The CORRUPTION of SCHOOL ACCOUNTABILITY

How experience with quantitative measurements in other sectors can inform the use of high-stakes test scores in education

BY RICHARD ROTHSTEIN

It has become conventional to say that holding educators accountable and paying for higher test scores will improve performance. When New York City Mayor Michael Bloomberg recently announced the city would pay teachers bonuses where 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."

Real estate developer Eli Broad, whose foundation promotes incentive pay 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."

Yet the two billionaires' statements were misleading about how other industries and sectors behave. In the private sector, pay is almost never based primarily on quantitative performance measures. Fewer firms than in the past now use commissions and piece rates for sales and production workers, and more firms award bonuses to professionals based largely on subjective supervisory evaluations.

NCLB Defenders

It is not hard to see why. Under No Child Left Behind, reliance solely on numerical measures, principally math and reading scores, to evaluate performance has corrupted schooling. Educators have responded rationally to incentives that, to devote more time to math and reading, spur reductions in social studies, science, art, music, physical education, cooperative learning and other character-building activities. Reductions have been most severe for disadvantaged students who are most in need of a balanced curriculum.

Schools now focus on "bubble" students (those just below the proficiency point) to the detriment of students who are far behind or already above proficiency. Accountability for an unreliable single test score results in arbitrary classifications of some fine schools as failing, and some poor schools as adequate. Drills leading to limited long-term learning and "teaching to the test" have become commonplace. Some schools manipulate data — for example, by opportunistic assignment of students to subgroups where they do the least harm to school ratings.

Defenders of NCLB are quick to denounce teachers for such tactics. They suggest if only teacher quality were higher, teachers would get high test scores without teaching to the test or engaging in other shortcuts to high test scores that do not reflect true learning. But teachers' responses to quantitative accountability systems are no worse than the responses of professionals and nonprofessionals in other fields to similar performance incentives. Such responses are inherent in simplistic quantitative evaluations.

Corruption of education from NCLB-type accountability, therefore, should have been foreseen. After all, there are many commonplace examples of the harm that such systems can do. Those familiar with the instructional distortions and gaming that have characterized schools' responses to NCLB and similar state test-score accountability policies should see obvious analogies in these examples from other fields.

▶ During the Vietnam War, Secretary of Defense Robert McNamara believed strongly in numerical measures and demanded reports of American and North Vietnamese body counts. Just as high reading test scores usually indicate reading proficiency, relative casualties usually indicate the fortunes of nations at war. But an army can be corrupted if local commanders are judged primarily by this relatively easily measured indicator, losing sight of political and economic objectives. High enemy body count numbers (sometimes contrived) misled American leaders into believing the war was being won.

▶ Motorists cited for trivial traffic violations may have experienced an accountability system in which commanders evaluate police officers by ticket quotas. Certainly, issuing citations for violations is one measure of good policing, but when officers are judged by this easily quantifiable outcome, they have incentives to focus on trivial offenses that meet a numerical goal rather than investigating serious crimes where payoffs may be less certain.

▶ The Federal Bureau of Investigation tracks local police clearance rates as a basis for evaluating effectiveness. The clearance rate is the percentage of reported crimes resulting in convictions. Just as high math scores characterize effective schools, high clearance rates characterize effective police departments. But as with math scores, the clearance rate indicator is corrupted when it becomes an end in itself.

Police increase the rate by offering reduced charges to suspects who confess other crimes, including those they have not actually committed. Such plea bargains give detectives big boosts in clearance rates. Meanwhile, those pleading guilty only to the crime for which they were arrested typically get harsher sentences than those who falsely confess to multiple crimes.

▶ Television stations sell advertising at rates determined by viewership during designated sweeps months when a survey company, Nielsen, determines what programs typical viewers watch. The system assumes that sweeps programming is representative of programming throughout the year. Yet stations respond to these high-stakes surveys by scheduling programs that are more attention-grabbing than a typical month's shows. When viewership numbers during sweeps months become ends in themselves rather than a reflection of year-round program popularity, they distort advertising rates.

▶ Several newspapers, most notably *The New York Times*, publish weekly best-seller lists. Books that make the list get special promotional displays in book stores, resulting in increased sales and authors' royalties. The best-seller list is compiled from sales reports collected by the newspaper from a national sample of book outlets. But publishers can "teach to the test," identifying stores to be sampled and organizing bulk purchases at them. The *Times* cannot always successfully monitor store sales to identify such artificial purchases that corrupt the representativeness of the index.

▶ *U.S. News and World Report* publishes an annual ranking of colleges, truly an accountability system because college boards of trustees sometimes consider the rankings when determining presidential compensation. Rankings are based partly on how selective a college is, determined by the percentage of admitted applicants. (More selective colleges admit a smaller percentage of applicants.)

Selectivity would be a reasonable indicator if there were no stakes attached to measuring it. Colleges that accept relatively few applicants are likely to have higher quality, but once this indicator became an accountability measure, colleges had incentives to boost their own rejection rates. Some send promotional mailings to unqualified applicants, drop application fees or send already-completed applications to high school seniors to sign. The indicator has thus lost much of its value.

▶ The U.S. Department of Transportation requires airlines to report the percentage of flights that departed and arrived on time, defined as within 15 minutes of the published schedule. The department, consumer groups and members of Congress who advocated such reporting believed that travelers would be more likely to choose airlines with better on-time performance, and this would be an incentive for the airlines to

ILLUSTRATION © BY DAVID CLARK

improve. To avoid incentives for airlines to hurry departures in unsafe conditions, flights delayed because of mechanical difficulties were excluded from the calculations.

Airlines responded by reporting more phony mechanical difficulties when flights were late. And they padded schedules — when more time was allotted, flights' on-time performance improved. This did nothing to accomplish the Transportation Department's stated objective, to improve on-time performance on previously published schedules, which were purported to be realistic.

▶ As a presidential candidate, Richard M. Nixon promised in 1968 to reduce crime nationwide. After his election, the Federal Bureau of Investigation publicly reported crime statistics by city. It judged whether police departments were effective by tracking the sum of crimes in several categories, one of which was serious larceny (where the loss was worth at least $50). Many cities subsequently posted significant reductions in crime. The biggest reductions were in larcenies of $50 and over in value.

Valuing larceny is a matter of judgment, so police departments placed lower values on losses after the accountability system was implemented. Although crime reportedly declined, the number of $49 larcenies increased.

▶ Other public sectors have had similar experiences. The federal government has held local job training agencies accountable for placing unemployed workers in jobs that last at least 90 days. Some agencies responded by providing child care and transportation to workers for the first 90 days of employment, terminating these services on the 91st day. Other agencies simply refused to enroll the most difficult-to-place unemployed workers. Others cut back on educational activities designed to train workers for higher-paying and longer-lasting jobs because only short-term employment counted in the accountability system.

▶ The Medicare system has issued report cards on health providers. One has been based on mortality rates of patients who undergo open heart surgery. Some hospitals and physicians responded simply by refusing to operate on the sickest patients. Because the accountability system attempted "risk adjustment," statistically controlling for patient characteristics, other providers simply claimed the patients were sicker than they were. The distortions were so great that Medicare abandoned the system in 1993. The cur-

rent administration has reinstituted it.

▶ Medicare also issues report cards on nursing homes, based on whether they meet 15 recognized quality standards — for example, the percentage of residents who have pressure sores (from being turned in bed too infrequently). Because nurses' time is limited, if they spend more time complying with the turning-patients-in-bed standard for which they are held accountable, they may have less time to maintain hygienic standards by washing hands regularly, something for which they are not held accountable. Following the introduction of the report card, performance on the accountability indicators improved, but adherence to many other standards (like hand washing) declined, resulting in poorer overall quality in nursing homes.

The U.S. General Accountability Office reviewed health care report cards and concluded: "[A]dministrators will place all their organizations' resources in areas that are being measured. Areas that are not highlighted in report cards will be ignored."

Gaming Incentives

For business organizations generally, quantitative measures of performance are used warily and never exclusively. Even stock prices or profit are not simple guides to public companies' performance and potential. The Securities and Exchange Commission has complex regulations to prevent publicly traded firms from using numerical indicators to mislead investors. Yet financial data are still too complex for laypersons to interpret, which is why investors rely on sophisticated analysts, employed to discern the underlying and often nonquantifiable potential that stock prices or other easily measured characteristics might obscure. Equity markets can only exist because easily measured indicators are not transparent — buyers and sellers have different interpretations of what firms' financial indicators mean.

Corporate gamesmanship further limits the ability of the SEC and private accounting standards to prevent the distortion of numerical performance incentives. Executives whose compensation is based partly on corporate earnings can maximize their bonuses by manipulating depreciation schedules for long-term assets; by varying whether shipments to or from inventories should be accelerated or delayed at the end of an accounting period; by transferring other revenues or expenses from one accounting period to another; by allocat-

ing overhead to inventories; and by shifting whether major repair activities, research and development or even advertising expenses should be capitalized or expensed.

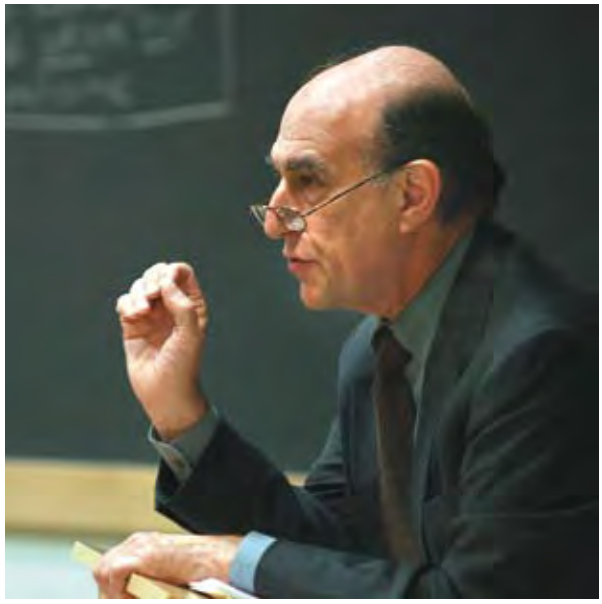
Most private-sector jobs, like teaching, include a composite of easily measured and less-easily measured responsibilities. Adding multiple measures of accountability is, by itself, insufficient to minimize goal distortion if the added measures are also quantitative. For example, one of the nation's largest banks determined that branch managers should not be rewarded only for short-term branch financials, but also for other measures that contributed to long-term profitability, such as customer satisfaction as determined by surveying customers. One manager boosted his ratings, and thus his bonuses, by serving free food and drinks, but this did nothing to boost the bank's long-term financial prospects.

Because of the ease with which most employees game purely quantitative incentives, most private-sector accountability systems blend quantitative and qualitative measures, with emphasis on the latter. McDonald's does not evaluate store managers by sales volume or profitability alone. Instead, a manager and his or her supervisor establish targets for easily quantifiable measures, but also less easily quantifiable product quality, service, cleanliness and personnel training because these factors may affect long-term profitability, as well as the reputation of other outlets over which a local manager has no control. Wal-Mart uses a similar system for professional employees, as do most other private organizations that engage in employee evaluation for purposes of pay.

Certainly, supervisory evaluations of employees are less reliable than objective, quantitative indicators. Supervisory evaluations may be tainted by favoritism, bias, inflation and even kickbacks or other forms of corruption. Yet the fact that subjective evaluations are so widely used, despite these flaws, suggests that most private-sector employers consider quantitative judgment even worse.

Accountability Scorecards

Managing accountability in the private sector is labor intensive. Bain and Co., the management consulting firm, advises clients that judgment of results should always focus on long-term, not short-term (and more easily quantifiable), goals. A company director estimated that at Bain itself, each manager devotes about 100 hours a year to evaluating five employees for purposes of its incentive pay system. "When I try to imagine a school principal doing 30 reviews, I have trouble," he observed.



Richard Rothstein, a research associate with the Economic Policy Institute, speaks frequently to education audiences.

A widespread business reform in recent decades has been total quality management, inspired by W. Edwards Deming, who warned that businesses seeking to improve quality and thus long-term performance should eliminate work standards (quotas), eliminate management by numbers and numerical goals and abolish merit ratings and management by objective because all of these encourage employees to focus on short-term results.

A corporate accountability tool that has grown more recently in popularity is the balanced scorecard, also first proposed in the early 1990s because business management theorists concluded that quantifiable short-term financial results were not an accurate guide to future profitability. Firms' goals were too complex to be reduced to a few quantifiable measures because predicting future performance relies not only on past financial success, but on subjective judgments of product quality, employee motivation, internal corporate cohesion and customer satisfaction and loyalty.

Curiously, the federal government administers a balanced scorecard approach, simultaneously with its test score-based No Child Left Behind Act. Each year since 1988, the U.S. Department of Commerce has made Malcolm Baldrige National Quality Awards for exemplary institutions in manufacturing and other business sectors. Numerical performance indicators play only a small role in award decisions. For the private sector, 450 out of 1,000 points are for "results," although even here, some results, such as ethical behavior, social responsibility, trust in senior leadership, workforce capability and capacity, and customer satisfaction and loyalty are difficult or impossible to quantify.

The Baldrige award program and its principles were extended to health and education institutions in 1999. For school districts, 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 was given in 2005 to the Jenks, Okla., school district. The Department of Commerce cited the school district's test scores as well as low teacher turnover and innovative programs such as an exchange relationship with schools in China and the enlistment of residents in a long-term care facility to mentor kindergartners and pre-kindergartners. Yet in 2006 the Jenks district was deemed by NCLB to be substandard because students had failed for two consecutive years to make adequate yearly progress in reading test scores.

A good accountability system in education — one that takes account of both the easily measured and the subjectively evaluated indicators of quality — will be expensive, far more expensive than NCLB's reliance on flawed standardized tests. In designing a good accountability system, policymakers should take to heart the calls of Mayor Bloomberg and Eli Broad to model incentives after those that are actually in use in the private sector. ■

Richard Rothstein is a research associate at the Economic Policy Institute in Washington, D.C.
E-mail: rrothstein@epi.org

PHOTO COURTESY OF CARLTON FORBES