaluate particular problems or curricular areas—for example, music instruction, physical education, the underachievement of minority students, or disparate punishments meted out to them.[23] For these, HMIs visited only a representative group of schools. There were enough of these special inspections, however, that schools were likely to have experienced an inspection for some purpose more frequently than was required by the regular schedule.[24]

England's inspection system may not be perfect—and even if it were, we could not simply adopt it in this country. But it does offer a compelling alternative to our test-based accountability. In the United States, there have been attempts to create a similar inspection system. In the late 1990s, a student of the English inspection system designed a school visit system for the state of Rhode Island.[25] But with the advent of NCLB, it lost importance as schools came to be judged solely on progress toward universal proficiency levels in math and reading. The Chicago school system hired a former English HMI to design a school review system for the district.[26] New York City hired an Ofsted contractor to visit and evaluate all New York City schools; the evaluations resulting from these visits apparently have credibility with both district administrators and teachers.[27] But these efforts are in conflict with contemporary state and federal accountability standards, which make schools almost exclusively accountable for math and reading test scores.

Such attempts to create better accountability systems shouldn't be allowed to collapse under the weight of our obsession with reading and math scores. To fulfill our desire to hold American schools and their supporting public institutions accountable, it makes sense to design a system that draws upon the best elements of standardized testing and inspection systems.

A Better Model: What Would It Look Like?

It is not our intent to present a fully developed accountability proposal; that is a task for policymakers, public officials, and citizens. We only hope to provoke discussion that will help move American policy beyond an exclusive reliance on standardized testing of basic skills.[28]

To begin, we assume that accountability should be a state, not federal, responsibility. Not only do we have a constitutional tradition of state control of education, but the failure of No Child Left Behind has made it apparent that in this large country, the U.S. Congress and U.S. Department of Education are too distant to micromanage school performance.

There are, however, two important tasks for the federal government: (1) to ensure that each state has the fiscal capacity to provide adequate education and other youth services, and (2) to expand the National Assessment of Educational Progress (NAEP) to provide state policymakers with information on the achieve-

ment of their states' young adults and 17-, 13-, and 9-year-olds in the eight broad areas we presented earlier. These two tasks are prerequisite to an accountability system that ensures we, as a nation, are raising the performance of disadvantaged children—and of middle-class children as well. We'll briefly discuss each.

For the last 30 years, reformers concerned with the inadequate resources devoted to the education of disadvantaged children have directed attention almost entirely to intrastate equalization—trying to see that districts serving poor students have as much if not more money to spend as districts serving middle-class children in the same state. These reformers have largely ignored the vast resource inequalities that exist between states. Yet about two-thirds of nationwide spending inequality is between states and only one-third is within them.[29] Efforts to redistribute education funds within states cannot address the most serious fiscal inequalities. Consider one of the most extreme cases, Mississippi: no matter how deep the commitment of its leaders may be to improving achievement, its tax base is too small to raise revenues in the way that wealthier states can, while its challenges—the number of its low-income minority children relative to the size of its population—are much greater than those of many states that are considered more progressive. In general, fewer dollars are spent on the education of the wealthiest children in Mississippi than on the poorest children in New York or New Jersey.

In general, fewer dollars are spent on the education of the wealthiest children in Mississippi than on the poorest children in New York. Yet **federal aid exacerbates inequality** in states' fiscal capacities.

Yet federal aid exacerbates inequality in states' fiscal capacities. Federal school aid—to districts serving poor children—is proportional to states' own spending.[30] New Jersey, which needs less aid, gets more aid per poor pupil than Mississippi, which needs more.

It is politically tough to fix this, because sensible redistribution, with aid given to states in proportion to need and in inverse proportion to capacity, must take tax revenues from states like New Jersey (whose representatives tend to favor federal spending) and direct them to states like Mississippi (whose representatives tend to oppose it).[31] Nonetheless, it is unreasonable to expect states that lack sufficient resources to hold their schools and other institutions of youth development accountable for adequate and equitable performance in each of the eight goal areas.

The second critical task for the federal government should be gathering valid and reliable information on the relative performance of students in the different states. One helpful aspect of No Child Left Behind was the requirement that every state participate in NAEP reading and math assessments for the fourth and eighth grades every two years. Because these are the only assessments administered in common to representative samples of students in all states, they provide a way to compare how each state ensures that its elementary school children gain these two academic skills. To spur effective state-level accountability, the NAEP state-level assessment should:

- *Assess representative samples of students at the state level and on a regular schedule, not only in math and reading, but in other academic subject areas*—science, history, other social studies, writing, foreign language—as well as in the arts, citizenship, social skills, and health behavior. These assessments should include paper-and-pencil test items, survey questions, and performance observations.

- *Gather better demographic data.* NAEP has collected systematic demographic data from its samples of test takers only for race, Hispanic ethnicity, and free or reduced-price lunch eligibility. The range of characteristics within these categories is wide. For example, first- and second-generation Hispanic immigrant children are in different circumstances from those who are third generation and beyond, and students eligible for free meals come from families that may be considerably poorer than those in the reduced-price program. Since 2000, NAEP has collected data on maternal educational attainment, and it would be relatively easy to collect a few other critical characteristics—most notably family structure (e.g., single parent) and the mother's country of birth. Such data could be collected by schools upon a child's initial enrollment and become part of a student's permanent record. Adding these demographic characteristics to state-level NAEP may require minimal expansion of sample sizes, but the payoff to this relatively modest expansion would be substantial, and it would facilitate the ability of state leaders to draw valid conclusions about their policy needs.

- *Report NAEP scores on scales, not achievement levels.* Reports of average scale scores at different points in the distribution, such as quartiles, could be published in language easily understood by the public. State policymakers should then be interested in how the average scale scores of students in each quartile of each relevant demographic subgroup compare with scores of similar students in other states. Successful progress should then be judged by whether such average scores in each achievement quartile make progress toward the scores of comparable students in better-performing states. Note that this approach does away with today's ill-considered achievement levels (which are based on fanciful definitions of "proficiency" that vary wildly from state to state). Since there would be no all-or-nothing cut score, there would then be no "bubble" of students just below the cut score, and teachers and schools would have no incentive to concentrate instruction only on these students. All students would be expected to make progress.

- *Use age-level, not grade-level, sampling.* Age-level assessment is the only way to get an accurate reading of the relative effectiveness of state education and youth policies. With the current grade-level assessment, one state's eighth-grade scores may be higher than another's only because more low-performing seventh graders were held back, not because its ultimate outcomes are superior. If 13-year-olds were assessed regardless of grade, this distortion would be avoided. With age-level sampling, results from states with different promotion and school-age policies could be compared accurately.*

- *Supplement in-school samples with out-of-school samples.* The best evidence of the quality of our education and youth development policies is the performance of 17-year-olds, for whom states are completing their normal institutional responsibility, and of young adults, to see whether knowledge and skills developed earlier are being retained. To get representative samples of 17-year-olds and young adults, assessments should include an out-of-school household survey that covers each of the eight broad goals.

A **full accountability system** requires judgment about whether schools, along with other institutions of youth development, are likely to generate **balanced outcomes** across the eight goals.

Dramatic expansion of NAEP in this fashion need not have the harmful effects that standardized testing under contemporary state and federal accountability policies has produced. Incentives for teachers to "teach to the test" are avoided because NAEP is a sampled assessment, with any one particular school rarely chosen, only a few students in the selected schools assessed, and those students given only portions of a complete exam. There are no consequences for students or schools who do well or poorly, because results are generated only at the state level; nobody knows how particular students or schools performed. Because an expanded NAEP should assess the full range of cognitive and noncognitive knowledge and skills encompassed by the eight broad goals of education, NAEP can give state policymakers and educators no incentives to ignore untested curricular areas.

With this federal support, states can design accountability systems that include academic testing in core subject areas and in those nonacademic fields where standardization is possible, such as health awareness and physical fitness. State accountability systems can supplement such testing and provide detailed school-level data by use of inspection procedures that ensure that adequate performance in each of the eight goal areas is achieved, and that schools and other institutions of youth development implement strategies likely to improve that performance.

State-Level Accountability That Encourages School Improvement

An expanded NAEP can tell governors, legislators, and citizens the extent to which their states are doing an adequate job of generating student success in each of the eight goal areas. Then, citizens and state policymakers can use this information to guide the refinement of state policy. They will want to ensure that particular schools and school districts, children's health care institutions, early childhood and preschool programs, parental support and education programs, after-school and summer programs, and community redevelopment agencies are contributing to, not impeding, the achievement of such success. This requires ways for state government to hold these school districts, schools, and other supporting institutions accountable.

The following proposals sound like a great deal of testing, but keep in mind that it is not necessary to test each subject in each grade level each year. Decisions about what to test, in which grade, and how often should be made at the state level, but a great deal of useful information can be gathered without more tests than students currently take. With that in mind, we propose that states:

- *Cover all eight goals of public education* to avoid the goal distortion that results from accountability for only a few basic skills. Many standardized tests in subjects other than math and reading now exist, but few include constructed-response items, in which students are not given multiple choices but must work out factual or prose answers on their own. Certainly, higher-quality academic tests in history, writing, the sciences, and other academic areas should be deployed, as should standardized assessment instruments, where possible, in nonacademic areas. For example, instruments exist that can assess a student's upper-body strength and, combined with data on the student's weight and height, inform the evaluation of a school's physical education program.[32]

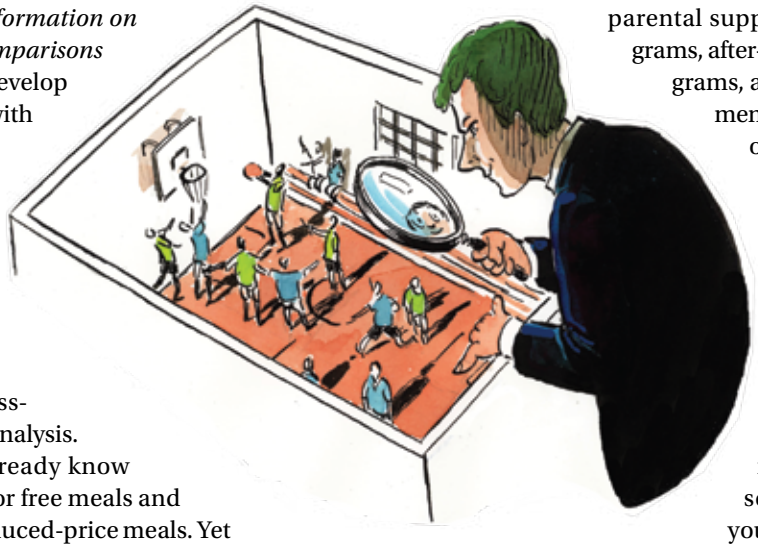
- *Use standardized test scores very cautiously to judge schools, and only in combination with other data.* If states' tests are improved, as they should be, to include higher-quality items that cannot be machine scored, the precision with which the tests can be scored will decline. Many schools are too small to generate reliable results for particular age groups even on

*Age-level sampling in NAEP need not mean that states' own tests used for school-level accountability must be standardized for age instead of grade level. Because states, if they choose, can standardize school entry ages and social promotion policies, grade-level test results are less subject to misinterpretation if confined to particular states. States have an interest in using tests to determine if mandated grade-level curricula are being implemented successfully. Provided that NAEP assesses samples of students of the same age, not grade, we will have the data we need to understand if the combination of age-to-grade policies in some states are more effective than they are elsewhere.

existing low-level tests of basic skills. With more complex items included, reliability will decline further.

- *Supplement information from standardized tests with expert evaluation of student work.* Even the most sophisticated test questions are not fully adequate to reveal students' abilities. NAEP exams include a large number of constructed-response items. But even these questions are no substitute for expert examination of drafts and redrafts of student essays for evidence of how students respond to critiques of their initial efforts and how they develop themes that are longer than those of a brief constructed response on an exam.

- *Collect richer background information on students to make test score comparisons meaningful.* As more states develop good student data systems, with unique student identification numbers and maintenance of cumulative records for each student in secure school databases for the student's entire school career, it will become easier to attach richer background information to student assessment results for purposes of analysis. As one example, schools already know which students are eligible for free meals and which are eligible only for reduced-price meals. Yet in their school "report cards," many (but not all) states and school districts combine these categories, rendering them less useful for understanding and comparing student performance. It would be a simple matter for elementary schools to record, upon a student's initial enrollment, not only the student's subsidized lunch eligibility but also the educational attainment of the mother (or primary caretaker), whether the mother was born in the U.S., and the number of parents or other responsible adults in the student's household.



- *Use NAEP to set realistic goals that inspire continuous improvement.* Goals are valuable, but they should always be feasible, not fanciful. Once NAEP has been expanded, states can establish goals based on the performance of students with similar characteristics in other states. Such goals should be established not only for average performance but also for NAEP performance at the higher and lower ends of the student achievement distribution. If all states regularly established and revised such realistic goals, it would result in a permanent process of continuous improvement.

But test scores and evaluations of student work, even for larger schools, and even when connected to more nuanced student background characteristics, are of only partial value. A full accountability system requires evaluation of student performance in areas more difficult to standardize (for example, cooperative behavior), and judgment about whether a school's curriculum and instruction, along with a community's other institutions of youth development, are likely to generate balanced and adequate outcomes across the eight goals.

To supplement test scores and evaluations of students' written work, states wanting to hold school districts, schools, and supporting institutions accountable require an inspection system. Each state should:

- *Conduct mandatory inspections in each school and in each related community institution* (children's health care services, early childhood and preschool programs, parental support and education programs, after-school and summer programs, and community development agencies) approximately once every three years. Where feasible, accreditation of all these institutions in a particular community should be coordinated. Once the system is firmly established, inspections might be conducted less frequently in communities and schools with satisfactory youth outcomes, and more frequently in communities and schools where outcomes are not satisfactory.

- *Design school inspections to determine primarily whether students are achieving adequate outcomes in all eight goals,* not whether schools are meeting the idiosyncratic goals of their faculties and administrations. Inspection teams should compare schools' performance to higher-performing schools with similar demographic characteristics. Such a standard necessarily will lead to continual improvement by all schools.

- *Make most inspectors professional evaluators,* not volunteers, trained to ensure consistency of judgment, and certified as competent by state (or regional) inspection agencies.

- *Include members of the public, representatives of the business community, or designees of elected officials on inspection teams.* Not only would such participation give inspection greater public credibility, but these members, with their varied backgrounds and perspectives, may detect aspects of school quality requiring improvement that may not be apparent to professional educators.

- *Conduct inspections with little or no advance notice, and give inspectors access to all classrooms for random observation.* Likewise, inspectors should choose random students to invite to interview, and whose work to review.

- *Have teams include in their reports an evaluation and interpretation of schools' standardized test scores, but supplement this* by examining student work, listening to student performances, observing student behavior, and interviewing students to gain insight into their knowledge and skills.

- *Require inspectors to make clear recommendations* about how curriculum, instruction, or other school practices should be improved if they find a school's performance to be inadequate in one or more goal areas. Although schools may choose not to follow the specific advice of inspectors, subsequent inspections (more frequent than once every three years in cases where performance is inadequate) should determine whether performance has improved and, if not, why schools did not follow recommendations for improvement. Inspections of other community institutions should employ similar procedures.

- *Make inspection reports public*, and in a timely fashion. Reports should include responses by administrators or teachers to inspectors' criticisms.

- *Establish consequences.* States should assume direct control of schools and other public institutions of youth development when improvement does not follow repeated inspection reports that indicate severe problems.

The accountability system outlined here would not be cheap. But neither would it be so expensive that this proposal is unrealistic, as the following "back-of-the-envelope" estimate shows. At present, the federal government spends about $40 million annually to administer a state-level NAEP exam in math or reading in grades 4, 8, and 12. Assessing 9-, 13-, and 17-year-olds instead could add a little, but not much, to the cost (because, for example, a few 13-year-olds might be found in high schools, not middle schools). Design costs (including substituting new items as old items are rotated out) also add relatively little cost. Expanding samples so that state-level information can be disaggregated into finer demographic subgroups also adds relatively little cost. Adding additional academic and nonacademic subjects (writing, history, other social studies, science, foreign language, health knowledge, physical fitness, and understanding of the arts and vocations) at the state level need not duplicate the full cost for each subject if only paper-and-pencil items are used, because NAEP could use many of the same schools that it samples for math and reading. There would, however, be additional costs for preparing test booklets that included sophisticated multicolor maps or art reproductions. Adding performance and other nontraditional items that can easily be standardized (for example, tests of upper-body strength or identification of musical themes) would incur substantial additional expense. As a very rough estimate, expanding regular state-level NAEP into all eight goals and into all subject areas within the academic categories, and administering such assessments every three years, with appropriate subgroup reporting, might cost a total of $500 million annually.

Supplementing these in-school assessments with a NAEP for out-of-school 17-year-olds and young adults, requiring a household survey conducted once every three years, might cost as much as an additional $20 million annually.

The **total cost** of the accountability system we have outlined here would be no more than **1 percent** of total elementary and secondary public school spending in the U.S.

In England, when inspections in each school took place approximately every six years, the school inspection system cost about one-quarter of 1 percent of total elementary and secondary school spending. If we assume a similar ratio for a system in the U.S., with teams visiting schools approximately every three years, the annual cost would be about $2.5 billion, or one-half of 1 percent of current federal, state, and local spending on elementary and secondary education. Additional costs would be incurred for inspecting other institutions of youth development.

Even with the additional costs of an expanded in-school state NAEP, and of a young adult and 17-year-old out-of-school state NAEP, the total cost of the accountability system we have outlined here would still be no more than 1 percent of total elementary and secondary public school spending in the U.S. This is not an unreasonable price for an accountability system that measures whether schools in every state, in coordination with other institutions of youth development, are preparing young adults to have adequate academic knowledge and skills, appreciation of the arts and literature, preparation for skilled work, social skills and work ethic, citizenship and community responsibility, physical health, and emotional health. If this system succeeded in correcting even some of the unproductive practices in schools and other institutions, the gains in efficiency would more than justify this expenditure. When accountability funds are spent correctly, they eliminate waste and save funds.

But saving money, probable though that might be in the long run, is not the primary purpose of an accountability system. If we truly want to hold institutions accountable for fulfilling the missions to which they have been assigned by the nation, and if we are determined to reverse the corruptions we have visited on schools by narrow test-based accountability policies, we should willingly entertain a system of accountability that might require higher expenditures in the short run.

No Child Left Behind has given accountability a bad name. An alternative program along the lines suggested here could redeem accountability's reputation. And it could give the citizens of this nation a better means to fulfill our responsibilities to provide for our youth and the nation's future. ❑

See last page for endnotes for this excerpt.

Endnotes

What's Wrong with Accountability by the Numbers?

1. Herbert A. Simon, "Rational Decision-Making in Business Organizations" (Nobel Memorial Lecture, December 8, 1978), 352, 366.

2. Donald T. Campbell, "Assessing the Impact of Planned Social Change," *Evaluation and Program Planning* 2 (1979): 67–90 (reprinted, with minor revisions and additions, from *Social Research and Public Policies,* ed. Gene M. Lyons (Hanover, NH: University Press of New England, 1975), 85.

3. Michael Murray, "Why Arrest Quotas Are Wrong," *PBA Magazine*, Spring 2005.

4. Scott Jaschik, "Should U.S. News Make Presidents Rich?" *Inside Higher Ed*, March 19, 2007.

5. Alan Finder, "College Ratings Race Roars on Despite Concerns," *New York Times*, August 17, 2007.

6. David Seidman and Michael Couzens,"Getting the Crime Rate Down: Political Pressure and Crime Reporting," *Law & Society Review* 8, no. 3 (1974): 457–494.

7. Seidman and Couzens, "Getting the Crime Rate Down," 462.

8. Lisa I. Iezzoni, "Risk and Outcomes," in *Risk Adjustment for Measuring Health Care Outcomes*, ed. Lisa I. Iezzoni (Ann Arbor, MI: Health Administration Press, 1994), 4.

9. Allen Schick, "Getting Performance Measures to Measure Up," in *Quicker, Better, Cheaper?: Managing Performance in American Government*, ed. Dall W. Forsythe (Albany: Rockefeller Institute Press, 2001), 41.

10. Lawrence P. Casalino et al., "General Internists' Views on Pay-for-Performance and Public Reporting of Quality Scores: A National Survey," *Health Affairs* 26, no. 2 (2007): 492–499, 495.

11. Marc Santora, "Cardiologists Say Rankings Sway Choices on Surgery," *New York Times*, January 11, 2005; and Casalino et al., "General Internists' Views," 496.

12. Lawrence K. Altman, "Heart-Surgery Death Rates Decline in New York," *New York Times*, December 5, 1990.

13. This article is not the first, or only, discussion of the applicability of Campbell's law to contemporary test-based educational accountability policies. The following have made similar observations: Sharon L. Nichols and David C. Berliner, *Collateral Damage: How High-Stakes Testing Corrupts America's Schools* (Cambridge, MA: Harvard Education Press, 2007); Daniel Koretz, "Inflation of Scores in Educational Accountability Systems: Empirical Findings and a Psychometric Framework" (powerpoint prepared for the Eric M. Mindich Conference on Experimental Social Science, in *Biases from Behavioral Responses to Measurement: Perspectives From Theoretical Economics, Health Care, Education, and Social Services*, Cambridge, MA, May 4, 2007); and Daniel Koretz, *Measuring Up: What Educational Testing Really Tells Us* (Cambridge, MA: Harvard University Press, 2008).

14. Martin West, "Testing, Learning, and Teaching: The Effects of Test-Based Accountability on Student Achievement and Instructional Time in Core Academic Subjects," in *Beyond the Basics: Achieving a Liberal Education for All Children*, eds. Chester E. Finn Jr. and Diane Ravitch (Washington, DC: Thomas B. Fordham Institute, 2007), 45–62, 57.

15. Steven Kerr, "On the Folly of Rewarding A While Hoping for B," *Academy of Management Journal* 18, no. 4 (1975): 769–783.

What Really Happens in the Private Sector?

1. Elissa Gootman, "Teachers Agree to Bonus Pay Tied to Scores," *New York Times*, October 18, 2007.

2. Michael Bloomberg, "Mayor Bloomberg, Chancellor Klein and UFT President Weingarten Announce Schoolwide Bonus Plan to Reward Teachers at Schools that Raise Student Acheivement," Mayor's Press Release No. 375, October 17, 2007.

3. Scott J. Adams and John S. Heywood, "Performance Pay in the U.S.: Concepts, Measurement and Trends" (2nd Draft, Economic Policy Institute, November 19, 2007), Tables 2 and 7.

4. Robert S. Kaplan and Anthony A. Atkinson, *Advanced Management Accounting*, 3rd ed. (Englewood Cliffs, NJ: Prentice Hall, 1998), 692–693.

5. Christopher D. Ittner, David F. Larcker, and Marshall W. Meyer, "Performance, Compensation, and the Balanced Scorecard" (Philadelphia: Wharton School, University of Pennsylvania, November 1, 1997), 9. That labor market success seems to be correlated with employees' physical attractiveness confirms that supervisory evaluations are flawed tools for objective evaluations of performance. See Daniel S. Hamermesh and Jeff E. Biddle, "Beauty and the Labor Market," *American Economic Review* 84, no. 5 (1994): 1174–1194.

6. William H. Bommer et al., "On the Interchangeability of Objective and Subjective Measures of Employee Performance: A Meta-Analysis," *Personnel Psychology* 48, no. 3 (1995): 587–605, 602.

Grading Education

1. The conclusions of many researchers and policy experts on this point are summarized in Richard Rothstein, *Class and Schools: Using Social, Economic, and Educational Reform to Close the Black-White Achievement Gap* (New York: Teachers College Press, 2004).

2. Daniel Koretz, Karen Mitchell, Sheila Barron, and Sarah Keith, *Final Report: The Perceived Effects of the Maryland School Performance Assessment Program*, CSE Technical Report No. 409 (Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California, 1996), Table 6.

3. Linda M. McNeil, *Contradictions of School Reform: Educational Costs of Standardized Testing* (New York: Routledge, 2000), 242–243 and passim.

4. Helen F. Ladd and Arnaldo Zelli, "School-Based Accountability in North Carolina: The Responses of School Principals," *Educational Administration Quarterly* 38, no. 4 (2002): 494–529, Figures 5 and 11.

5. Claus Von Zastrow, with Helen Janc, *Academic Atrophy: The Condition of the Liberal Arts in America's Public Schools* (Washington, DC: Council for Basic Education, 2004), Figure 17.

6. Jennifer McMurrer, *Choices, Changes, and Challenges: Curriculum and Instruction in the NCLB Era* (Washington, DC: Center on Education Policy, July [revised December] 2007), Table 3; and Jennifer McMurrer, *Instructional Time in Elementary Schools: A Closer Look at Changes in Specific Subjects* (Washington, DC: Center on Education Policy, February 2008).

7. These problems are discussed at length in Rothstein, *Class and Schools*, chapter 1. More recent and eloquent treatments of these issues are in Susan B. Neuman, *Changing the Odds for Children at Risk: Seven Essential Principles of Education Programs that Break the Cycle of Poverty* (Westport, CT: Praeger, 2008); and Susan B. Neuman, "Education Should Lift All Children," *Detroit Free Press*, July 31, 2008.

8. For estimate of effect of mobility: Eric A. Hanushek, John F. Kain, and Steven G. Rivkin, "Disruption Versus Tiebout Improvement: The Costs and Benefits of Switching Schools," *Journal of Public Economics* 88, nos. 9–10 (2004): 1721–1746; for estimate of effect of child and maternal health: Janet Currie, "Health Disparities and Gaps in School Readiness," *The Future of Children* 15, no. 1 (2005): 117–138.

9. Melanie C. M. Ehren and A. J. Visscher, "The Relationships Between School Inspections, School Characteristics, and School Improvement," *British Journal of Educational Studies* 56, no. 2 (2008): 205–227.

10. Peter Matthews and Pam Sammons, *Improvement Through Inspection: An Evaluation of the Impact of Ofsted's Work* (London: Institute of Education, University of London, and Office for Standards in Education [Ofsted], July 2004), 83–84.

11. Matthews and Sammons, *Improvement Through Inspection*, 9; Thomas A. Wilson, *Reaching for a Better Standard: English School Inspection and the Dilemma of Accountability for American Public Schools* (New York: Teachers College Press, 1996), 134; and Office for Standards in Education (Ofsted), *Every Child Matters: Framework for the Inspection of Schools in England from September 2005* (London: Ofsted, April 2008).

12. W. Norton Grubb, "Opening Classrooms and Improving Teaching: Lessons from School Inspections in England," *Teachers College Record* 102, no. 4 (2000): 696–723, 709.

13. Tim Brighouse (visiting professor of education at the Institute of Education, London University, former chief adviser to London Schools and former chief education officer for Birmingham), personal correspondence and telephone interview with author (various dates, and May 8, 2008).

14. Grubb, "Opening Classrooms and Improving Teaching," 701, 703; and Wilson, *Reaching for a Better Standard*, 127.

15. Grubb, "Opening Classrooms and Improving Teaching," 703; and Wilson, *Reaching for a Better Standard*, 71.

16. Brighouse, personal correspondence and telephone interview with author.

17. Matthews and Sammons, *Improvement Through Inspection*, 14, 34; and Grubb, "Opening Classrooms and Improving Teaching," 701.

18. Grubb, "Opening Classrooms and Improving Teaching," 701; and Brighouse, personal correspondence and telephone interview with author.

19. Ofsted, *Every Child Matters*, 22.

20. Ofsted, *Every Child Matters*.

21. Brighouse, personal correspondence and telephone interview with author.

22. Ofsted, *Every Child Matters*, 9.

23. Matthews and Sammons, *Improvement Through Inspection*, 112, 108; and Rebecca Smithers, "Punishment for Black Pupils Appears Harsher: Watchdog's Report Points to Inconsistency Over Exclusions," *Guardian*, March 1, 2001.

24. Matthews and Sammons, *Improvement Through Inspection*, 150; Smithers, "Punishment for Black Pupils"; and Ofsted, *Every Child Matters*.

25. The system was designed, and then implemented, by Thomas A. Wilson, whose study (Wilson, *Reaching for a Better Standard*) of the English system prior to the 1993 reforms made it familiar to American education experts. See Rhode Island Department of Elementary and Secondary Education, "School Accountability for Learning and Teaching (SALT)" (Providence, RI: RIDE, 2008).

26. Grubb, "Opening Classrooms and Improving Teaching," 718.

27. Randi Weingarten, Keynote Address (33rd Annual Conference of the American Education Finance Association, Denver, CO, April 10, 2008).

28. Others are also helping to provoke this discussion. The proposal set forth here joins a conversation in which Ladd (Helen F. Ladd, "Holding Schools Accountable Revisited" [Spencer Foundation Lecture in Education Policy and Management, Association for Public Policy Analysis and Management, 2007]), Nichols and Berliner (Sharon L. Nichols and David C. Berliner, *Collateral Damage: How High-Stakes Testing Corrupts America's Schools* [Cambridge, MA: Harvard Education Press, 2007]), and Dorn (Sherman Dorn, *Accountability Frankenstein: Understanding and Taming the Monster* [Charlotte, NC: Information Age Publishing, 2007]) have engaged. Jones (Ken Jones, "Thinking Ahead," in *Democratic School Accountability*, ed. Ken Jones [Lanham, MD: Rowman and Littlefield Education, 2006]), and Fruchter and Mediratta (Norm Fruchter and Kavitha Mediratta, "Bottom-Up Accountability: An Urban Perspective," in *Democratic School Accountability,* ed. Ken Jones [Lanham, MD: Rowman and Littlefield Education, 2006]) envision an accountability system with elements similar to those proposed here, but where accountability is primarily to local governing bodies (school boards or parent councils), not state government.

29. Sheila E. Murray, William N. Evans, and Robert M. Schwab, "Education-Finance Reform and the Distribution of Education Resources," *American Economic Review* 88, no. 4 (1998): 789–812, 808.

30. Goodwin Liu, "Improving Title I Funding Equity Across States, Districts, and Schools," *Iowa Law Review* 93 (2008): 973–1013.

31. Rothstein (Richard Rothstein, "Equalizing Education Resources on Behalf of Disadvantaged Children," in *A Notion at Risk: Preserving Public Education as an Engine of Social Mobility*, ed. Richard D. Kahlenberg [New York: Century Foundation Press, 2000]) and Liu (Goodwin Liu, "Interstate Inequality in Educational Opportunity," *New York University Law Review* 81, no. 6 [2006]: 2044–2128) offer proposals for interstate finance equalization. They differ in that Liu proposes an adjustment for state tax effort, and Rothstein does not.

32. Cooper Institute, *Fitnessgram/Activitygram* (Dallas: Cooper Institute, 2008).