



## Richard Rothstein

**A**ccountability based exclusively on test score gains has corrupted American education. To spend more time raising reading and math scores, schools pay less attention to nontested academic subjects, as well as to the arts and the development of citizenship and character.

Reading and math curricula have been narrowed to the most easily tested basic skills. Meanwhile, teachers have generated inflated test scores and a false sense of progress by substituting drill, test preparation, and test-taking tricks for good instruction.

To meet accountability targets, schools focus on children just below pre-determined passing points, overlooking those already above those points and too far below to pass. And test-based accountability has thrown schools into needless turmoil by falsely identifying high-quality schools as needing improvement (and falsely identifying inadequate schools as satisfactory).

These perverse consequences of accountability are particularly disturbing because similar corruption in other fields has been thoroughly documented by journalists and social scientists. Designers of No Child Left Behind (NCLB) and similar

narrow, test-based state accountability plans simply ignored well-established research that demonstrates the ineffectiveness of using quantitative measures alone.

### Goal distortion

Before the Soviet Union collapsed in 1991, Western analysts often described goal distortion and corruption endemic to Soviet attempts to manage an economy by establishing targets for production and punishing managers who fell short.

There were targets, for example, for shoe production. Certainly, increasing output was important, but it was not the only shoe industry goal—factories also needed to produce a variety of sizes. Instead, managers responded to the quota by using limited supplies of leather to produce a glut of useless small sizes.

Attempts in America today to hold schools accountable for math and reading test scores have produced similar goal distortion, reducing attention paid to other important goals. In the October 2006 issue of *American School Board Journal*, I described a Los Angeles teacher who said:

“The pressure became so intense that we had to show how every single lesson we taught connected to a standard that was going to be tested. This meant that art, music, and even science and social studies were not a priority and were hardly ever taught. We were forced to spend 90 percent of the instructional time on reading and math. This made



# Taking Aim at Testing

*Accountability targets are only one measure—and an inaccurate one at that—of student success*

teaching boring for me and was a huge part of why I decided to leave the profession.”

I have since interviewed many others with similar experiences. The consistency with which professionals and their institutions respond in this fashion in all fields should persuade us that this is not a peculiar failure of teachers, but an inevitable consequence of any narrowly quantitative incentive system.

## A few familiar examples

Perhaps the most tragic example of goal distortion from quantifiable accountability stemmed from the work of a former business executive turned public official, U.S. Secretary of Defense Robert McNamara. In the Vietnam War, he sought quantitative success measures and demanded that generals report relative “body counts” of Americans and North Vietnamese.

Just as high test scores usually indicate reading proficiency, relative casualties usually indicate military success. But when imposing casualties became an end in itself, the army was corrupted. In Vietnam, officers pleased superiors by reporting fewer American deaths than those of the enemy; enemy death numbers were inflated (for example, by counting civilians killed as enemy deaths). Our leaders confused superiority in body counts with achievement of political objectives, contributing to the loss of the war.

Other familiar examples of goal distortion include:

- Motorists stopped for trivial traffic violations experi-

ence an accountability system that evaluates the police by whether or not they meet ticket quotas. Officers judged by this easily quantifiable outcome can't afford to concentrate on more serious infractions.

■ *U.S. News and World Report's* annual ranking of colleges truly is an accountability system; college presidents have received bonuses for raising rankings. The ranking considers how selective a college is, determined partly by SAT scores of entering freshmen and the percentage of applicants who are admitted (higher-quality colleges accept relatively fewer applicants).

Once these indicators became high-stakes accountability measures, colleges had incentives to boost the number of rejected applicants. Promotional mailings were sent to unqualified applicants with application fees waived. One college offered monetary rewards to already-admitted high school seniors if they re-took the SAT and got higher scores, raising the college's rank. These indicators have lost validity.

## Health care report cards

The corruption of health care from quantitative accountability has been documented in Great Britain and in the United States, where governments created “report cards” to compare doctors' and hospitals' open-heart surgery survival rates.

Like schools, health care providers must balance multiple goals. For heart surgery, one is to reduce mortality.

Another is to give terminally ill patients the choice to avoid artificial technology that prolongs life only briefly, so federal legislation requires hospitals to promote living wills. The two goals can be reconciled only by difficult judgments of physicians and families.

Heart surgery report cards undermined this balancing process. Rewarding hospitals for reducing easily measured mortality, the government created incentives to discourage the use of living wills.

Britain's National Health Service also rated maternity services, publishing data on newborn death rates. These are easily measured, but obstetricians have other objectives such as reducing handicaps for high-risk infants. This requires maternity services to devote resources to prenatal care.

To raise ratings, maternity services invested less in prenatal care and more in hospital services. The mortality rate declined, just as the incentives intended. But developmental outcomes for live births were worse—more low birthweight deliveries and more lifelong learning difficulties and behavioral problems, from less attention paid to prenatal care.

Test-based accountability in education should account for differences in student characteristics. Schools with many low-income children who have serious health problems, high residential mobility, great family stress, and little literacy support at home may be great schools—even with low test scores.

Some educators try to compare only “similar” schools—those with similar proportions of minority students and those eligible for free and reduced-price lunches. But these comparisons still don't work too well, because background differences can be subtle. Stable working class families with incomes nearly double the poverty line are eligible for subsidized lunches; schools with such students can easily get higher scores than schools with poorer students, yet the latter may be more effective. Charters can enroll minority students whose parents are more highly motivated, leading to false claims of superiority when test scores rise.

Medicine faces similar problems; some patients are sicker and thus harder to cure than others with similar disease. Patients' ages, other illnesses, prior treatment, health habits (such as smoking), diet, and home environment all matter. So health care report cards have been “risk-adjusted” for patients' initial conditions. Yet health policy experts still conclude that accountability systems cannot adjust comparisons adequately for patient characteristics.

### Taking another look

Researchers who re-analyzed federal Medicare data, controlling for additional background factors, found that nearly half of the hospitals that Medicare identified as having high mortality for cardiac surgery, purportedly because of poor-quality care, no longer would be in the high-mortality group if patient characteristics were more adequately controlled.

In 1994, the U.S. General Accounting Office (GAO) ana-

lyzed health care report cards and concluded that no adjustments for patient characteristics were sophisticated enough to be “valid and reliable.” The GAO added that “administrators will place all their organizations' resources in areas that are being measured. Areas that are not highlighted in report cards will be ignored.”

In 2003, a team of economists concluded that health care report cards “may give doctors and hospitals incentives to decline to treat more difficult, severely ill patients.” Quantitative medical accountability has “led to higher levels of resource use [because delaying surgery for sicker patients necessitated more expensive treatment later] and to worse outcomes, particularly for sicker patients... [A]t least in the short run, these report cards decreased patient and social welfare.”

One of the paper's authors was Mark McClellan, a member of President George W. Bush's Council of Economic Advisers. Although report cards advertised that some hospitals got better outcomes, the paper concluded that “on net,” they “were particularly harmful. ... Report cards on the performance of schools raise the same issues and therefore also need empirical evaluation.”

McClellan was apparently not consulted about the design of NCLB.

The same reliance on quantitative indicators has corrupted job training and welfare agencies. Under the Job Training Partnership Act (JTPA) of 1982, local agencies that had better job placement records got financial rewards. The U.S. Department of Labor defined successful placements as those lasting at least 90 days. This created incentives to place workers in low-skill and short-term jobs that might last not much longer than 90 days.

In some cases, job training agencies provided special services to support employment, such as child care, transportation, or clothing allowances. Such services were terminated after the 90th day of employment. James Heckman, a Nobel laureate in economics, concluded that JTPA “performance standards based on short-term outcome levels likely do little to encourage the provision of services to those who benefit most from them.”

### The private sector

Calls for test-based accountability in education are frequently accompanied by claims that the private sector works this way. New York City Mayor Michael Bloomberg, announcing a plan to pay cash bonuses at schools where test scores increase, 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.”

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 link-

ing performance and pay is a powerful incentive.”

This misrepresents how firms motivate employees. Although incentive pay is commonplace for private sector professionals, it is almost never based primarily on quantitative output measurement. Indeed, while the share of employees who get performance pay has increased, the share who get it based on numerical output has declined. Instead, performance pay is now usually based largely on subjective supervisory evaluations. Business management journals are filled with warnings about incentives that rely heavily on quantitative rather than qualitative measures.

Employees can easily game purely quantitative incentives, so most corporate accountability systems blend quantitative and qualitative indicators, with emphasis on the latter. Even McDonald's and Wal-Mart do not evaluate store and other managers by sales volume or profitability alone. Instead, managers and supervisors negotiate targets for easily quantifiable measures such as sales volume and costs, but also less easily quantifiable product quality, service, cleanliness, and personnel training. Managers are judged by balancing these factors. Similar systems are also common for professionals in the private sector.

Certainly, supervisory evaluations may be tainted by favoritism, bias, even kickbacks or other corruption. Yet the fact that subjective evaluations are so widely used, 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.” Or, according to another management review, “The prevalence of subjectivity in the performance measurement systems of virtually all [business] organizations suggests that exclusive reliance on distorted and risky objective measures is not an efficient alternative.”

Bain and Company, a consulting firm, advises clients that results should be judged based on long-term, not short-term (and more easily quantifiable) goals. A company director estimated that Bain's managers devote about 100 hours a year to evaluating five employees for its incentive pay system. “When I try to imagine a school principal doing 30 reviews, I have trouble,” he observed.

### The balanced scorecard

A now popular corporate accountability tool is the balanced scorecard, first proposed in the early 1990s because management theorists concluded that quantifiable short-term financial results were not accurate guides to profitability. Goals are too complex to reduce to quantifiable measures because, as business theorist Robert Kaplan puts it, financial success relies on “intangible and intellectual assets, such as high-quality products and services, motivated and skilled employees, responsive and predictable internal processes, and satisfied and loyal customers.”

Kaplan says “best-practice firms” that employ balanced

scorecards and use subjective judgments believe results-based compensation may not always be the ideal scheme for rewarding managers. The reason: “Many factors not under the control or influence of managers also affect reported performance [and] many managerial actions create (or destroy) economic value but may not be measured.”

Curiously, the federal government uses balanced scorecards for education simultaneously with its quantitatively based NCLB. Since 1988, the U.S. Department of Commerce has given annual Malcolm Baldrige National Quality Awards to exemplary institutions. Numerical output indicators play only a small role in award decisions; for the private sector, 450 out of 1,000 points are for results.

Even here, some indicators termed “results”—ethical behavior, social responsibility, trust in senior leadership, workforce capability and capacity, and customer satisfaction and loyalty—are based on points awarded for qualitative judgments. Other criteria, also relying on qualitative evaluation, comprise the other 550 points, such as how senior leaders “set organizational vision and values,” and “protection of stakeholder and stockholder interests, as appropriate.”

For educational institutions, only 100 of 1,000 points are for “student learning outcomes,” with other points awarded for subjectively evaluated measures, such as “how senior leaders’ personal actions reflect a commitment to the organization’s values.”

The most recent Baldrige award in elementary and secondary education went in 2005 to Oklahoma’s Jenks School District. The Department of Commerce cited the district’s test scores, low teacher turnover, and innovative programs such as an exchange relationship with schools in China and the enlistment of residents of a local long-term care facility to mentor kindergartners.

Yet in 2006, the Jenks district was deemed by the U.S. Department of Education to be substandard under NCLB because economically disadvantaged and special education students failed for two consecutive years to make Adequate Yearly Progress in reading scores.

It is possible, indeed practical, to have accountability in education that ensures that educators meet their responsibilities to deliver the broad outcomes that the public demands, without relying exclusively on measures as imperfect as test scores. But 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 expect quantitative measures like test scores to provide. ■

---

Richard Rothstein (riroth@epi.org) is a research associate of the Economic Policy Institute. This article is adapted from his book, coauthored with Rebecca Jacobsen and Tamara Wilder, *Grading Education: Getting Accountability Right* (Teachers College Press, 2008).